Acta Logopaedica (2025), Vol. 2 | 98–120
DOI: 10.58986/al.2025.41653

Acta Logopaedica

ORIGINAL RESEARCH

Auditory-Perceptual Assessment of Gender Expression in Voice (PAGE-V): development and initial evaluation of a clinical rating protocol

Jenny Holmberg[1,2], Maria Södersten[3,4], Fredrik Nylén[1]

[1]Department of Clinical Sciences, Division of Speech and Language Pathology, Umeå University, Umeå, SE-901 87, Sweden
[2]Umeå Centre for Gender Studies, Umeå University, Umeå, SE-90187 Umeå, Sweden
[3]Division of Speech and Language Pathology, Department of Clinical Science, Intervention and Technology, Karolinska Institutet, Stockholm SE-171 77, Sweden
[4]Medical Unit Allied Health Professionals, Section Speech and Language Pathology, Karolinska University Hospital, Stockholm SE-141 86, Sweden

***Corresponding author:** Jenny Holmberg, Department of Clinical Sciences, Division of Speech and Language Pathology, Umeå University, Umeå, SE-901 87, Sweden. Email: jenny.c.holmberg@umu.se

Publication date: 1 September 2025

Abstract

Gender-affirming voice training aims to assist transgender and gender-diverse clients in reaching a voice in better alignment with their gender identity. We describe the development and evaluation of the Auditory-Perceptual Assessment of Gender Expression in Voice (PAGE-V) protocol for assessing voice and speech features commonly targeted in gender-affirming voice training. Features previously identified to influence the perception of femininity and masculinity were evaluated in two steps. First, seven speech and language pathologists (SLPs) reviewed PAGE-V based on their experience in providing gender-affirming voice training. Second, six SLPs completed listener training and assessed 45 voice samples with varying gender expressions using PAGE-V. The results showed that individual SLPs rated items relatively consistently; agreement among SLPs was lower. The PAGE-V was concluded to provide a clinically useful tool, and the insights gained support further development of PAGE-V. Listener training sessions are needed to strengthen SLPs' internal representations and consensus on protocol items.

Keywords: auditory-perceptual voice assessment; listener training; voice quality; resonance; intonation

Sammanfattning
Könsbekräftande röstbehandling syftar till att hjälpa transpersoner och personer med varierande könsidentiter att nå en röst som bättre överensstämmer med deras könsidentitet. Vi beskriver och utvärderar protokollet ”Auditory-Perceptual Assessment of Gender Expression in Voice” (PAGE-V) för att bedöma röst- och talegenskaper av betydelse för att utvärdera könsbekräftande röstbehandling. Utvärderingen av protokollet genomfördes i två steg. Först granskade sju logopeder PAGE-V baserat på deras erfarenhet av att ge könsbekräftande röstbehandling. Därefter genomgick sex logopeder en lyssnarutbildning och bedömde 45 röster enligt PAGE-V. Resultaten visar att enskilda logopeder betygsatte protokollets egenskaper relativt konsekvent, men att överenstämmelsen mellan logopederna var lägre. PAGE-V sågs av logopederna som ett kliniskt användbart verktyg, och återkopplingen stärker vidareutvecklingen av PAGE-V. Lyssnarträning kommer dock att krävas för att säkerställa god konsensus i användning av protokollet.

Keywords: auditiv-perceptuell röstbedömning; lyssnarträning; röstkvalitet; resonans; intonation

Introduction

For transgender and gender diverse (TGD) people, their gender identity does not align with the gender they were assigned at birth. This incongruence may lead to discomfort and distress, *gender dysphoria*, that has negative effects on well-being and participation in social life (American Psychiatric Association [APA], 2022; World Health Organization [WHO], 1993). For many, the experience of a voice that does not match their gender identity can be a prominent part of gender dysphoria (Ziltzer *et al.*, 2023). TGD people may therefore wish to modify their voice in better alignment with their gender identity.

Gender-affirming voice training aims to assist the TGD client towards a preferred gender expression in voice in a way that is not harmful for the voice mechanism (Coleman *et al.*, 2022). Clients' preferred expression in voice may align with binary cisgender norms on voice as signalling either a female or male speaker gender, or they may want to express varying degrees of femininity and masculinity, according to a gender identity outside the binary gender categories of woman and man (Holmberg *et al.*, 2023). To match varying gender expressions, gender-affirming voice training commonly focuses on voice and speech aspects that influence the perception of femininity and masculinity in voice (Davies and Goldberg, 2006; Davies *et al.*, 2015; Oates and Dacakis, 1997). These aspects include pitch, loudness, resonance and voice quality as well as intonational patterns, extent and frequency of pitch change in intonations (Davies and Goldberg, 2006; Oates and Dacakis, 1997). Aspects related to the production of speech sounds include articulatory preciseness and durational characteristics in terms of words and phrases being produced separately or with sustained voicing through speech sounds (Davies *et al.*, 2015). Structured assessments of these voice and speech features are needed to decide on which features to target in training to best fit the client's expressed goals. For auditory-perceptual voice assessment, the standard rating scales commonly used in speech and language pathology (SLP) practice were developed primarily to capture aspects of dysphonia in voice disorders (Hammarberg *et al.*, 1980; Hirano, 1981; Kempster *et al.*, 2009). There is

no existing published rating protocol specifically constructed to capture voice features relevant for assessing gender expression in voice.

While auditory-perceptual analyses are fundamental in voice assessment and clinical voice training, their applicability may be reduced by low agreement among raters (De Bodt *et al.*, 1997; Iwarsson and Petersen, 2012; Kempster *et al.*, 2009), even in highly experienced listeners (Kreiman *et al.*, 1993). Previous research has suggested that variability in rater agreement is an issue of rating task design, rather than listeners being unreliable (Kreiman *et al.*, 2007). Familiarity with rating scales from auditory-perceptual rating protocols with congruent definitions may be a supporting factor for increased agreement. The 'Stockholm Voice Evaluation Approach' (SVEA) (Hammarberg, 2000) is the predominant auditory-perceptual rating protocol for dysphonia used by SLPs in Sweden and may therefore support development of a new rating protocol based on familiarity with the SVEA protocol and the assessed voice dimensions. Increased agreement among raters may further be supported by thoroughly considering which type of scale and scale granularity is appropriate for the rating of a specific perceptual parameter (Kreiman *et al.*, 2007; Pearse, 2011). When a high level of granularity is desired, the visual analogue scales (VA-scales) have been demonstrated to result in comparably high agreement, whereas rivalling scales, such as the equally appearing interval (EAI) scale, have indicated lower agreement (Kreiman *et al.*, 2007) and drifts in listeners' ratings over time (Gerratt *et al.*, 1993; Kreiman *et al.*, 1993).

An additional factor to consider is that agreement among raters may increase if listeners are provided with anchor voices representing defined scale intervals during the listening task (Gerratt *et al.*, 1993; Kreiman *et al.*, 2007). Agreement has also been seen to increase with listener training prior to the rating task, where the pre-listening task preparation having included written definitions of the features to be rated, the use of example voices representing voice features of different grades of severity, and listeners' joint discussions on their ratings (Chan and Yiu, 2002; Eadie and Baylor, 2006; Hammarberg, 2000; Iwarsson and Petersen, 2012). Therefore, the presentation of a rating task, choice of rating scales and endpoint labels, and the information provided to the raters need to be thoroughly considered when constructing an auditory-perceptual rating protocol.

The aim of this study was to construct a clinical auditory-perceptual rating protocol for use in gender-affirming voice training and outcome assessment, named the Auditory-Perceptual Assessment of Gender Expression in Voice (PAGE-V). We report on the initial evaluation of the applicability of the protocol in terms of intra-rater consistency and inter-rater agreement among SLPs experienced in assessing TGD clients' voices from the perspective of their vocal gender expression.

Method

Ethical approval for this study was obtained from the Swedish Ethical Review Authority (Case No. 2019-05374). All speakers provided a written informed consent for their recordings to be used in a listening task. The study procedure was divided into two phases: the development of the rating protocol (phase 1) and the evaluation of the protocol (phase 2).

Phase 1—Development of the PAGE-V protocol

Protocol items

The choice of items represented in the PAGE-V protocol was primarily based on acoustic and perceptual aspects of voice and speech that have been identified in a systematic review by Leung *et al.* (2018) to influence the perception of gender expression in voice. Protocol items comprising auditory-perceptual parameters that may be relevant to address in gender-affirming voice assessment and training targeting voice feminisation or masculinisation were identified. Additionally, items related to the functional aspects of voice were added, as these aspects may need to be addressed in voice training. For example, hyperfunction may be heard in trans men's voices when attempting to reach a lowered speaking pitch at the physiological boundaries of their pitch range, and in trans women who lack an efficient voice technique for raising their pitch. A hyperfunctional vocal behaviour may result in vocal fatigue, which has been seen to occasionally follow from hormonal treatment in trans men (Azul *et al.*, 2017; Nygren *et al.*, 2016), from frequency-raising vocal fold surgery (Kelly *et al.*, 2019) and voice training (Leyns *et al.*, 2022) in trans women. Conversely, a habitual hyperfunctional vocal behaviour may prevent the TGD client from reaching their preferred pitch, resonance, or voice quality and is therefore often addressed in voice training.

In total, 18 items were grouped into the domains *pitch*, *loudness*, *voice quality*, *resonance*, *intonation*, and *articulation* in the protocol. Two additional items were summative ratings of *perceived level of femininity* and *perceived level of masculinity* at the end of the protocol. The rating protocol further allowed the rater to note other auditory-perceptual features perceived in a voice. Written descriptions of the perceptual, acoustic, and/or physiological correlates of each item (Table 1), inspired by Hammarberg (2000), were provided together with the protocol.

Rating scales

The protocol used two rating scales based on the item characteristics: VA-scales for continuous ratings, and tick-boxes for categorical or ordinal ratings. The VA-scales of 100 mm in length were used to rate the perceived degree of the feature in a voice sample, for example breathiness or vocal fry. The scales were labelled with specified endpoints that matched the item characteristics. For example, the item *pitch variability* was labelled with the endpoints '*monotonous*' and '*very varied*', while the item *articulatory preciseness* was labelled with the endpoints '*unprecise*' and '*precise*'. Items related to voice quality, such as *breathiness*, were labelled with the endpoints '*lack of*' and '*high degree of*', in agreement with Hammarberg (2000). For a few items, we viewed it more suitable to provide the rater with labelled options that described, for example, vocal register or phrase-final intonation patterns most frequently used by the speaker. Protocol items and their corresponding rating scales are presented in Table 1.

Review of the rating protocol by a group of experts

Seven SLPs working with TGD voice clients were asked to review the rating protocol regarding the relevance of the proposed items and scales as well as the clarity of the item

https://doi.org/10.58986/al.2025.41653

Table 1. A description of the protocol items, the vocal domains they relate to, their perceptual, acoustic, and/or physiological correlates, and the type of scale used for the item in the PAGE-V protocol.

Domain	Item	Description	Scale
Pitch	Gender normative pitch*	The perception of speaking pitch to be within cisgender-normative values for men, women, or corresponding to a gender-neutral speaking pitch.	Categorical scale with three options *Corresponding to cisgender norms for women, men,* and *gender neutral*.
	Pitch	The main auditory correlate of fundamental frequency,[#] related to the rate of vocal fold vibrations.[¤]	VAS
Vocal loudness	Vocal loudness	The main auditory correlate of sound pressure level.[#] Regulated mainly by the subglottal pressure.[¤]	VAS
Voice quality	Breathiness	Audible turbulent noise due to insufficient glottal closure during phonation.[#]	VAS
	Hypofunction	Insufficient vocal fold tension, resulting in a weak, lax voice.[#]	VAS
	Hyperfunction/ tense voice	Strained phonation due to constriction of vocal folds and laryngeal tube during phonation.[#]	VAS
	Vocal fry	Low-frequency, periodic vocal fold vibrations with long closed phases and short open phases.[#]	VAS
	Flow phonation	An increased transglottal air flow during phonation, leading to a complete glottal closure with little adductive force resulting in large vocal fold amplitude and a high relative level of the fundamental.[¤]	VAS
	Instability	**Unstable pitch:** The speaking pitch is mostly stable with occasional fluctuations into a distinctly higher or lower pitch range.[#] **Unstable voice quality:** Fluctuations from modal register into middle or falsetto, or from a predominantly middle or falsetto register into modal register.[#] **Voice breaks:** Intermittent sudden breaks between registers, usually from modal to falsetto.[#]	Categorical scale on which either *no instability* or one or more of the options *unstable pitch, unstable voice quality*, and *voice breaks* were selected.

(continues)

Table 1. Continued.

Domain	Item	Description	Scale
	Vocal register	**Modal register:** Vocal folds vibrate with a nearly or complete glottal closure, long closed phase, and a mucosal wave generating a sonorous voice.[#] A perceptually different sound quality, usually produced within a lower frequency range, compared to middle and falsetto register. **Middle register:** The vocal folds are stretched with smaller vibrating mass, compared to modal register. The relative level of the fundamental is higher,[¤] and the voice sounds less sonorous than in modal register. **Falsetto:** Stretched and thin vocal folds resulting in short or incomplete vocal fold closure and no mucosal wave, generating a thin and slightly breathy voice.[#]	Categorical scale with options *modal register, falsetto, middle register,* and *cannot decide*.
Intonation	Pitch variability	Pitch variation/liveliness, related to extent and rate of pitch change in connected speech.[ϕ]	VAS
	Phrase-final intonation pattern	Patterns of pitch variation predominantly used by the speaker.[‡]	Categorical scale with options *predominantly falling, rising, level*, and *no predominant intonation pattern*
	Extent of pitch change in phrase-final intonations	Extent of pitch change in phrase-final intonations.[ϕ]	Ordinal scale with options *small, moderate*, and *large pitch change*.
Resonance	Resonant voice	Tuning of the supraglottic cavities, often resulting in a sensation of vibrations in the face and speaker-perceived "easy phonation". The vocal folds barely ab- and adduct, leading to a gentle and efficient voice production.[ſ]	VAS
	Oral resonance	The size of the oral resonant space (related to the position of the jaw/tongue/lips) influences the perception of a bright or dark resonance.[§]	VAS

(continues)

https://doi.org/10.58986/al.2025.41653

Table 1. Continued.

Domain	Item	Description	Scale
	Pharyngeal resonance	The size of the pharyngeal resonant space from the vocal folds to the base of the tongue influences the perception of a bright or dark resonance.	VAS
Speech	Articulatory preciseness	A precise articulation is characterised by distinct phonetic contrasts between vowels, and precise articulation of consonant sounds. Imprecise articulation may include reduction to central vowel [ə], reduced or distorted phonemes.[§]	VAS
	Articulatory transitions	Smooth, blended transitions between words in connected speech gives an impression of an overall legato sound, in contrast to hard glottal attacks and heavy articulatory contacts in word initial consonants.[§]	VAS
	Speech rate variability	Speech rate variations.[$]	VAS
Overall gender expression in voice	Femininity	The perceived level of femininity in connected speech.	VAS
	Masculinity	The perceived level of masculinity in connected speech.	VAS

Notes: VAS = visual analogue scale.
[*]Item added after the initial review by seven voice experts; [#]Hammarberg (2000); [¤]Sundberg (2001); [ł]Henton (1995); [‡]Avery and Liss (1996); [J]Verdolini *et al*. (1998); [§]Hirsch et al. (2019); [$]Leung *et al*. (2018).

labels and descriptions. All SLPs were experienced in performing auditory-perceptual voice evaluations and familiar with the structure and terminology used in the SVEA protocol (Hammarberg, 2000). The descriptions of the auditory-perceptual items and their physiological base were discussed to reach a shared understanding (consensus) of the terminology and scale labels (Hammarberg *et al.*, 1980; Iwarsson and Petersen, 2012). The SLPs regarded all proposed items relevant and meaningful to use in voice training for TGD clients. Following the discussions, one item (pitch) was transformed into a two-step rating. First, the raters indicated the pitch as perceived to be within the gender norms for female or male voices, or to be gender neutral. Second, pitch was rated on a VA-scale, with the gender range norms serving as references. The endpoints of the pitch rating scale were labelled '*very low*' and '*very high*', respectively.

Phase 2—Evaluation of the rating protocol

The protocol was evaluated in a listening task 2½ weeks after phase 1. The evaluation was based on the intra-rater consistency and inter-rater agreement within

https://doi.org/10.58986/al.2025.41653

and among the SLPs' ratings. Further, a topical survey (Sandelowski and Barroso, 2003) was performed orally in which the SLPs were asked about their experiences using the protocol and their thoughts about the potential further development of the protocol. The survey was performed directly following the rating procedure as a group discussion, with all participants present in the same room. Six of the seven SLPs who participated in phase 1 agreed to participate in the listening task in phase 2. The six SLPs had on average 9.8 (1–16) years of experience in providing gender-affirming voice training to TGD clients. A training session was conducted to ensure consensus regarding terminology in relation to the rating protocol prior to the listening task.

Voice recordings

Speakers

Voice recordings representing a variety of gender expressions in voice were selected for listener training and ratings using the protocol. The speakers were recruited from the voice client load at two SLP clinics and by convenience sampling as part of a larger project (Holmberg *et al.*, 2024; Nylen *et al.*, 2024). Out of the 59 selected voices, 14 were selected for listener training and thus not included in the listening task. The remaining 45 voice samples included 34 recordings from TGD speakers and 11 from cisgender speakers (6 men and 5 women). Among the TGD speakers, 17 identified as women, among whom 3 were re-transitioning women, assigned females at birth but who had experienced voice change due to testosterone treatment during previous gender identification as men. Of the 17 TGD speakers who did not identify as women, 10 identified as men, whereas 7 reported a non-binary gender identity. All recorded participants were native speakers of Swedish between 19–60 years of age.

Recording procedure

The recordings of the speech samples were made at a location convenient to the speaker, most often at a hospital in a separate room to reduce ambient noise. The equipment was calibrated to a reference tone prior to recording. The recording context did, however, not allow the inclusion of a reliable reference tone into the recording. The recordings were made with an omnidirectional RØDE SmartLav+ microphone with a frequency range of 20 Hz–20 kHz and a signal-to-noise ratio of 67 dB. The microphone was head-mounted at a distance of 5 cm from the angle of the speaker's mouth (Svec and Granqvist, 2010) and connected to an Android mobile phone that allowed for automatic gain control to be disabled. The mobile application *Noise* (Swedish Work Environment Authority, 2019) was used to ensure that the surrounding sound level was less than 38 dB(A) to achieve good sound quality according to recommendations for instrumental voice assessment (Patel *et al.*, 2018).

Speech material

The collected speech material consisted of spontaneous speech, as it was considered to best represent speakers' habitual expression in voice and speech and to show a

https://doi.org/10.58986/al.2025.41653

variety in speaker intonation patterns, articulatory preciseness, and other features of speech. The speech samples presented in the listening task were approximately 20–30-seconds long. Sentences that could potentially identify the speaker or lead the listener to the perception of the speaker being of a specific gender, for example '*my son's father*', or '*me and the other tenors*', were removed. In addition, non-speech sounds, such as laughter and low-frequency coughs, were also removed to avoid these influencing the perception of a voice.

Listener training and forming a consensus

The revised PAGE-V protocol and the definitions of the 21 protocol items were presented to the SLPs. The participants took part in two practice rounds, in which the PAGE-V protocol was applied in individual ratings of two recorded voices. After each practice round, the SLPs compared their ratings in pairs and within the group (Eadie and Baylor, 2006). When consensus was not reached, additional voice recordings were provided as anchors for a shared understanding of the expected range of each particular voice feature. To form a consensus, the SLPs were instructed to imitate the voice quality and resonance types to use proprioception for their internal understanding of the correlation between production and perception of the voice feature, as has been suggested by Iwarsson and Petersen (2012). Directly following the consensus procedure, the SLPs did the listening task individually.

Assessment of PAGE-V items after consensus

A structured listening task was performed in which the SLPs rated PAGE-V items. The evaluation took place in a room that allowed each SLP to be seated at a stationary computer at a sufficient distance from other SLPs so that they would not see each other's ratings. Over-ear headphones (SONY MDR-ZX660) were used to facilitate acoustically appropriate and comparable listening conditions, and to avoid that the SLPs would be disturbed by surrounding noise. The listening task included 45 unique voice recordings, of which nine were presented twice (54 voice samples in total). The voice samples had been randomized into five blocks consisting of 10 or 11 voice recordings each. The five blocks were presented in a randomized order to the SLPs who were encouraged to take shorter breaks after every block. Three longer breaks of 15, 60, and 25 minutes, respectively, were scheduled every 1½–2 hour during the day. With the breaks excluded, the rating protocol evaluation lasted 5½ hours (and approximately 6 minutes per recording on average). Each sample was set to play in a continuous loop after the start of the recording. Hence, the SLPs could listen to a voice sample as many times as needed to rate all protocol items and decide when to start the next recording (Helou *et al.*, 2010). To resemble a clinical setting, the SLPs were not allowed to go back and revise the ratings for a previous voice sample.

Statistical analysis of intra-rater consistency and inter-rater agreement

To evaluate the intra-rater consistency within the SLPs and inter-rater agreement among SLPs of items in the PAGE-V protocol, separate statistical approaches were

used depending on the type of response given for each protocol item. For continuous data (from VA-scales, 16 items; see Table 1), an intra-class correlation coefficient (ICC) two-way mixed effects model (with an absolute agreement definition and calculations based on single measurements and single raters) was used for analysing intra-rater consistency and inter-rater agreement, respectively (Koo and Li, 2016). Values less than 0.5 were considered to be indicative of poor intra-rater consistency/inter-rater agreement, values between 0.5 and 0.75 indicated moderate, values between 0.75 and 0.9 good, and values greater than 0.9 excellent intra-rater consistency/inter-rater agreement (Koo and Li, 2016).

The protocol items *gender normative pitch*, *instability*, *vocal register*, and *phrase-final intonation pattern* were rated on nominal scales and assessed in terms of intra-rater consistency and inter-rater agreement using Cohen's, and Fleiss' kappa (κ), respectively. The level of intra-rater consistency and inter-rater agreement was considered in relation to the discrete ranges none or minimal ($\kappa < 0.39$), weak ($\kappa = 0.40–0.59$), moderate ($\kappa = 0.60–0.79$), strong ($\kappa = 0.80–0.90$), or almost perfect ($\kappa > 0.90$). The items *instability* and *vocal register* were assumed to show little variety due to the voice samples representing non-dysphonic voices. As little variety may lead to decreased κ values regardless of agreement or disagreement among raters (Tinsley and Weiss, 1975), intra-rater consistency and inter-rater agreement were also evaluated in terms of per cent exact agreement for these items. Based on McHugh (2012), a per cent exact agreement of less than 50% was considered poor, and a per cent exact agreement of 50–80% was considered to show a moderately strong intra-rater consistency or inter-rater agreement. A per cent exact agreement level of 80% was considered as the minimum acceptable agreement, and >90% excellent agreement (McHugh, 2012).

The protocol item *extent of pitch change in phrase-final intonations* was rated on a three-step ranking scale with options *small*, *medium*, and *large extent of pitch change*. For these ordinal data, Cohen's weighted kappa with quadratic weighting was used for analysing the intra-rater consistency, while the inter-rater agreement was evaluated in terms of per cent exact agreement. Similar to the nominal scales, the level of intra-rater consistency was interpreted as none or minimal ($\kappa < 0.39$), weak ($\kappa = 0.40–0.59$), moderate ($\kappa = 0.60–0.79$), strong ($\kappa = 0.80–0.90$), or almost perfect ($\kappa > 0.90$).

Intra-class correlation coefficient and kappa estimates, and their 95% confidence intervals, were calculated using SPSS, version 29.

Results

Four of the six SLPs rated all 54 voice samples, whereas two SLPs completed 44 of the 54 ratings during the time provided to complete the ratings (5½ hours). All listeners omitted to rate one or more of the 21 protocol items. Hence, the total number of ratings in the statistical analyses varied for different protocol items (see Table 2).

Overall distribution of ratings

Most items rated on VA-scales showed a distribution of ratings along the full length of the scale (Figure 1). However, for the items *breathiness*, *hypofunction*

https://doi.org/10.58986/al.2025.41653

Table 2. The number of ratings and missing data for all the rated protocol items divided into seven domains.

Domain	Item	Ratings (N)	Missing (N)
Pitch	Gender normative pitch	304	20
	Pitch	303	21
Vocal loudness	Vocal loudness	302	22
Voice quality	Breathiness	302	22
	Hypofunction	301	23
	Hyperfunction	300	24
	Vocal fry	302	22
	Flow phonation	299	25
	Instability	298	26
	Vocal register	301	23
Resonance	Resonant voice	304	20
	Oral resonance	303	21
	Pharyngeal resonance	302	22
Intonation	Pitch variability	301	23
	Phrase-final intonation pattern	303	21
	Extent of pitch change in phrase-final intonations	285	39
Speech	Articulatory preciseness	303	21
	Articulatory transitions	303	21
	Speech rate variability	301	23
Overall gender expression in voice	Femininity	303	21
	Masculinity	303	21

and *hyperfunction*, a skewed distribution was indicated, with ratings mainly reflecting that these voice features were absent or infrequently occurring in the provided voice samples. Ratings of vocal *loudness* were centred round the midpoint of the scale (for which *very weak* and *very loud* formed the endpoints). The new protocol items, not included in the SVEA protocol, showed a wide distribution of ratings.

Among the protocol items rated using categories, *instability* and *vocal register* showed a small variability of ratings. The SLPs rated 89% of the voice samples to show *no instability*, while *unstable voice quality*, *unstable pitch*, and *voice breaks* were reported in only 9.4%, 3.3%, and 2.4% of the rated samples, respectively. The protocol item *register* showed a preponderance of the rated voice samples corresponding to the perception of a *modal register* characterising the speaker's voice (77%), compared to a *middle register* (21.4%) or *falsetto* (0.4%).

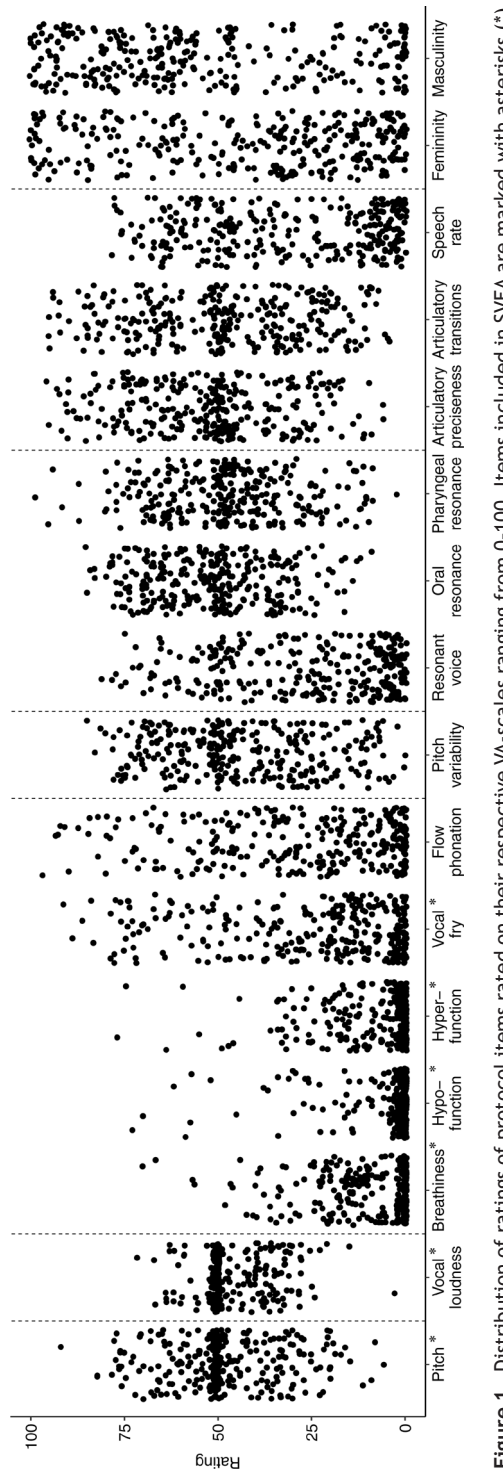

Figure 1. Distribution of ratings of protocol items rated on their respective VA-scales ranging from 0-100. Items included in SVEA are marked with asterisks (*).

https://doi.org/10.58986/al.2025.41653 Acta Logopaedica is published according to open access under a Creative Commons CC-BY 4.0 license

Intra-rater consistency and inter-rater agreement

In general, the level of intra-rater consistency was stronger than the level of inter-rater agreement, both for items rated on VA-scales and items rated on nominal or ordinal scales. Intra-rater consistency and inter-rater agreement are presented in more detail below.

The 16 items rated on VA-scales showed that the six SLPs were, to a high degree, consistent in their ratings of the nine duplicated voice samples. Excellent intra-rater consistency was seen for the items *femininity* (ICC = 0.96, 95% CI = [0.93–0.97]), *masculinity* (ICC = 0.94, 95% CI = [0.89–0.96]), and *pitch variability* (ICC = 0.91, 95% CI = [0.85–0.95]). Good intra-rater consistency was seen for items related to voice quality, resonance, and pitch; *vocal fry* (ICC = 0.84, 95% CI = [0.74–0.91]), *flow phonation* (ICC = 0.81, 95% CI = [0.70–0.89]), *breathiness* (ICC = 0.80, 95% CI = [0.68–0.88]), *resonant voice* (ICC = 0.87, 95% CI = [0.79–0.92]), *oral resonance* (ICC = 0.79, 95% CI = [0.67–0.88]), and *pitch* (ICC = 0.75, 95% CI = [0.61–0.85]). The SLPs were least consistent in their ratings of *pharyngeal resonance* (ICC = 0.55, 95% CI = [0.32–0.71]) and *vocal loudness* (ICC = 0.55, 95% CI = [0.34–0.71]), although with varying results for individual SLPs. Good inter-rater agreement was observed for the protocol items *femininity* (ICC = 0.76, 95% CI = [0.65–0.85]) and *masculinity* (ICC = 0.80, 95% CI = [0.70–0.88]). An overview of the level of intra-rater consistency and inter-rater agreement is presented in Table 3.

The item *gender normative pitch* was rated on a categorical scale with three options (*corresponding to cis men's voices*, *corresponding to cis women's voices*, and *gender neutral*), and showed almost perfect intra-rater consistency (κ = 0.94, 95% CI = [0.85–1.0]) and moderate inter-rater agreement (κ = 0.68, 95% CI = [0.62–0.74]). The items *instability* and *vocal register* showed excellent intra-rater consistency (96% and 91% exact agreement, respectively), while only the item *instability* showed good inter-rater agreement (81% exact agreement). Two items were related to speakers' intonation patterns. For item *phrase-final intonation pattern*, for which the SLPs indicated if they perceived the speech sample to show predominantly *falling*, *rising*, *level*, or *no predominant intonation pattern*, the raters showed moderate intra-rater consistency (κ = 0.75, 95% CI = [0.61–0.89]) and weak inter-rater agreement (κ = 0.42, 95% CI = [0.37–0.48]).

The protocol item *extent of pitch change in phrase-final intonations* was the only item rated on an ordinal scale, for which the SLPs were asked to categorise each voice sample according to their perception of the speakers' phrase-final intonation changes (*small*, *moderate*, or *large pitch changes*). The SLPs were, to a high degree, consistent in their individual ratings (κ = 0.85, 95% CI = [0.75–0.95]). The per cent exact agreement across the 45 voices and six raters was only 16%.

Experiences of the participating SLPs

The shared opinion among the participating SLPs was that the rating protocol would be a useful tool in clinical work with TGD voice clients. The new protocol items not included in SVEA were, in general, considered somewhat difficult to rate as the SLPs were not accustomed to assessing them in the structured way, and with the rating scales presented in the protocol. Items related to resonance were considered

Table 3. The intra-rater consistency and inter-rater agreement for assessed items in each domain. The range categories of agreement levels (A–D) are further indicated together with an indicative background colouring to support visual separation of levels.

Domain	Item	Level of intra-rater consistency	Metric	Level of inter-rater agreement	Metric
Pitch	Gender normative pitch	A	κ = 0.94	C	κ = 0.68
	Pitch	B	ICC = 0.75	D	ICC = 0.18
Vocal loudness	Vocal loudness	C	ICC = 0.55	D	ICC = 0.47
Voice quality	Breathiness	B	ICC = 0.80	D	ICC = 0.27
	Hypofunction	C	ICC = 0.70	D	ICC = 0.29
	Hyperfunction	C	ICC = 0.68	D	ICC = 0.29
	Vocal fry	B	ICC = 0.84	C	ICC = 0.61
	Flow phonation	B	ICC = 0.81	D	ICC = 0.25
	Instability	A	96%*	B	81%*
	Vocal register	A	91%*	D	44%*
Resonance	Resonant voice	B	ICC = 0.87	D	ICC = 0.06
	Oral resonance	B	ICC = 0.79	D	ICC = 0.24
	Pharyngeal resonance	C	ICC = 0.55	D	ICC = 0.28
Intonation	Pitch variability	A	ICC = 0.91	D	ICC = 0.32
	Phrase-final intonation pattern	C	κ = 0.75	D	κ = 0.42
	Extent of pitch change in phrase-final intonations	B	κ = 0.85	D	16%*
Speech	Articulatory preciseness	C	ICC = 0.63	D	ICC = 0.39
	Articulatory transitions	C	ICC = 0.63	D	ICC = 0.20
	Speech rate variability	C	ICC = 0.73	D	ICC = 0.12
Overall gender expression in voice	Femininity	A	ICC = 0.96	B	ICC = 0.76
	Masculinity	A	ICC = 0.94	B	ICC = 0.80

Notes: A = excellent/almost perfect (ICC > 0.90, κ > 0.90, per cent exact agreement >90%).
B = good/strong (ICC = 0.75–0.90, κ = 0.80–0.90, per cent exact agreement >80%).
C = moderate (ICC = 0.50–0.75, κ = 0.60–0.79, per cent exact agreement >50%).
D = poor/weak (ICC < 0.50, κ < 0.60, per cent exact agreement <50%).
*Per cent exact agreement.

especially challenging, and the perceptual characteristics of a *resonant voice* were discussed. The items *oral* and *pharyngeal resonance* were considered difficult to differentiate by many of the SLPs. However, the SLPs viewed the separate ratings of these items as providing a more detailed analysis that may assist in identifying relevant targets for gender-affirming voice training. The varying endpoints of VA-scales, where the scale midpoint was sometimes thought to represent a neutral position (e.g. pitch or loudness) and sometimes seen to represent a mid-value (50 out of 100; e.g. for the breathiness and vocal fry items) was mentioned by the SLPs to somewhat increase the cognitive load. There was a reported need for more practice sessions in using the protocol. Anchor voices and repeated consensus listener training with colleagues were suggested to assist in becoming more skilled at performing structured auditory-perceptual assessments of voice and speech within gender-affirming voice training.

Discussion

Auditory-perceptual assessments are an essential part of clinical voice evaluations, and validated rating protocols are employed to conduct these evaluations in a structured and consistent manner. In the absence of a rating protocol that includes voice and speech features of importance for the perception of femininity and masculinity, the PAGE-V protocol was developed to afford specifically the assessment of perceptual features commonly targeted in gender-affirming voice training. The initial evaluation of the PAGE-V protocol provides support for structured clinical ratings of the included voice and speech features to be meaningful and reliable in terms of intra-rater consistency.

The PAGE-V includes some new items compared to the SVEA-protocol (Hammarberg, 2000), which the participating SLPs were not used to assessing. These items showed a wide distribution of ratings. The fact that the SLPs placed ratings throughout the full length of the rating scales indicates that they perceived between-speaker differences and that the rating scales were able to capture perceived item variability. Inter-rater agreement has been seen to increase with raters' experience and training in using a rating scale (Dejonckere *et al.*, 1998). Therefore, a stronger inter-rater agreement would have been expected for the items included in the SVEA protocol and, thus, familiar to the expert listeners. However, this was not always the case, and the results did not show all new items to be associated with a lower level of inter-rater agreement.

The new items related to the resonance domain proved particularly challenging to the raters. The responses to the item *resonant voice* rendered discussions among the participating SLPs related to the expected variability in an item for a representative sample. Furthermore, the discussion centred on which scale endpoints would be most suitable for revealing this variety. In addition, discussions about how to perceptually differentiate between resonant voice, middle register, and flow phonation indicated that the raters differed in the extent to which they used these concepts. For a reliable rating protocol to be constructed, agreement on definitions is essential (Iwarsson and Petersen, 2012), and the variability observed among the participating SLPs, therefore, indicates a need for establishing clear definitions and further collegial discussions about these features. The items *oral resonance* and *pharyngeal*

resonance were considered difficult to differentiate by the expert listeners. However, the separation of the two items was viewed as valuable for performing a voice assessment that is appropriate to guide many TDG clients' training goals. Overall, increased listener training and the use of more example voices were considered needed to increase confidence in rating the items of the *resonance* domain.

The item *speech rate variability* was also found to create highly variable ratings due to the listening task not being sufficiently clear and lacking information about which temporal aspects of speech to consider, and whether or not, to consider pausing. A clear definition of the item *speech rate variability* is needed in future work on PAGE-V. Alternatively, the lack of strong support for speech rate variability signalling a degree of femininity or masculinity (Dacakis *et al.*, 2012; Leung *et al.*, 2018) may warrant the exclusion of this item in future revisions of the protocol.

The items rated on VA-scales that showed the most narrow distribution of ratings were *vocal loudness*, *breathiness*, *hypo-*, and *hyperfunction*. For these items, ICC computed within and among raters may be misleadingly low due to the ICC being sensitive to low variance (Graham *et al.*, 2012). All four items are included among the voice features that the SLPs were familiar in evaluating on VA-scales. The skewed distribution of ratings of these features may, therefore, less easily be explained in terms of familiarity among the raters or the rating scales being ill-fitted to capture potential variability within the voice features. Rather, the skewness reflects that the rated voice samples belonged to speakers without voice disorders. Vocal loudness, breathiness, hypo-, and hyperfunction may, however, be signs of inefficient voice use that may hinder or be the effect of voice modification (Azul *et al.*, 2017). For example, when transfeminine clients raise their pitch, they may increase vocal loudness (Dahl and Mahler, 2020), potentially leading to hyperfunction. Hyperfunction may, in turn, contribute to vocal fatigue or vocal fold lesions (Hillman *et al.*, 2020) or difficulties with voice projection. Similarly, breathiness can be a sign of feminine voice quality as well as a pathological voice (Leung *et al.* 2018). As excessive or reduced loudness, increased breathiness, hypo-, or hyperfunction may indicate a less efficient or potentially harmful voice use, these items are suggested to be included in auditory-perceptual voice assessments performed within gender-affirming voice training.

All items showed moderate to excellent intra-rater consistency, while inter-rater agreement was good only for items *masculinity*, *femininity*, and *instability*, and moderate for items *vocal fry* and *gender normative pitch*. The observation of a stronger intra-rater consistency than inter-rater agreement is in line with previous studies on listeners' auditory-perceptual voice ratings (Granqvist, 2003; Iwarsson and Petersen, 2012). Low levels of inter-rater agreement when performing auditory-perceptual assessments have been suggested in previous research to be in part due to listeners' differing internal representations of voice qualities (Kreiman *et al.*, 1993). In fact, clinical training and experience of a wide variety of voices have been suggested to lead expert listeners to develop individual prototypes, or 'internal representations', for different voices, which may cause listeners to differ in which voice features they primarily attend to (Kreiman *et al.*, 1990). Although expert listeners in our study were not used to formally assess all items using the provided rating scales, they were not unfamiliar with the included voice and speech features and may have transferred their clinical experience to internal representations also of the new items, leading to differing rating patterns among listeners. To make individual raters' internal

representations better matched, joint listener training has been recommended to increase intra-rater consistency and inter-rater agreement (Chan and Yiu, 2002; dos Santos *et al.*, 2019; Eadie and Baylor, 2006; Iwarsson and Petersen, 2012). Although the training approaches described in previous studies have not provided uniform evidence regarding the type of training, stimuli, and length of training that best support listeners' formation of a shared understanding (Walden and Khayumov, 2022), there is support for the use of anchor voices to further increase inter-rater agreement (Chan and Yiu, 2002). The listener training performed in the present study used example voices to illustrate the expected range of the new items indicated by the SLPs to be most challenging to assess consistently. More time being set aside for listener training would likely have provided a stronger basis for listeners to build a robust shared understanding of the items.

The evaluation of PAGE-V was conducted using natural voices and spontaneous speech. In previous studies, the type of stimuli presented in the listener training have been both natural voices (dos Santos *et al.*, 2019; Eadie and Baylor, 2006; Iwarsson and Petersen, 2012) and synthetic voices (e.g., Granqvist, 2003). The use of natural voices in the present study is argued to provide a more ecologically valid training, close to clinical voice assessments. However, the multi-dimensional nature of natural voices complicates for listeners to separate features to be rated individually (Kreiman *et al.*, 2007). Had the study used synthetic voices, it would have been possible to tune individual acoustic voice properties separately, without other aspects also being affected. For example, the frequencies of formants related to the oral resonance space (and affecting perceived *oral resonance)* could have been adjusted separately from the formants affected also by the size of the pharyngeal resonance space (and affecting perceived *pharyngeal resonance*). Example stimuli that differ in terms of either perceived o*ral resonance* or perceived *pharyngeal resonance* may thus be created and used in future listener training of these PAGE-V items.

Since spontaneous speech samples were used, we propose additionally that the material represent the individual speakers' intonation patterns, and other aspects related to speaking style, to a degree that scaffold the ability to transfer observations to clinical applications of the PAGE-V assessment tool. However, the varying linguistic content in speech samples may have increased the cognitive load and contributed to a lower inter-rater agreement. The use of standardised recording material consisting of sentences or text reading that the listeners are well-acquainted with may be advantageous when many voice and speech features are to be rated, especially in less experienced listeners. While inter-rater agreement among experienced listeners has been seen to be slightly better in spontaneous speech compared to read sentences, the opposite has been seen for SLPs less experienced in performing voice assessments (Alves *et al.*, 2024), who may find it easier if the linguistic content is held constant. Further, the use of a standardised material could facilitate the rating of speech features included in PAGE-V; adapting the material to ensure, for example, combinations of word-final and word-initial consonants would facilitate the assessment of articulatory transitions. In future evaluations of PAGE-V, speech samples should thus comprise both spontaneous and controlled speech, as standardised speech material may facilitate the listening task.

It is proposed that a revised listener training procedure may further strengthen the shared understanding of rated voice items among raters. In the present evaluation

of PAGE-V, the raters were not allowed to go back to previously rated voice samples, which may have inflated the level of inter-rater disagreement and reduced intra-rater consistency. The listening task can also be characterised as demanding when performed in concentrated sessions, which may further have influenced intra-rater consistency due to effects of fatigue. These design choices were, however, made (1) to ensure a base level of transfer to a clinical setting, where clients are assessed individually, and (2) in response to the need of assessing PAGE-V in a wide range of voices while providing otherwise identical contexts, in terms of instruction and listening environment, for all raters. Intra-rater consistency and inter-rater agreement have, however, been seen to improve with the use of the Visual Sort and Rate (VSR) method (Granqvist, 2003), in which training samples are sorted and ranked based on the perceived degree of a certain voice feature. With this method, listeners shift their references from their internal representations to external representations constituted by the other voice samples (Granqvist, 2003). Moreover, rating can be assumed to become easier if voice samples are first sorted (Granqvist, 2003). We did not use the VSR method in the current investigation, and future evaluations of PAGE-V should, therefore, consider raters being able to compare voice samples during the rating procedure. It is proposed that the VSR method could enhance raters' ability to benefit from the shared external references provided by the voice feature variety across samples to support increased intra-rater consistency and inter-rater agreement.

Previous research has indicated detailed perceptual assessments to be challenging due to the complex relationships and covariation among voice features making it hard for listeners to separate individual voice features (Kreiman *et al.*, 2007). On the other hand, a more detailed assessment distinguishing between related voice and speech features has been recommended for a first-visit assessment and relevant choice of training content (Iwarsson *et al.*, 2018). In line with Iwarsson *et al.* (2018), we argue that detailed assessments of individual features are necessary to provide a gender-affirming voice training that starts from the individual TGD client's current voice expression and targets voice and speech features relevant to the client's preferred gender expression. However, a detailed assessment of voice and speech features that may be closely related requires explicit definitions of individual items. Our findings suggest that revisiting item definitions as well as scale labels and appropriateness of scale granularity (Pearse, 2011) by the group of active practitioners will be the key to forming a strong consensus (Iwarsson and Petersen, 2012) to drive development going forward. The continued work on PAGE-V should, therefore, aim for a wider inclusion of expert listeners within the national network of SLPs (within the Swedish Association for Transgender Health, SPATH). We further suggest that smaller clinics, where extensive experience in providing gender-affirming voice training may not be available, require careful consideration and support to achieve reliable use of PAGE-V. We argue that a broader spectrum of practitioners should be included in consensus listener training sessions and joint discussions to build shared understanding of the PAGE-V items among SLPs in general.

Conclusions

The initial evaluation of PAGE-V provides support for the clinical relevance of standardised auditory-perceptual assessments of voice and speech features influencing

the perception of gender expression in voice and speech. A majority of items were reliably assessed by multiple raters, even with the limited training provided. Some items showed lower intra-rater consistency or inter-rater agreement. Listener training and consensus procedures involving joint listening and discussions about the included items are, therefore, needed to strengthen and equalise listeners' internal representations of the items. The results of the intra-rater consistency and inter-rater agreement analysis, and comments from the participating listener experts, support the need for further development and evaluation of PAGE-V.

References

Alves J DN, Almeida AAF, Yamasaki R and Lopes LW. 2024. The influence of listener experience, measurement scale and speech task on the reliability of auditory-perceptual evaluation of vocal quality. *Codas* 36(3), e20230175. https://doi.org/10.1590/2317-1782/20232023175

American Psychiatric Association (APA). 2022. Diagnostic and Statistical Manual of Mental Disorders, 5th ed., Text Revision (DSM-5-TR). Washington, DC: APA.

Avery JD and Liss JM. 1996. Acoustic characteristics of less-masculine-sounding male speech. *Journal of the Acoustic Society of America* 99(6), 3738–3748. https://doi.org/10.1121/1.414970

Azul D, Nygren U, Södersten M and Neuschaefer-Rube C. 2017. Transmasculine people's voice function: a review of the currently available evidence. *Journal of Voice* 31(2), 261 e269–261 e223. https://doi.org/10.1016/j.jvoice.2016.05.005

Chan KM and Yiu EM. 2002. The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research* 45(1), 111–126. https://doi.org/10.1044/1092-4388(2002/009)

Coleman E, Radix AE, Bouman WP, Brown GR, de Vries ALC, Deutsch MB, Ettner R, Fraser L, Goodman M, Green J, Hancock AB, Johnson TW, Karasic DH, Knudson GA, Leibowitz SF, Meyer-Bahlburg HFL, Monstrey SJ, Motmans J, Nahata L, . . . Arcelus J. 2022. Standards of care for the health of transgender and gender diverse people, version 8. *International Journal of Transgender Health* 23 (Supl.1), S1–S259. https://doi.org/10.1080/26895269.2022.2100644

Dacakis G, Oates J and Douglas J. 2012. Beyond voice: perceptions of gender in male-to-female transsexuals. *Current Opinion in Otolaryngology & Head and Neck Surgery* 20(3), 165–170. https://doi.org/10.1097/MOO.0b013e3283530f85

Dahl KL and Mahler LA. 2020. Acoustic features of transfeminine voices and perceptions of voice femininity. *Journal of Voice 34*(6), 961 e919–961 e926. https://doi.org/10.1016/j.jvoice.2019.05.012

Davies S and Goldberg, JM. 2006. Clinical aspects of transgender speech feminization and masculinization. *International Journal of Transgenderism* 9(3–4), 167–196. https://doi.org/10.1300/J485v09n03_08

Davies S, Papp VG and Antoni C. 2015. Voice and communication change for gender nonconforming individuals: giving voice to the person inside. *International Journal of Transgenderism* 16(3), 117–159. https://doi.org/10.1080/15532739.2015.1075931

De Bodt MS, Wuyts FL, Van de Heyning PH and Croux C. 1997. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice* 11(1), 74–80. https://doi.org/10.1016/s0892-1997(97)80026-4

Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier L and Millet B. 1998. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Revue de Laryngologie – Otologie – Rhinologie (Bordeaux)* 119(4), 247–248. https://www.ncbi.nlm.nih.gov/pubmed/9865100

dos Santos PCM, Vieira MN, Sansão JPH and Gama ACC. 2019. Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. *Journal of Voice* 33(2), 220–225. https://doi.org/https://doi.org/10.1016/j.jvoice.2017.10.020

Eadie TL and Baylor CR 2006. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice* 20(4), 527–544. https://doi.org/10.1016/j.jvoice.2005.08.007

Gerratt BR, Kreiman J, Antonanzas-Barroso N and Berke GS. 1993. Comparing internal and external standards in voice quality judgments. *Journal of Speech, Language, and Hearing Research* 36(1), 14–20. https://doi.org/10.1044/jshr.3601.14

Graham M, Milanowski A and Miller J. 2012. Measuring and promoting inter-rateragreement of teacher and principal performance ratings. Available at: https://files.eric.ed.gov/fulltext/ ED532068.pdf (Accessed: 6 Dec 2024).

Granqvist S. 2003. The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology* 28(3), 109–116. https://doi.org/10.1080/14015430310015255

Hammarberg B. 2000. Voice research and clinical needs. *Folia Phoniatrica et Logopaedica* 52(1–3), 93–102. https://doi.org/10.1159/000021517

Hammarberg B, Fritzell B, Gauffin J, Sundberg J and Wedin L. 1980. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica* 90(5–6), 441–451. https://doi.org/10.3109/00016488009131746

Helou LB, Solomon NP, Henry LR, Coppit GL, Howard RS and Stojadinovic A. 2010. The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V) ratings of post-thyroidectomy voice. *American Journal of Speech-Language Pathology* 19(3), 248–258. https://doi.org/10.1044/1058-0360(2010/09-0012)

Henton C. 1995. Pitch dynamism in female and male speech. *Language & Communication* 15(1), 43–61. https://doi.org/10.1016/0271-5309(94)00011-Z

Hillman RE, Stepp CE, Van Stan JH, Zanartu M and Mehta DD. 2020. An updated theoretical framework for vocal hyperfunction. *American Journal of Speech-Language Pathology*, 29(4), 2254–2260. https://doi.org/10.1044/2020_AJSLP-20-00104

Hirano M. 1981. Clinical Examination of Voice. Cham: Springer.

Hirsch S, Pausewang Gelfer M and Boonin J. 2019. The art and science of resonance, articulation, and volume. In Adler RK, Hirsch S and Pickering J (eds.), Voice and Communication Therapy for the Transgender/Gender Diverse Client: A Comprehensive Clinical Guide, 3rd ed. (pp 217–248). Queensland: Plural Publishing.

Holmberg J, Linander I, Södersten M and Karlsson F. 2023. Exploring motives and perceived barriers for voice modification: the views of transgender and gender-diverse voice clients. *Journal of Speech, Language, and Hearing Research* 66(7), 2246–2259. https://doi.org/10.1044/2023_JSLHR-23-00042

Holmberg J, Södersten M, Linander I and Nylen F. 2024. Perception of femininity and masculinity in voices as rated by transgender and gender diverse people, professional speech and language pathologists, and cisgender naive listeners. *Journal of Voice*. Advance online publication. https://doi.org/10.1016/j.jvoice.2024.07.034

Iwarsson J, Bingen-Jakobsen A, Johansen DS, Kolle IE, Pedersen SG, Thorsen SL and Petersen NR. 2018. Auditory-perceptual evaluation of dysphonia: a comparison between narrow and broad terminology systems. *Journal of Voice* 32(4), 428–436. https://doi.org/10.1016/j.jvoice.2017.07.006

Iwarsson J and Petersen NR. 2012. Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice* 26(3), 304–312. https://doi.org/10.1016/j.jvoice.2011.06.003

Kelly V, Hertegard S, Eriksson J, Nygren U. and Södersten M. 2019. Effects of gender-confirming pitch-raising surgery in transgender women a long-term follow-up study of acoustic and patient-reported data. *Journal of Voice* 33(5), 781–791. https://doi.org/10.1016/j.jvoice.2018.03.005

Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J and Hillman RE. 2009. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology* 18(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Koo TK and Li MY. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kreiman J, Gerratt BR and Ito M. 2007. When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America* 122(4), 2354–2364. https://doi.org/10.1121/1.2770547

Kreiman J, Gerratt BR, Kempster, GB, Erman A and Berke GS. 1993. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research* 36(1), 21–40. https://doi.org/10.1044/jshr.3601.21

Kreiman J, Gerratt BR and Precoda K. 1990. Listener experience and perception of voice quality. *Journal of Speech, Language, and Hearing Research* 33(1), 103–115. https://doi.org/10.1044/jshr.3301.103

Leung Y, Oates J and Chan SP. 2018. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: a systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research* 61(2), 266–297. https://doi.org/10.1044/2017_JSLHR-S-17-0067

Leyns C, Alighieri C, De Wilde J, Van Lierde K, T'Sjoen G and D'haeseleer E. 2022. Experiences of transgender women with speech feminization training: a qualitative study. *Healthcare* 10(2295). https://doi.org/10.3390/healthcare10112295

McHugh ML. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22(3), 276–282. https://doi.org/10.11613/bm.2012.031

Nygren U, Nordenskjold A, Arver S and Södersten M. 2016. Effects on voice fundamental frequency and satisfaction with voice in trans men during testosterone treatment—a longitudinal study. *Journal of Voice* 30(6), 766.e23–766.e34. https://doi.org/UNSP 766.e2310.1016/j.jvoice.2015.10.016

Nylen F, Holmberg J and Södersten M. 2024. Acoustic cues to femininity and masculinity in spontaneous speech. *Journal of the Acoustical Society of America* 155(5), 3090–3100. https://doi.org/10.1121/10.0025932

Oates J and Dacakis G. 1997. Voice change in transsexuals. *Venerology* 10(3), 178–187.

Patel RR, Awan SN, Barkmeier-Kraemer J, Courey M, Deliyski D, Eadie T, Paul D, Svec JG and Hillman R. 2018. Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology* 27(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009

Pearse N. 2011. Deciding on the scale granularity of response categories of Likert-type scales: the case of a 21-point scale. *Electronic Journal of Business Research Methods* 9, 159–171.

Sandelowski M and Barroso J. 2003. Classifying the findings in qualitative studies. *Qualitative Health Research* 13(7), 905–923. https://doi.org/10.1177/1049732303253488

Sundberg J. 2001. Röstlära: Fakta om Rösten i Tal och Sång [Science of the Singing Voice], 3rd ed. Stockholm: Proprius Förlag.

Svec JG and Granqvist S. 2010. Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology* 19(4), 356–368. https://doi.org/10.1044/1058-0360(2010/09-0091)

Swedish Work Environment Authority. 2019. Buller (version 2.2) [mobile application] Google Play. https://play.google.com/store/apps

Tinsley HE and Weiss DJ. 1975. Interrater reliability and inter-rater agreement of subjective judgments. *Journal of Counseling Psychology* 22(4), 358–376. https://doi.org/10.1037/h0076640

Verdolini K, Druker DG, Palmer PM and Samawi H. 1998. Laryngeal adduction in resonant voice. *Journal of Voice* 12(3), 315–327. https://doi.org/10.1016/s0892-1997(98)80021-0

Walden PR and Khayumov J. 2022. The use of auditory-perceptual training as a research method: a summary review. *Journal of Voice* 36(3), 322–334. https://doi.org/10.1016/j.jvoice.2020.06.032

World Health Organisation (WHO). 1993. The ICD-10 Classification of Mental and Behavioural Disorders. Genève: WHO.

Ziltzer RS, Lett E, Su-Genyk P, Chambers T and Moayer R. 2023. Needs assessment of gender-affirming face, neck, and voice procedures and the role of gender dysphoria. *Otolaryngology–Head and Neck Surgery* 169(4), 906–916. https://doi.org/10.1002/ohn.329

Auditivt-perceptuell bedömning av rösten vid könsdysfori

Baserad på SVEA, Hammarberg 2006, modifierad och kompletterad

Röstkvalitet

Läckage
avsaknad — *hög grad av*

Hypofunktion
avsaknad — *hög grad av*

Hyperfunktion, press
avsaknad — *hög grad av*

Knarr
avsaknad — *hög grad av*

Flödig fonation
avsaknad — *hög grad av*

Instabilitet
☐ ingen instabilitet ☐ instabilt läge ☐ instabil klang ☐ registerbrott

Register
☐ modalregister ☐ falsettregister ☐ mellanregister ☐ går ej avgöra

Röstläge

☐ i enlighet med normen för kvinnoröster ☐ könsneutral ☐ i enlighet med normen för mansröster

mycket lågt — *mycket högt*

Röststyrka

mycket låg — *mycket hög*

Intonation

Variation i satsmelodi
monoton — *mycket varierad*

Intonationsmönster i frasslut
☐ främst fallande ☐ främst stigande ☐ främst jämn ☐ lika ofta fallande/stigande/jämn

Storlek på röstlägesvariationen i fallande/stigande frasslut
☐ liten ☐ måttlig ☐ stor röstlägesvariation

https://doi.org/10.58986/al.2025.41653

Resonans

Resonant röst
inte alls resonant — *mycket resonant*

Oral resonans
mycket mörk — *mycket ljus*

Faryngeal resonans
mycket mörk — *mycket ljus*

Tal

Artikulation
oprecis — *precis*

Artikulatoriska övergångar (transitioner mellan stavelser/ord)
separerade/markerade — *sammanbundna/jämna*

Taltempo
jämnt — *mycket varierat*

Övergripande bedömning av rösten

Femininitet i rösten
inte alls feminin — *mycket feminin*

Maskulinitet i rösten
inte alls maskulin — *mycket maskulin*

Tilläggsparametrar (som bedöms inverka feminiserande/maskuliniserande på röst och tal)

_ _ _ _ _ _ _ _ _ _ _
inte alls maskulint — *mycket maskulint*

_ _ _ _ _ _ _ _ _ _ _
inte alls feminint — *mycket feminint*

_ _ _ _ _ _ _ _ _ _ _