Acta Logopaedica (2025), Vol. 2 | 58–78
DOI: 10.58986/al.2025.41881

Acta Logopaedica

ORIGINAL RESEARCH

Students' use of e-learning in developing auditory-perceptual skills in speech language pathology

Sofia Strömbergsson[1,2*], Maria Södersten[1,3], Anette Lohmander[1]

[1]Department of Clinical Science, Intervention and Technology (CLINTEC), Division of Speech and Language Pathology, Karolinska Institutet, Stockholm, Sweden; [2]Department of Neurology, Division of Speech and Language Pathology, Danderyd Hospital, Stockholm, Sweden; [3]Speech and Language Pathology, Allied Health Professionals Medical Unit, Karolinska University Hospital, Stockholm, Sweden

***Corresponding author:** Sofia Strömbergsson, Department of Clinical Science, Intervention and Technology (CLINTEC), Division of Speech and Language Pathology, Karolinska Institutet, Huddinge F67, 141 86 Stockholm, Sweden. Email: sofia.strombergsson@ki.se

Publication date: 1 August 2025

Abstract

Auditory-perceptual assessment is central in speech-language pathology (SLP) practice, and hence an important target in SLP education. Self-supervised e-learning may serve as a means to achieve sufficient practice. The web-based platform, Practical education Using Multimedia Application (PUMA) has been developed to this end. The aim of this study was to evaluate students' use of e-learning via PUMA, potential improvement in their assessment performance, and their experiences of using PUMA. Students performed perceptual ratings at start (n = 30) and end (n = 31) of the course, and in between, they participated in course activities and self-supervised e-learning, where their activity was logged. By the end of the course, a higher proportion of student ratings were 'expert-like', indicating improved performance. However, as time logs showed low e-learning activity, improved performance was likely explained by other activities. This suggests that for self-supervised learning to take place, it needs to be integrated into the course.

Keywords: voice disorder; assessment skills; training; e-learning

Abstract

Auditiv-perceptuell bedömning är central inom logopedisk praktik och därmed ett viktigt mål inom logopedutbildning. Självstyrt e-lärande kan fungera som ett sätt att uppnå tillräcklig övning. Den webbaserade plattformen Practical education Using Multimedia Application (PUMA) har utvecklats just för detta syfte. Målet med projektet som här

beskrivs är att utvärdera studenters användning av e-lärande via PUMA, eventuell förbättring i deras bedömningsförmåga samt deras erfarenheter av att använda PUMA. Studenter registrerade på en kurs om röststörningar genomförde perceptuella bedömningar i början (n = 30) och slutet (n = 31) av kursen, och däremellan deltog de i kursaktiviteter och självstyrt e-lärande, där deras aktivitet loggades. Vid kursens slut var en större andel av studenternas bedömningar "expertlika", vilket indikerar förbättrad prestation. Eftersom tidsloggarna visade låg aktivitet i e-lärandet, förklaras dock förbättringen troligen av andra aktiviteter än av aktivitet i PUMA. Detta tyder på att för att självstyrt lärande faktiskt ska bli av, behöver det integreras i kursen.

Keywords: röststörningar; bedömningsfärdighet; färdighetsträning; e-lärande

Background

Perceptual assessment is central to the clinical practice of speech-language pathologists. In areas such as speech, voice, and swallowing disorders, it underpins medical decisions concerning diagnostics and intervention. Therefore, ensuring assessment accuracy and reliability is crucial for safe care. Acquiring perceptual assessment skills requires practice, making the training of these skills an important component of speech-language pathology (SLP) education. E-learning has been suggested as a means to provide students with opportunities to train perceptual assessment skills in some SLP areas, such as cleft palate speech (Bruneel *et al.*, 2022; Lohmander *et al.*, 2021). However, e-learning in educational training of perceptual assessment in other areas, such as voice disorders, has received less attention.

Across patient areas, SLP education typically includes both theoretical knowledge concerning the nature and clinical care of different disorders, and perceptual skills training to identify perceptual traits that signal different types and degrees of disorders. Ideally, by the end of a course, students should have acquired skills to make reliable perceptual assessments at a basic level. However, the subjective nature of perceptual assessment poses a threat to the reliability of ratings. Within the area of voice disorders, for example, reliability of auditory-perceptual assessment has been a concern for many decades (Kent, 1996; Kreiman *et al.*, 1993; Oates, 2009). One important factor in achieving reliable assessments is listener training (Hammarberg *et al.*, 1980; Iwarsson and Reinholt Petersen, 2012; Lee *et al.*, 2009).

Auditory-perceptual assessment requires listeners to compare what they hear to an internal standard (Goldstone, 1998; Kreiman *et al.*, 1993). When evaluating the degree of breathiness in a voice, for example, the listener must first learn to identify the parameter (*i.e.* 'breathiness') and then develop an internal standard of what constitutes high and low degrees of this parameter. Listeners with little or no experience of attending to such dimensions of voice have less robust internal standards than more experienced listeners. As such, training in auditory-perceptual assessment can be seen as a process of shaping reliable internal standards (Kreiman *et al.*, 1993). In their review of auditory-perceptual training procedures used in education and research, Walden and Khayumov (2022) identified commonly employed training components. The most frequently used training components were multiple exposures to rating and utilisation of external references (or 'anchor stimuli').

In training based on multiple exposures to rating, the authors further noted that practice or group consensus was frequently used. Practice, in the form of repeated rating, is often emphasised as important for stabilising listeners' internal standards, especially if feedback on listeners' ratings is provided (Eadie and Baylor, 2006). Consensus procedures are collaborative processes, enabling listeners to calibrate against other listeners. For example, consensus discussions are an important ingredient in the Stockholm Voice Evaluation Approach (SVEA; Hammarberg, 2000; Hammarberg and Gauffin, 1995), which involves group discussions regarding terminology and perceptual ratings, with the aim of achieving a shared understanding of concepts and consensus in ratings.

Regarding educational activities aimed at providing training in auditory-perceptual assessment of voice, published descriptions are rare. Of the 36 studies included in Walden and Khayumov (2022), only four represent descriptions of SLP education course activities. Nevertheless, a closer look at these provides insights into how the perceptual training strategies mentioned above have been used in the education of SLP students. Iwarsson and Reinholt Petersen (2012) describe educational activities in an advanced-level course aimed at training SLP students in auditory-perceptual assessment of voice. Inspired by SVEA, the course training included, for each of 10 voice quality parameters, its definition, underlying physiology, presentation of selected voice samples representing different degrees of severity, group discussion, and practical exercises imitating the voice qualities. The duration of the consensus training was 20 hours, after which the students conducted individual training and homework. Another educational programme focusing on multiple voice parameters is described by Silva and colleagues (2012). Here, too, training sessions focused on one voice parameter at a time – its definition and acoustic presentation in voice samples representing different degrees of the parameter in focus. However, instead of being included in the same course, the nine training sessions were spread out across two semesters, carried out during the last 15–30 minutes of class lectures. It is not clear from the description what type of feedback was provided or whether the students also conducted individual training at home. More condensed educational training programmes are described in Chan and Yiu (2002) and Eadie and Baylor (2006), both aimed at advancing students' perceptual assessment of three selected voice parameters. Here, SLP students were engaged in two 30–60 minute sessions of listener training. The sessions included a presentation of the definition of the voice parameter in focus and the presentation of anchor stimuli selected to represent different degrees of that voice parameter. In both training programmes, the students first rated samples individually and were then provided with suggested expert ratings as feedback. Eadie and Baylor (2006) also included a consensus element, where students compared and discussed their ratings with peers. Taken together, these examples illustrate variation in auditory-perceptual training strategies used in SLP education. Multiple exposures to rating were employed in all of them; this is not surprising, given its central role in auditory-perceptual training in general (Walden and Khayumov, 2022). Apart from that, activities varied quite extensively in terms of training duration, training scope (*i.e.* targeting few or multiple voice parameters), and whether practice or consensus, or both, was included.

Regarding the effect of the auditory-perceptual training, three of the four educational programmes for auditory-perceptual training of voice assessment described

above observed improved rating performance in the students post-training, often reflected as a more consistent use of the rating scale within an individual student and/or as higher agreement in the student group (Chan and Yiu, 2002; Eadie and Baylor, 2006; Iwarsson and Reinholt Petersen, 2012). In one case, no sign of improvement was observed (Silva *et al.*, 2012). Interestingly, where improvements were observed, those were often more pronounced for some voice parameters than others. For example, Eadie and Baylor (2006) observed a steeper increase in rating reliability for the parameters 'roughness' and 'breathiness', compared to the parameter 'overall severity', where relatively high reliability was observed already before training. From the perspective of evaluating whether students' internal standards have been stabilised and/or equalised, it makes sense to evaluate students' rating performance with respect to intra- and inter-rater reliability. From an educational perspective, however, the initial goal is not necessarily to achieve high inter- or intra-reliability. After all, consistency in ratings is of little worth unless the ratings also align with what is clinically relevant. Typically, the expertise to perceptually assess clinically relevant traits lies with experienced clinicians.

Given the goal of SLP education to educate new speech-language pathologists, one may argue that the intended learning outcome in perceptual training is for students to 'rate like an expert'. From this perspective, it seems reasonable to evaluate student performance with reference to expert clinicians; the more 'expert-like' the student ratings are, the better their performance. Such evaluation needs to account for the rating variability that exists also among experts; there is rarely a single correct rating, but rather a range within which expert ratings fall (Holmberg *et al.*, 2001). This is especially true when ratings are based on a scale of high granularity (*i.e.* with many scale steps) or a visual analogue scale (Eadie and Baylor, 2006; Holmberg *et al.*, 2001), and when multiple dimensions are rated, rather than a few (Iwarsson *et al.*, 2018). To reflect the wide variety of voice qualities that are encountered in clinical practice, however, a multiple dimension approach in voice assessment is difficult to escape.

In terms of course activities, education on voice disorders typically includes a variety of synchronous on-campus activities, such as lectures, seminars, workshops, and laboratory work. Achieving multiple exposures to rating can be difficult without schedule overload. Asynchronous e-learning may potentially relieve these demands by allowing students to practice at times and places of their own choosing (Kimura *et al.*, 2023; Lawn *et al.*, 2017; Lohmander *et al.*, 2021). Still, few descriptions of e-learning in the area of voice disorders exist. A pioneering example was the interactive multi-media package 'A Sound Judgement' (Oates and Russell, 1998). Here, students were presented with recordings of authentic patients, their case histories, video-recorded interviews, and laryngoscopy findings. In addition, students could make perceptual ratings of audio recordings of patient voices and receive feedback on their ratings from a virtual expert clinical supervisor. Although an informal evaluation of students' experiences of using the package showed positive results, no evaluation of the students' assessment skills was presented (Oates and Russell, 1998). Around the same time, another multimedia platform, 'The Project for the Development of Multimedia Methods in Logopedics and Phoniatrics' (PUMP) was developed (Lyberg Åhlander *et al.*, 1999), also containing recordings and information from patients with voice disorders. Although preliminary evaluation of student

experiences was positive (Lyberg Åhlander *et al.*, 1999), we have not found any more detailed descriptions or subsequent evaluations of PUMP. Apart from the custom-made quiz mentioned in passing in Iwarsson and Reinholt Petersen (2012), we are not aware of any later cases where e-learning has been used in education on voice disorders.

Practical education Using Multimedia Application (PUMA) is a website developed at the University of Gothenburg and at Karolinska Institutet to provide students with opportunities to practice perceptual assessment (Lohmander *et al.*, 2021). Through collaboration across SLP education programmes in Sweden and Swedish-speaking Finland, PUMA also serves to foster national uniformity in documentation and analysis (Lohmander *et al.*, 2021). Initially focused on the assessment of cleft palate speech, PUMA has expanded to cover other patient areas, such as neuromotor speech and swallowing disorders. An evaluation of e-learning through PUMA has shown significant gains for students' perceptual skills in evaluating speech function and deviances associated with cleft palate (Lohmander *et al.*, 2021). As PUMA expands to new areas, evaluation continues in parallel. One of the recent additions is the area of voice disorders, where the need for evaluation serves as rationale for the present project.

Aim and research questions

The aim of the project was to evaluate the result of self-supervised training with the e-learning tool PUMA on students' auditory-perceptual voice assessment performance, whether results could be linked to time spent on training, and how the students perceived e-learning in PUMA. Specifically, the research questions were as follows:

1. Does students' auditory-perceptual voice assessment performance improve after self-supervised training in PUMA? This question is operationalized by comparing the proportion of 'expert-like' ratings before and after training.
2. Is there an association between students' potential improvement in assessment performance, and their time spent with self-supervised training in PUMA?
3. What are the students' experiences from working with PUMA?

Materials and method

Evaluation of auditory-perceptual skills training

The evaluation was conducted in the fall of 2023, during the 4-week course module Voice disorders, which runs during the 5th semester at the Study Programme in Speech and Language Pathology at Karolinska Institutet. The evaluation involved a pre-test and a post-test, to track the students' performance in voice assessment at the beginning and end of the course module (see Figure 1).

Informants

Thirty-two students participated in the course module. Thirty of them participated in the pre-test and 32 in the post-test. Due to technical error at post-test, however,

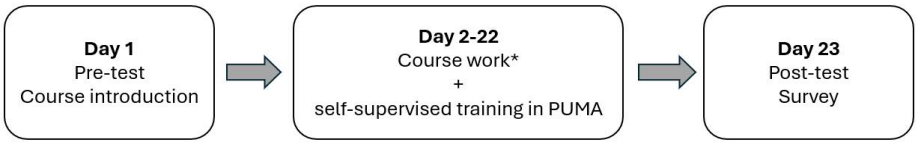

Figure 1. Flow chart of the steps in the evaluation of the learning impact on students' performance in auditory-perceptual voice assessment.

the results for one of the students were corrupted and therefore excluded. Twenty-nine students responded to a post-course evaluation survey.

Materials for pre- and post-test

The perceptual task at pre- and post-test was based on seven recorded voice samples and a digital evaluation form for reporting auditory-perceptual ratings for each sample. The voice samples were approximately half-minute long recordings of adult patients reading a standard passage. The samples represented different types of voice disorders and voice qualities, selected to reflect variability in voice characteristics. For example, the recorded voice samples varied regarding parameters such as breathiness, vocal fry/creakiness, and aphonia.

The evaluation form was a digital version of the SVEA form (Hammarberg, 1986, 2000), implemented in PUMA (Figure 2). The SVEA consists of 11 voice quality parameters rated on a visual analogue scale (VAS) with the endpoints to the left 'not at all' (Swe. 'Avsaknad av') and to the right 'high degree of' (Swe. 'Hög grad av'). Pitch and loudness are judged on 200-mm lines with 0 in the middle representing normality. Vocal registers are rated using three categories (modal, falsetto, and 'cannot judge'). For evaluation in the present study, only the11 VAS-rated parameters were included in the analysis.

The seven voice samples were evaluated by a group of expert raters; eight SLPs who had been working as voice clinicians for 3–40 years. The expert raters evaluated the voice samples using SVEA (in its original analog version) during a shared listening session. The session was introduced by a reminder of the definitions of the voice parameters in SVEA, followed by shared listening to two voice samples (not included in the present study), where the raters discussed their ratings for consensus as suggested by Hammarberg (2000). After this introduction, they started their evaluation individually of the seven voice samples included in the present study. The sound files were played from a computer and a Fostex 6301ND loudspeaker and the listeners used paper and pen for their individual ratings. The samples were played and replayed until all were satisfied. Figure 3 illustrates the distribution of expert ratings across the seven voice samples, each reflecting a graphical voice 'profile'.

Training

The 4-week course module included a variety of educational activities: lectures, seminars, laboratory work, and self-supervised training in PUMA. None of these

https://doi.org/10.58986/al.2025.41881

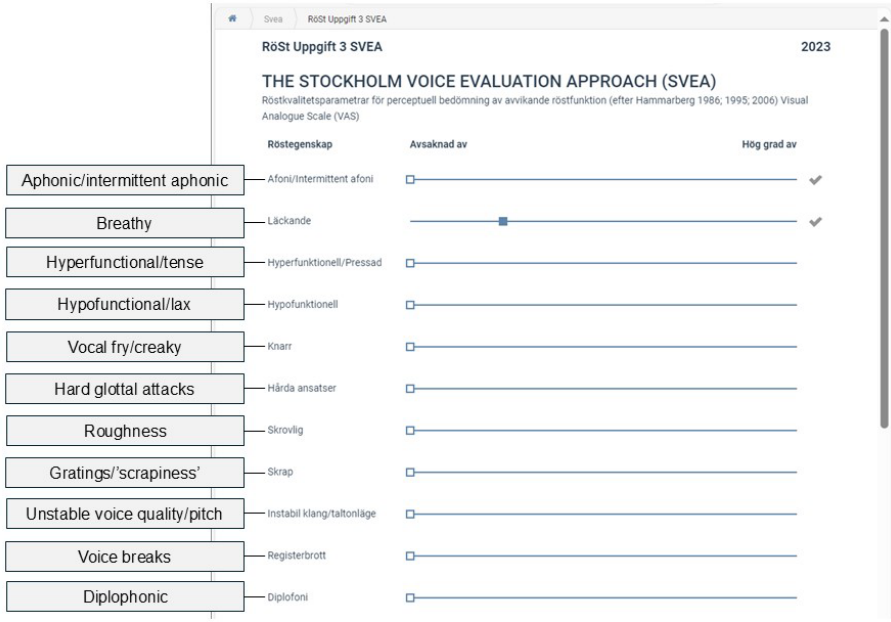

Figure 2. The digital Stockholm Voice Evaluation Approach (SVEA) auditory-perceptual voice assessment form, in Swedish, with English translations of voice parameters. The figure shows all 11 parameters included in the analysis. The figure as it appears shows the status when a user has rated two parameters (as indicated by the check marks on the right), and none of the others. Please note the difference between an intentional 0 rating for the first voice parameter (aphonic/intermittent aphonic), and the default 0 rating for parameters 3–11, where rating decisions have not yet been made; this is commented in 'Analysis.'

activities included the voice samples evaluated at pre- and post-test. During the course, the students were introduced to audio-perceptual voice assessment, including theoretical definitions of the voice quality parameters in SVEA as well as practical experience with listening to examples of these parameters. In an interactive workshop, they evaluated different voice samples using the SVEA protocol and compared and discussed their individual ratings with each other, with teacher support and with reference to expert ratings. Hence, the workshop included elements of both consensus and individual practice. For the workshop, voice samples were selected to illustrate that an audio-perceptual analysis can detect important changes in voices after intervention, assuming this would motivate the students to practice auditory-perceptual assessment using PUMA. The exercises in PUMA were, however, not formally presented or introduced to the students, as they were expected to carry out the training on their own.

Three different types of exercises were available in PUMA for the students to conduct whenever they wished and as many times as they wanted:

- *Identifying the most salient voice feature.* This task was designed as a multiple-choice task, where the students were presented with voice samples and asked to select the most salient of three voice features (*e.g.* hyperfunction, breathiness,

https://doi.org/10.58986/al.2025.41881 Acta Logopaedica is published according to open access under a Creative Commons CC-BY 4.0 license

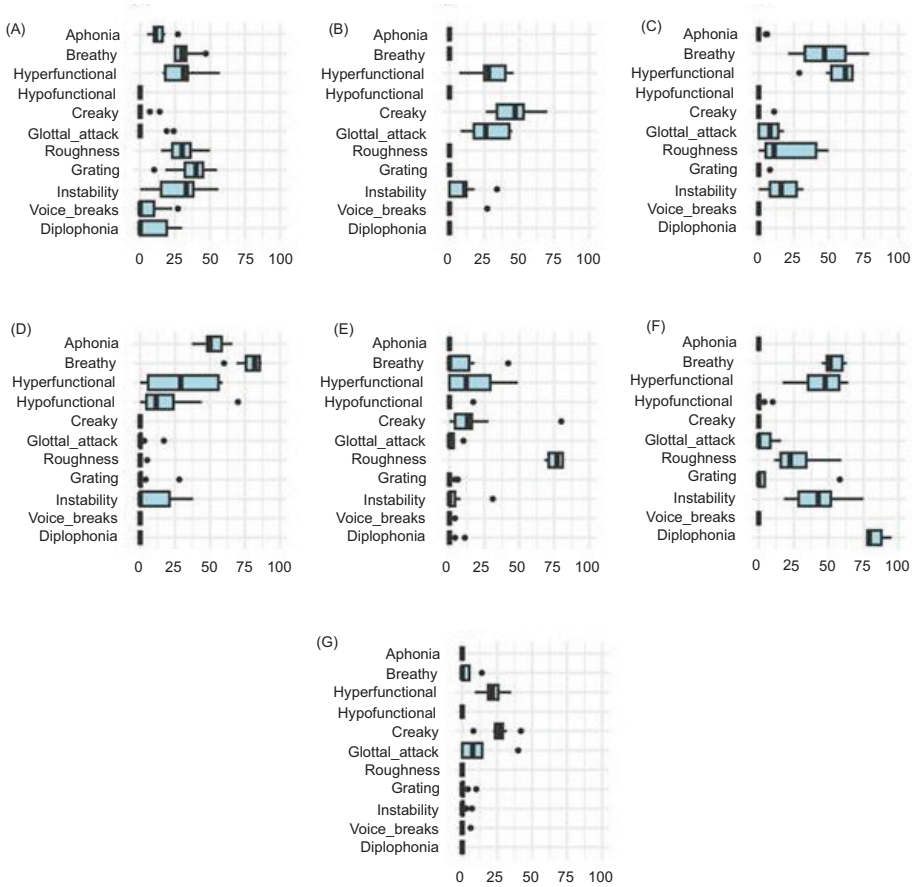

Figure 3. The distribution of expert evaluations of the seven voice samples A–G regarding the 11 rated voice parameters in SVEA, each reflected as a graphical voice 'profile'. Median values are represented as black lines within the boxes; boxes extend between 25th and 75th percentile, and lines outside boxes extend to the most extreme values within 1.5 interquartile range from 25th to 75th percentile. Dots represent outliers.

vocal fry). The task included 19 voice samples, and the students received immediate feedback on whether their response was correct. A screenshot illustrating this exercise is available as Figure S1.

- *Visual sort and rate (VSR)*. VSR is a method for perceptual rating developed by Granqvist (2003) which was implemented in PUMA. The task for the students was to focus on one voice parameter at a time (*e.g.* hyperfunction, breathiness, or instability) and visually sort voice samples in relation to each other with respect to an arbitrary scale represented by endpoints 'no or none' to 'a high degree of' that voice parameter. There was no feedback available for this task. Therefore, the students were encouraged to do it together with peers so that they could compare their ratings. A screenshot illustrating this exercise is available as Figure S2.

https://doi.org/10.58986/al.2025.41881

- *Voice assessment using the SVEA protocol.* Five voice samples were available, along with information about the diagnoses of voice disorders. The task was, for each voice sample, to provide an auditory-perceptual assessment with regards to the SVEA protocol. Expert ratings were available for reference, allowing the students to evaluate their own ratings.

The students' activity in the self-supervised training was logged (see 'Analysis').

Procedure

The pre- and post-test sessions followed the same procedure. On both occasions, the test sessions were conducted in two computer rooms at campus, with half of the students placed in each room. The students were seated by individual computers with some distance between each other and fitted with headphones (Sony MDR-ZX660AP). To participate in the test, the students had to log on to their computer with their personal details. They received written instructions on how to do the test of seven voices (as described in 'Materials'). A technician or a teacher was present in each room to answer questions about the procedure. After completing their assessment of one voice, the students were instructed to tick a box ('share with teacher') to make the assessment available for analysis. For technical reasons, three SVEA forms were missing in the pre-test, which resulted in a data loss of three SVEA forms (leaving with 207 forms). For the post-test, technical issues were resolved, and all students' evaluations of all seven samples were included in the analysis (adding up to 217 forms). At the post-test session, after completing the test and before leaving the lecture hall, the students were asked to fill out a digital post-course evaluation survey.

Post-course evaluation survey

The digital post-course evaluation survey contained six questions. The first two questions concerned the students' views regarding whether the exercises in PUMA had contributed to their development of perceptual skills ('*The exercises in PUMA have contributed to my development of perceptual voice analysis skills*, and *I feel well-prepared for conducting perceptual assessment of voice in patients during clinical practicum*'). Responses to these questions were indicated as a rating from 0 ('*To a very little extent*') to 5 ('*To a very large extent*'). For the third question, the students were asked to rank the three exercises with respect to perceived difficulty, and for question four, to elaborate their response in free text. For question five and six, the students were asked to provide free-text responses concerning what they liked most about PUMA-Voice ('*What was best with PUMA-Voice?*') and provide suggestions for improvement ('*How can the exercises in PUMA-Voice be improved?*'). The full post-course evaluation survey, in the original Swedish version with English translations, is available as Table S1.

Analysis

The students' responses to 11 VAS items in the SVEA forms were included in the analysis of rating accuracy. Rating responses were documented as a number between 0 and 100.

For the pre-test, it was possible for respondents to leave items without response. Unfortunately, there was no separate documentation of (a) non-responses, and (b) responses intentionally left at the default marker position at the zero end of the scale. (For an illustration, see Figure 2, where a zero rating has been provided for the first parameter, but where no ratings have yet been provided for the parameters 3–11. The tick mark on the right indicates an active response; however, this tick mark was not yet implemented at the pre-test.) Instead, despite being ambiguous, all such responses were documented as non-responses. This issue was addressed before the post-test, such that active responses were required for all items.

Student ratings were categorised as 'expert-like' if they fitted within the range of the ratings provided by the expert group. The handling of non-responses at the pre-test required careful consideration. To avoid over-classification of non-responses as 'non-expert-like', non-responses left at the default marker position at zero were interpreted as intentional 0 ratings, and classified as 'expert-like' if (at least) one of the expert ratings was 0. Although this decision entails a risk of over-classifying non-responses as 'expert-like', it was motivated by an effort not to exaggerate a potential increase in performance between pre- and post-test. For the examination of potential change between pre- and post-test, two separate analyses were conducted; one with all responses totaled across rating parameters, and another split up per rating parameter. The latter was motivated by an educational interest in revealing whether some voice parameters are more difficult to learn than others. χ^2-analyses were calculated to explore potential differences in the distribution of 'expert-like' and 'non-expert-like' ratings at pre- and post-test.

For the analysis of students' time spent in PUMA, usage statistics were extracted for all students from 25 September 2023 at 4 PM (*i.e.* end of the pre-test day) to 18 October 2023 at 9 AM (*i.e.* starting time of the post-test). The usage logs contained time stamps for a user entering a new page; however, there was no logged information concerning the user's activity on each page. As an approximation of the time spent on voice-related PUMA activities, therefore, the following principles were applied:

- Time stamps where the duration since the preceding time stamp was <15 minutes were handled as a continuation of the same activity session.
- Time stamps that were not followed by a new time stamp within the 15-minute frame were assigned a duration of 30 seconds.

The motivation for this conservative operationalisation (where the assigned duration of 30 seconds assumes very little activity) was an effort to minimise the risk of assuming activity when there was none. Duration of a user's activity on voice-related pages in PUMA was summed to a total duration (in minutes).

The students' responses to the post-course evaluation survey were analysed descriptively.

Ethical considerations

The PUMA website is hosted on a server with the highest level of security, with RapidSSL certificate. Login credentials to PUMA are granted to SLP students and to

practicing clinicians after signing an agreement regarding the restricted and secure use of PUMA. Individuals who have contributed to PUMA with audio and/or video recordings have provided their written informed consent regarding the use of the material in education. The streaming is encrypted, and files cannot be downloaded to personal computers. The handling of all personal data follows the Personal Data Act/General Data Protection Regulation. As for the students participating in the course described, no personal data was collected for the purpose of the study.

Results

Evaluation of skills training

As described above, settings at pre-test allowed informants to leave items unrated, that is, without confirmation that a default 0 rating was an active decision. Of all 2,277 responses (207 forms × 11 parameters) at pre-test, 1,566 (69%) were 'real' responses whereas 711 were non-responses. Notably, non-responses were not equally distributed across all students; 12 students provided fewer than two non-responses in total whereas equally manyprovided 40 non-responses or more. For the post-test, technical issues were resolved, so responses were required. Hence, the number of 'real' ratings at post-test was consistently 217 responses per rating parameter, that is, 2,387 rating responses in total.

Overall, there was a higher proportion of 'expert-like' ratings at the post-test (68.7%), compared to that at the pre-test (56.8%). This was statistically confirmed by χ^2 analysis: $\chi^2(2) = 70.72$, $p < 0.001$. Figure 4 illustrates the distribution of 'expert-like' and 'non-expert-like' ratings at pre- and post-test across 11 rating parameters. Collectively, the figure confirms the pattern of 'expert-like' ratings being more frequent at post-test, compared to pre-test. However, the pattern is more pronounced for some rated parameters than for others. For example, for the parameters Vocal fry/Creaky, Instability, and Gratings, the difference is quite pronounced whereas for the parameter Roughness, there is no difference between pre- and post-test. As shown in Table 1, this pattern was confirmed statistically. One can further note that for some parameters (*e.g.* Hard glottal attack, Instability, and Voice_breaks), many student ratings were 'expert-like' already at pre-test. For more details concerning variation in ratings (*e.g.* the observation of a wider spread in student ratings by course start, compared to by course end), see Figure S3.

Association between user activity and change in performance

Charting of students' self-paced activity in PUMA revealed substantial variation (see Figure 5). Whereas five students spent over 50 minutes in PUMA, eight students spent less than 5 minutes in PUMA. As Figure 5 illustrates, there was no association between students' time spent in PUMA and the change in performance from pre- to post-test. A closer inspection of the distribution of activity across different parts of the PUMA webpage further revealed that 37% of the summed student activity (262 out of 700 minutes) was spent on the PUMA-Voice landing page, 21% (147 minutes) on the SVEA exercise, 10% (70 minutes) on the VSR exercise, and 6% (43 minutes) on the multiple-choice exercise. The remaining 26% activity was

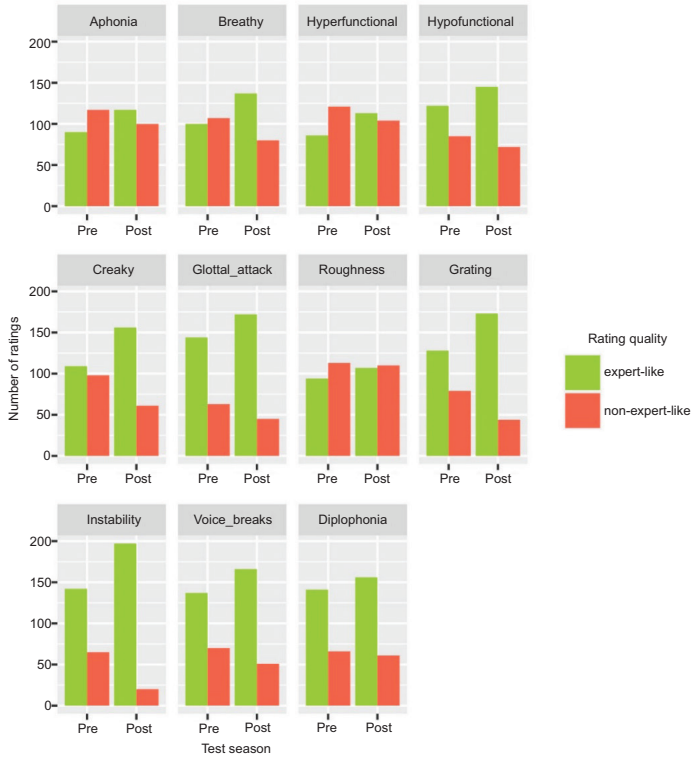

Figure 4. Number of responses categorised as 'expert-like' and 'non-expert-like' across the 11 rated parameters in SVEA from students' ratings at pre-test (n = 207) and post-test (n = 217).

Table 1. χ^2-statistics for the distribution of ratings as 'expert-like'/'non-expert-like' at pre-test (n = 207 per parameter) and post-test (n = 217 per parameter).

Rated parameter	χ^2	*df*	*p*
Aphonia	4.91	3	0.18
Breathy	9.73	3	0.02
Hyperfunction	5.00	3	0.17
Hypofunction	3.11	3	0.37
Vocal fry/creaky	17.01	3	$p < 0.01$
Glottal_attacks	5.54	3	0.13
Roughness	0.94	3	0.82
Gratings	16.76	3	$p < 0.01$
Instability	32.83	3	$p < 0.01$
Voice_breaks	5.83	3	0.12
Diplophonia	0.90	3	0.83

https://doi.org/10.58986/al.2025.41881

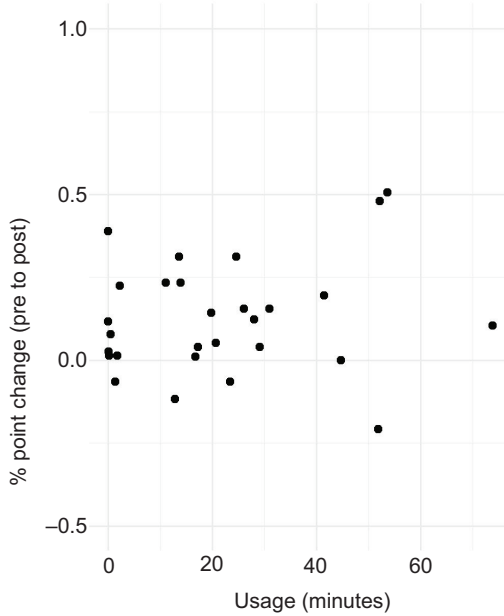

Figure 5. The association between change in rating performance between pre- and post-test, and the approximated time (in minutes) spent in PUMA. Change in performance from pre- to post-test is expressed as the difference in the proportion of expert-like ratings at pre- and post-test in percentage points.

distributed across other voice disorder-related pages, including five case reports (13%; 89 minutes) and a page containing a video-recorded fiberoptic laryngoscopy (4%; 27 minutes). In sum, it appears likely that other factors than the students' self-supervised work in PUMA contributed to the groups' improved performance at post-test.

Students' experiences of using PUMA

Twenty-nine students responded to the post-course evaluation questionnaire. The majority responded that the exercises in PUMA-Voice had contributed to their development of perceptual skills to a high (n = 12) or a very high (n = 6) extent whereas the remaining 11 students responded that PUMA-Voice had only contributed partly. Twenty-one (72%) students felt only partly prepared to conduct perceptual voice assessment in clinical practicum, and five of them to a small (n = 4) or very small (n = 1) extent. The VSR and SVEA exercises were ranked as more difficult than the multiple-choice task Select, the most salient voice characteristic. In their free-text responses concerning what they appreciated the most with PUMA-Voice, many highlighted the accessibility to real-life examples for self-guided practice. As for suggested improvements, the majority suggested the addition of a key reference for the VSR task to provide the students with feedback on their ranking of samples. Some also suggested that practice with PUMA-Voice should have been scheduled to enforce more practice.

https://doi.org/10.58986/al.2025.41881

Discussion

The present study aimed to evaluate students' use of self-paced e-learning on their performance in auditory-perceptual assessment. Improved assessment performance was indeed observed at the end of the course, compared to the start. However, no association was found with the students' time spent on self-learning e-learning activities, indicating that their increased performance was driven by other factors, such as in-class educational activities and self-studies outside of the e-learning platform. This pattern reflects the experience in around one-third of the students that the e-learning activities had only partly contributed to their development of perceptual skills. It should be noted, however, that the students' usage of PUMA varied extensively, with a few of them spending over 50 minutes on self-learning activities, and 25% spending less than 5 minutes. In addition, students' activity in PUMA was not primarily devoted to exercises specifically designed for auditory-perceptual training but also included other materials on the webpage. The results suggest several routes for development, both within the e-learning platform and in terms of general course development, which are discussed here.

From a general perspective, the improved assessment performance at post-test resembles similar positive results reported previously (Chan and Yiu, 2002; Eadie and Baylor, 2006; Iwarsson and Reinholt Petersen, 2012). As outcome measures differ, however, comparisons of the extent of improvement are difficult to make. Still, it can be noted that training components described in earlier examples of educational training in perceptual assessment of voice – such as multiple exposure to ratings, practice, and consensus – were available also to the students in the present study. These components were, however, available not only via PUMA but also through other course activities.

The finding that auditory-perceptual assessment performance increased for certain parameters but not for others provides insight into what aspects of voice assessment may be particularly challenging. As the results indicate, the proportion of non-expert-like student ratings regarding the parameters Aphonia, Hyperfunction, and Roughness is close to 50% also at post-test. From an educational perspective, the finding suggests that more efforts need to be spent on highlighting these voice aspects and offering students more practice opportunities. At the other end of the scale, a high proportion of expert-like ratings were observed for some parameters already before training, such as for Hard glottal attacks and Instability. Consequently, this could signal to course responsible teachers that educational efforts devoted to these parameters are sufficient and need not be prioritised. From a clinical perspective, however, such considerations need to be balanced with what parameters are most critical in clinical practice.

The operationalisation of assessment performance deserves commenting. A consequence of categorising all student ratings as expert-like as long as they fell within the range of ratings provided by experts is that for voice parameters with a wide range of expert ratings, student ratings were more likely to be categorised as expert-like. Indeed, the range of expert ratings for some parameters and voices was wide (*e.g.* for Roughness for speaker C, or for Hyperfunction for speaker D; see Figure 2), highlighting the challenging nature of auditory-perceptual assessment, even among experts (Brunnegård and Lohmander, 2007; Kent, 1996; Kreiman *et al.*, 1993; Oates,

2009; Watterson *et al.*, 2007). It can be noted that the 30 minutes of consensus discussions that the expert clinicians in the present study had is considerably less than the 20-hour consensus procedures for students in Iwarsson and Reinholt Petersen (2012). Although experienced clinicians can be expected to require less time for training than students do, allowing time for consensus discussion and redoing assessments at regular intervals for calibration of listener evaluations is needed and recommended (Klintö and Lohmander, 2023; Oates, 2009; Sell *et al.*, 2009). The wide range of expert ratings for some parameters/voices indicates a need for redoing the ratings, perhaps with more time allowed for consensus discussion to achieve higher agreement between raters, as a step for future development in PUMA. However, as for the interpretation of results in the present study, there are no systematic patterns of expert ratings varying over wider ranges for certain parameters/voices compared to others. Hence, the high proportion of student ratings being expert-like at both pre- and post-test for certain parameters (*e.g.* Hard glottal attacks, Instability, and Voice breaks) could not be attributed to more generous thresholds for qualifying as expert-like for these specific parameters.

Regarding the potential association between the students' time spent on self-learning in PUMA and their assessment performance, it is evident that any increase in performance could not be explained by more time spent in PUMA. It is important to note, however, that the time spent on self-learning in PUMA, as estimated from usage logs, was surprisingly low overall. As an attest of whether students' own motivation and/or capacity is sufficient for them to engage in self-supervised e-learning activities, this observation suggests that it is not. It should be noted that the estimation of usage time was deliberately conservative; with an intent not to assume usage without solid substantiation, there is a risk that usage time was underestimated. Even so, the usage is markedly low, compared to that reported in Lohmander and colleagues (2021), where exercises in PUMA were introduced and integrated into the course, and where the estimated average usage was 16.5 hours per student. For course development, this suggests that course-responsible teachers need to ensure sufficient room for self-training in the course schedule. In addition, students may need explicit instruction and guidance to engage more with self-training during the course. Notably, this was also suggested by the students themselves in the post-course evaluation survey. To encourage more engagement in self-supervised practice, PUMA activities could be more explicitly integrated into other course activities, with careful balance with the students' freedom to control their own learning to avoid the risk of hampering their creativity and own drive to learn (McAllister *et al.*, 2014).

It should be noted that the evaluation of potential increase in student performance between pre- and post-test is laden with noise, reflecting the continuous development of PUMA. For example, the technical settings at pre-test which disabled the disambiguation between intended 0-ratings and non-responses are a notable limitation. Methodologically, we addressed this limitation on the side of caution, to avoid exaggerating performance improvement between pre- and post-test. However, it should be borne in mind that student ratings at pre-test contain some noise in this respect. Another reflection of the continuous development of PUMA is the technical issues at pre-test (resulting in some data loss), which were resolved at post-test.

Although this does introduce noise in the data, we argue that evaluation of educational activities is important also in dynamically changing settings. The alternative, to postpone evaluation until settings are optimal, would entail the risk that evaluation is forever adjourned.

Regarding the design and contents of PUMA exercises, the student evaluation responses indicate a request for feedback. Given that feedback is an important vehicle in perceptual learning (Eadie and Baylor, 2006), the students' request for a key reference for the VSR exercise is understandable. As a recommendation during the course, teachers instead encouraged students to collaborate with peers, to compare and calibrate their ratings amongst each other (*i.e.* a consensus approach). This, however, runs counter to PUMA's ambition to enable self-supervised training. Finding a technical solution to the students' request for feedback therefore remains a priority for the future development of PUMA.

Students' self-evaluation of their auditory-perceptual assessment skills indicates that they still feel insecure in this respect. Considering that auditory-perceptual assessment is a skill that requires repeated practice and continuous calibration even after graduation (Klintö and Lohmander, 2023; Oates, 2009; Sell *et al.*, 2009), one may argue that for students still in the middle of their study programme, this insecurity is not unexpected. In fact, responses indicating certainty would perhaps be concerning. A purposive goal in SLP education is to make students well acquainted with protocols for auditory-perceptual analyses and for them to have developed strategies for perceptual assessments for use in clinical practice. The fact that PUMA is accessible also for clinical supervisors and other clinicians, makes it available as a platform for life-long learning, with the uttermost goal of providing reliable and safe patient care.

Expansion and future directions

Although the present evaluation has contributed several important suggestions for future course development, the need to evaluate the effects of educational activities in SLP education will never cease. In such evaluations, educators need to carefully balance ethical, educational, and methodological concerns. For example, in the present study, the educationally and ethically motivated concern of offering potentially valuable educational resources to all students was prioritized over a methodological concern of having a control group of students to whom access to these resources was denied. In future evaluations, other alternatives may be considered. For example, the documentation of student performance at a time when e-learning resources are unavailable to all students, may serve as a control condition. This approach would allow for a more rigorous assessment of the impact of e-learning tools while ensuring that all students ultimately benefit from educational resources.

In parallel with the evaluation of PUMA-Voice, the PUMA website was expanded with two new areas: Fluency disorders and Acquired language disorders. With these new additions, PUMA now includes eight patient areas, where material can be utilized by SLP students, at their convenience, to meet course requirements and achieve the overall learning objectives of SLP program. As PUMA is expanded with new patient areas, evaluation continues in parallel to ensure the efficient use of e-learning in PUMA in students' acquisition of clinical assessment skills.

Conclusion

In combination with other learning activities, self-supervised e-learning can contribute to improvement in SLP students' auditory-perceptual rating performances. For students to actually engage with self-supervised learning activities, however, activities need to be integrated into the course and allocated sufficient time in the course schedule. Additionally, perceptual e-learning tasks that do not offer feedback should either be developed to include feedback or supplemented by teacher-supported consensus discussions to ensure their value for students' perceptual learning.

Acknowledgements

This project was funded by Karolinska Institutet's Pedagogical Project Funding. We are very grateful to Johan Flodin for the administration of the website and for assistance in the conduction of the test sessions; and to Svante Granqvist for valuable feedback in the writing process.

References

Bruneel, L., Danhieux, A. and Van Lierde, K. 2022. Training speech pathology students in the perceptual evaluation of speech in patients with cleft palate: Reliability results and the students' perspective. *International Journal of Pediatric Otorhinolaryngology* 157, 111145. https://doi.org/10.1016/j.ijporl.2022.111145

Brunnegård, K. and Lohmander, A. 2007. A cross-sectional study of speech in 10-year-old children with cleft palate: Results and issues of rater reliability. *The Cleft Palate Craniofacial Journal* 44(1), 33–44. https://doi.org/10.1597/05-164

Chan, K.M.K. and Yiu, E.M.-L. 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

Eadie, T.L. and Baylor, C.R. 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

Goldstone, R.L. 1998. Perceptual learning. *Annual Review of Psychology* 49, 585–612. https://doi.org/10.1146/annurev.psych.49.1.585

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. 1986. Perceptual and Acoustic Analysis of Dysphonia (Vol. 1986). Solna: Karolinska Institutet.

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

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J. and Wedin, L. 1980. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica* 90(1–6), 441–451. https://doi.org/10.3109/00016488009131746

Hammarberg, B. and Gauffin, J. 1995. Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In O. Fujimura, and M. Hirano (Eds) *Vocal Fold Physiology: Voice Quality Control*. San Diego, CA: Singular Publishing Group, pp. 203–283.

Holmberg, E.B., Hillman, R.E., Hammarberg, B., Södersten, M. and Doyle, P. 2001. Efficacy of a behaviorally based voice therapy protocol for vocal nodules. *Journal of Voice* 15(3), 395–412. https://doi.org/10.1016/S0892-1997(01)00041-8

Iwarsson, J., Bingen-Jakobsen, A., Johansen, D.S., Kølle, I.E., Pedersen, S.G., Thorsen, S.L. and Petersen, N.R. 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 Reinholt Petersen, N. 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

Kent, R.D. 1996. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology* 5(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Kimura, R., Matsunaga, M., Barroga, E. and Hayashi, N. 2023. Asynchronous e-learning with technology-enabled and enhanced training for continuing education of nurses: A scoping review. *BMC Medical Education* 23(1), 505. https://doi.org/10.1186/s12909-023-04477-w

Klintö, K. and Lohmander, A. 2023. Perceptual assessment of cleft palate speech—Bridging the gap from research to clinical practice—The Swedish perspective. *Perspectives of the ASHA Special Interest Groups* 8(5), 986–1002. https://doi.org/10.1044/2023_PERSP-22-00271

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A. and Berke, G.S. 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

Lawn, S., Zhi, X. and Morello, A. 2017. An integrative review of e-learning in the delivery of self-management support training for health professionals. *BMC Medical Education* 17(1), 183. https://doi.org/10.1186/s12909-017-1022-0

Lee, A., Whitehill, T.L. and Ciocca, V. 2009. Effect of listener training on perceptual judgement of hypernasality. *Clinical Linguistics & Phonetics* 23(5), 319–334. https://doi.org/10.1080/02699200802688596

Lohmander, A., Klintö, K., Schalling, E., Szabo Portela, A., Johansson, K. and McAllister, A. 2021. Students take charge of Learning – Using e-learning in perceptual assessment in speech-language pathology. *Scandinavian Journal of Educational Research* 65(3), 468–480. https://doi.org/10.1080/00313831.2020.1716064

Lyberg Åhlander, V., Falk Nilsson, E., Wigforss, E. and Rydell, R. 1999. The project for the development of multimedia methods in logopedics and phoniatrics (PUMP). In: Proceedings of MATISSE – Method and Tool Innovations for Speech Science Education, 1999. pp. 37–40.

McAllister, A., Aanstoot, J., Hammarström, I.L., Samuelsson, C., Johannesson, E., Sandström, K. and Berglind, U. 2014. Learning in the tutorial group: A balance between individual freedom and institutional control. *Clinical Linguistics & Phonetics* 28(1–2), 47–59. https://doi.org/10.3109/02699206.2013.809148

Oates, J. 2009. Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatricaet Logopaedica* 61(1), 49–56. https://doi.org/10.1159/000200768

Oates, J. and Russell, A. 1998. Learning voice analysis using an interactive multi-media package: Development and preliminary evaluation. *Journal of Voice* 12(4), 500–512. https://doi.org/10.1016/S0892-1997(98)80059-3

Sell, D., John, A., Harding-Bell, A., Sweeney, T., Hegarty, F. and Freeman, J. 2009. Cleft audit protocol for speech (CAPS-A): A comprehensive training package for speech analysis. *International Journal of Language & Communication Disorders* 44(4), 529–548. https://doi.org/10.1080/13682820802196815

Silva, R.S.A., Simões-Zenari, M. and Nemr, N.K. 2012. Impacto de treinamento auditivo avaliaçã perceptivo-auditiva da voz realizada por estudantes de Fonoaudiologia. *Jornal Da Sociedade Brasileira de Fonoaudiologia* 24(1), 19–25. https://doi.org/10.1590/S2179-64912012000100005

Walden, P.R. 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

Watterson, T., Lewis, K., Allord, M., Sulprizio, S. and O'Neill, P. 2007. Effect of vowel type on reliability of nasality ratings. *Journal of Communication Disorders* 40(6), 503–512. https://doi.org/10.1016/j.jcomdis.2007.02.002

Supplementary

Figure S1. Screenshot illustrating the exercise 'Identifying the most salient feature' in PUMA. The task for the student is to listen to a voice sample (not visible in the figure), and to select which one of the three response alternatives represents the most salient voice quality parameter in the sample.

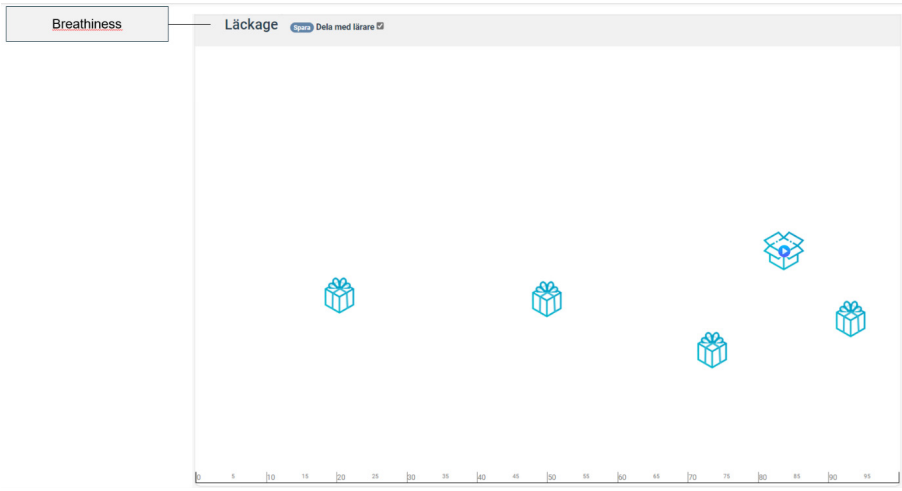

Figure S2. Screenshot illustrating the exercise 'Visual Sort and Rate (VSR)' in PUMA. The five packages represent five different voice samples. Each package can be moved to anywhere on the screen, and each sample can be played by pushing the play button on the package. The task is to rate and order the samples regarding a given voice parameter – in this case 'Breathiness' – along a scale from 0 to 100 (at the bottom of the screen).

https://doi.org/10.58986/al.2025.41881

Table S1. Questions in the evaluation form for recording students' experiences of using PUMA-Voice, in their original Swedish wording and translated into English.

Evaluation questions (Swedish)	English translation
1. Övningarna i PUMA har bidragit till att utveckla mina färdigheter i perceptuell röstanalys	1. The exercises in PUMA have contributed to my development of perceptual voice analysis skills.
2. Jag känner mig väl förberedd för att göra perceptuell bedömning av röst hos patienter i VFU.	2. I feel well-prepared for conducting perceptual assessment of voice in patients during clinical practicum.
3. Vilken av de tre övningsuppgifterna tyckte du var svårast/lättast? Rangordna uppgifterna med den svåraste högst upp.	3. Which of the three exercises did you find the most difficult/easiest? Rank the exercises with the most difficult one first.
4. Utveckla gärna ditt svar!	4. Please elaborate.
5. Vad var det bästa med PUMA-RöSt?	5. What was best with PUMA-Voice?
6. Hur kan uppgifterna i PUMA-RöSt förbättras?	6. How can the exercises in PUMA-Voice be improved?

https://doi.org/10.58986/al.2025.41881

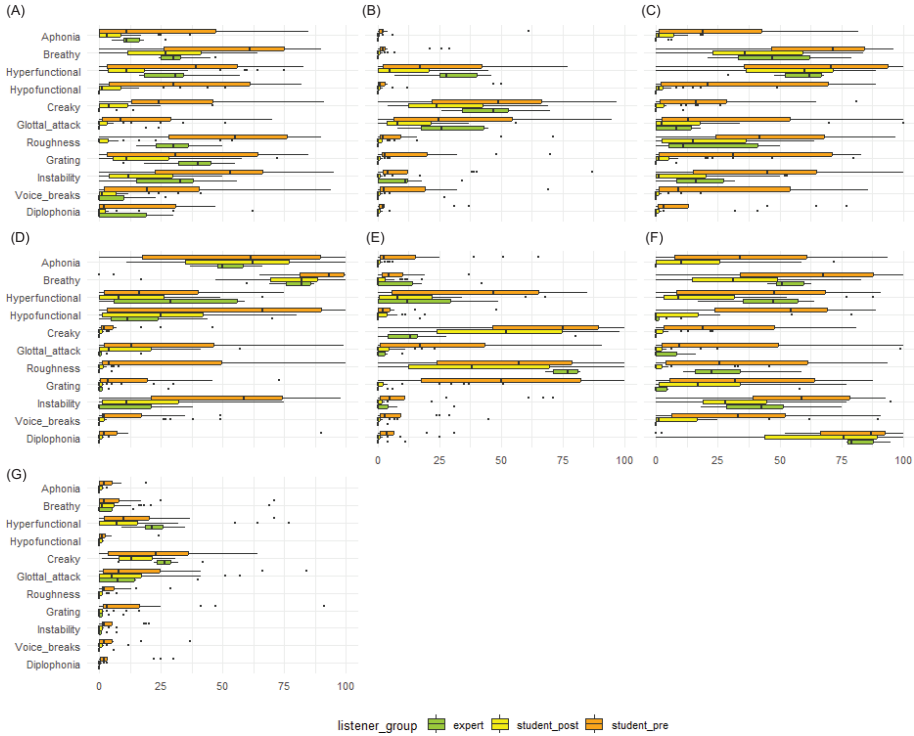

Figure S3. Box plot illustrating the distribution of ratings of seven voice samples A–G regarding the 11 rated voice parameters in SVEA, across expert listeners (expert), students at pre-test (student_pre), and students at post-test (student_post). The *x*-axis represents rating values varying between 0 and 100. Median values are represented as black lines within the boxes; boxes extend between the 25th and 75th percentiles, and lines outside boxes extend to the most extreme values within 1.5 interquartile range from 25th to 75th percentile. The distribution of expert ratings is thus identical to the distribution presented in Figure 3 in the manuscript. Note that for many of the students' ratings at course start, the variation was considerably larger than by course end (as seen by a wider extension of orange boxes, compared to yellow boxes).