Effect of Auditory-Perceptual Training With Natural Voice Anchors on Vocal Quality Evaluation

***Priscila Campos Martins dos Santos, †Maurílio Nunes Vieira, ‡João Pedro Hallack Sansão, and *Ana Cristina Côrtes Gama,** *†*Belo Horizonte and* ‡*São João Del Rey, Minas Gerais*

Summary: Purpose. To analyze the effects of auditory-perceptual training with anchor stimuli of natural voices on inter-rater agreement during the assessment of vocal quality.
Study design. This is a quantitative nature study.
Methods. An auditory-perceptual training site was developed consisting of Programming Interface A, an auditory training activity, and Programming Interface B, a control activity. Each interface had three stages: pre-training/pre-interval evaluation, training/interval, and post-training/post-interval evaluation. Two experienced evaluators classified 381 voices according to the GRBASI scale (G-grade, R-roughness, B-breathiness, A-asthenia, S-strain, I-instability). Voices were selected that received the same evaluation by both evaluators: 57 voices for evaluation and 56 for training were selected, with varying degrees of deviation across parameters. Fifteen inexperienced evaluators were then selected. In the pre-, post-training, pre-, and postinterval stages, evaluators listened to the voices and classified them via the GRBASI scale. In the stage interval evaluators read a text. In the stage training each parameter was trained separately. Evaluators analyzed the degrees of deviation of the GRBASI parameters based on anchor stimuli, and could only advance after correctly classifying the voices. To quantify inter-rater agreement and provide statistical analyses, the AC1 coefficient, confidence intervals, and percentage variation of agreement were employed.
Results. Except for the asthenia parameter, decreased agreement was observed in the control condition. Improved agreement was observed with auditory training, but this improvement did not achieve statistical significance.
Conclusion. Training with natural voice anchors suggest an increased inter-rater agreement during perceptual voice analysis, potentially indicating that new internal references were established.
Key Words: Voice–Voice quality–Dysphonia–Auditory perception–Voice training.

INTRODUCTION

Although the voice can be measured objectively, perceptual changes in voice quality are likely the primary stimulus for patients to seek clinical care.[1] Vocal quality is frequently judged based on the impression that a listener has of the voice of a speaker as a whole; auditory-perception of a voice is both holistic and integrative.[2] Auditory-perceptual evaluation is classically employed to evaluate vocal quality and is frequently employed in the clinical setting as instrumental analyses do not recapitulate human perception.[1]

Auditory-perceptual analyses have several advantages, including evaluation of vocal quality, a perceptual phenomenon in response to an auditory stimulus. As a consequence, perceptual descriptions of voice are potentially intuitive and significant.[2] The advantages of perceptual analyses are also directly related to limitations of instrumental approaches to voice measurement, including the acoustic conditions of the recording environment, characteristics of hardware and software systems, analysis protocols employed, and individual variability in acoustic parameters. Finally, auditory-perceptual analysis is fast, noninvasive, and inexpensive.[2]

However, auditory-perceptual evaluation has limitations. Primarily, it is subjective and based on internal standards to evaluate aberrant voices.[1,3] Although auditory-perceptual assessment is relatively easy and largely free from instrumentation, reliability and sensitivity may vary.[3] Increased variability in intra- and inter-rater agreement has long been a central issue in voice research.[1]

Recent studies outlined factors that may interfere with reliability of auditory-perceptual assessment such as the multidimensional characteristics of voice as well as auditory processing, speech task employed, and the internal standards of evaluators which can be reflective of their personal, professional experience, and previous training.[1,4] The fact that voices are inherently unstable and are often characterized by the presence of more than one aberrant parameter, such as roughness plus breathiness or breathiness plus tension confirmed the need for improved reliability of auditory-perceptual evaluation.[5–8]

Specifically, some tasks have been shown to hold promise to optimize auditory-perceptual evaluation, such as the use of anchor stimuli and auditory training.[3] Specifically, to increase the reliability of auditory-perceptual evaluation, recent research suggested the use of controlled anchored stimuli.[3–5] Anchors are predefined and selected to represent a determined type and/or degree of deviation. In addition, auditory training has also been shown to increase the reliability of auditory-perceptual evaluation, reducing variability and subjectivity of analyses.[4,6–8]

According to previous work, perceptual learning during auditory training is most effective when two key processes are

Accepted for publication October 26, 2017.
From the *Department of Speech-language Pathology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais; †Department of Electronic Engineering, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais; and the ‡Department of Telecommunications Engineering and Mechatronics, Federal University of São João Del Rey, São João Del Rey, Minas Gerais.
Address correspondence and reprint requests to Priscila Campos Martins dos Santos, Universidade Federal de Minas Gerais, Av. Alfredo Balena, 190/249, Belo Horizonte, Minas Gerais, Brazil. E-mail: priscila.fonoaudiologia@gmail.com
Journal of Voice, Vol. 33, No. 2, pp. 220–225
0892-1997

https://doi.org/10.1016/j.jvoice.2017.10.020

involved: stimulus imprinting and differentiation.[9] Stimulus imprinting refers to the development of internal detectors or standards formed by repeated exposure to the stimulus. The form of detector is shaped by the impinging stimulus, and it justifies the term stimulus imprinting. These detectors increase the speed, accuracy, and general fluency with which the stimuli are processed.[9] Differentiation of stimuli allows the listener to learn to differentiate signals; that is, to identify each voice parameter independently. When feedback is provided during training, this differentiation occurs more rapidly.[9]

Therefore, standardization of training to increase sensitivity, reliability, and agreement of auditory-perceptual evaluation of voice is of great relevance for both research and clinical care. The objective of the present study was to analyze the effect of auditory-perceptual training with anchor stimuli of natural voices on inter-rater agreement. It was hypothesized that the inter-rater agreement during auditory-perceptual evaluation would improve after the training employing anchor stimuli of natural voices.

MATERIALS AND METHODS

The current project was approved by the Research Ethics Committee (CAAE—37872314.2.0000.5149) at the Federal University of Minas Gerais. This is a quantitative nature study. All subjects provided written, informed consent before participation. Upon consent, subjects completed a brief questionnaire providing information regarding their experience with auditory-perceptual analysis including any previous training, as well as age and profession. Subjects were then randomly assigned to one of two programming interfaces.

Programming interface

The programming interfaces are websites created with explicit instructions for performing each stage of auditory-perceptive voice evaluation. The evaluators received the link of programming interface and accessed it by the Google Chrome browser. No other instruction was provided by the applicator, so that it could not influence the training. Two programming interfaces were created: Programming Interface A involved auditory training and Programming Interface B was the control activity. The second programming interface was performed 15 days after the first to avoid task memorization.[6,10] Each programming interface was composed of three stages (Figure 1).

Programming interface A

This programming interface involved auditory training.

Process. This activity was structured in three stages:

(A) Pretraining evaluation: in this stage, participants listened to voices and classified them according to the GRBASI scale (G-grade, R-roughness, B-breathiness, A-asthenia, S-strain, I-instability)[11] according the previously reported protocols.
(B) Training: Training focused on the parameters R, B, A, S, and I presented separately. Participants listened to voice stimuli, and subsequently, anchors. Participants then rated the voices on the degree of parameter in training. If their response was incorrect, they were encouraged to repeat the task until correct. It was considered incorrect answers those with classification of degree of parameter in training different from the classification given by the evaluators in the selection of voices to be evaluated—just voices that obtained the same classification of degree of parameter in training by two experienced evaluators were selected.
(C) Post-training evaluation: similar to the pretraining evaluation.

All participants were provided with written definitions of the parameters and degrees of vocal deviation during Programming Interface A. Analyses were completed individually, using

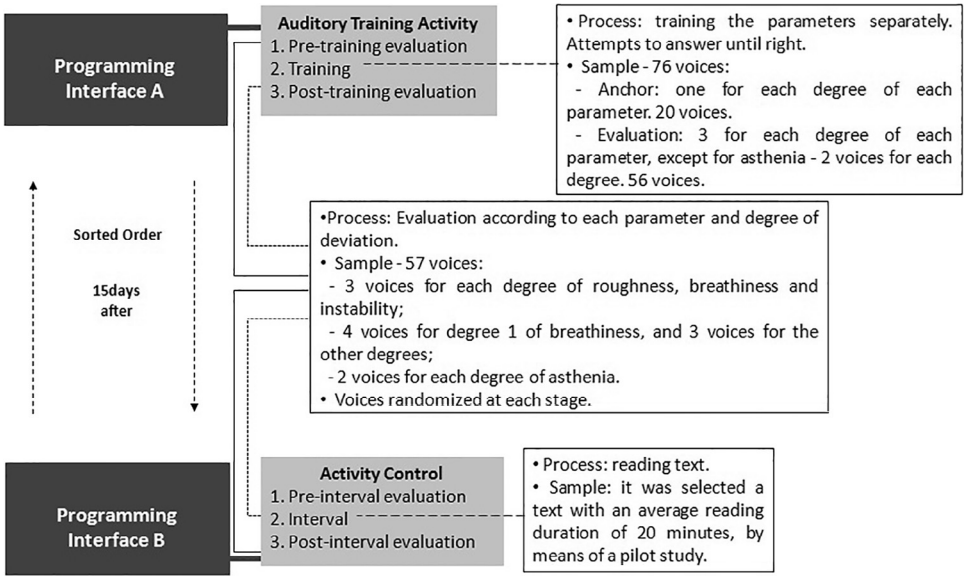

FIGURE 1. Experimental flow chart of auditory-perceptual training programming interfaces.

an over-ear headphone, in a silent classroom, which, however, was not acoustically treated. Each participant set the headphone at a comfortable volume for themselves and could hear voices as many times as necessary.

Selection of vocal stimuli for evaluation. The voice bank of the university clinic, composed of 381 samples of sustained /a/ from male and female adults was employed to compile the samples for evaluation. Two voice specialists with more than 5 years of experience in auditory-perceptual assessment analyzed the voices using an over-ear headphone, model Multilaser Vibe Stereo. Voice samples were classified according to the degree (0—neutral, 1—slight, 2—moderate, and 3—intense) of predominant parameter (R, B, A, S, or I).

Inclusion criteria considered were natural voices of female and male subjects, greater than 18 years of age, with a predominantly deviant vocal parameter. Voices were selected that obtained the same rating by two evaluators on the degree of predominant parameter.

Three stimuli for each parameter (R, B, S, and I) and severity were selected for pre- and post-training, and a grade of one of the parameters received four stimuli, totaling 49 samples. For asthenia, two stimuli were selected for each degree, totaling eight samples. The occurrence of asthenic voices in the bank used for research was reduced when compared with other parameters, which justified the decreased number of stimuli. Therefore, the total number of samples for Programming Interface A was 57.

For the training stage, three stimuli were selected for each degree of each parameter (R, B, S, I). For asthenia, two stimuli for each degree were selected, totaling 56 samples for this stage of Programming Interface A.

Selection of anchors for training. Anchor stimuli were derived from the same bank of voices used to select the stimuli to be evaluated in the pre- and post-training evaluation. The voices were previously classified by two voice specialists, with more than 5 years of experience in auditory-perceptual assessment, according to the degree (0—neutral, 1—slight, 2—moderate, and 3—intense) of predominant parameter (R, B, A, S, and I). Analyses were completed individually, using an over-ear headphone, model Multilaser Vibe Stereo.

The following inclusion criteria were considered: natural voices of women and men, 18 years of age or older, and with deviation degree 0—neutral, 1—slight, 2—moderate, and 3—intense for each predominant parameter. Voices were selected if they obtained the same rating by two evaluators on the degree of predominant parameter.

One sample was selected as an anchor for each degree of deviation for each parameter. Thus, the total number of anchor stimuli for training in Programming Interface A was 20.

Programming interface B

Programming Interface B is a control activity. This program includes the same stages as Program A. However, instead of training, the evaluators read a text, which is the control activity. It was included to allow for analysis of a training effect, eliminating the possibility of improved performance due to chance.

Process. Programming Interface B was also structured in three stages:

(A) Preinterval evaluation: performed in the same way as the pretraining evaluation stage of Programming Interface A.
(B) Interval: participant read text.
(C) Postinterval evaluation: performed in the same way as the preinterval evaluation stage.

Similar to Programming Interface A, written definitions of the parameters and degrees of vocal deviation were provided to the participants during the pre- and postinterval evaluation of Programming Interface B. Analyses were completed individually, using an over-ear headphone, in a silent classroom, which, however, was not acoustically treated. Each participant set the headphone at a comfortable volume for themselves and could hear voices as many times as necessary.

Selection of vocal stimuli to compose the task of evaluation. The same stimuli employed in pre- and post-training of Programming Interface A were used in pre- and postinterval evaluation stages. The samples were randomized for each stage.

Selection of the text to compose the interval step. The interval stage of Programming Interface B was determined to be the same average duration as the training stage of Programming Interface A. Therefore, a pilot study was conducted with three subjects with no prior experience in auditory-perceptual evaluation of voice. These subjects completed the training stage in an average of 21 minutes (standard deviation [SD] = 4.02).

These same subjects were then asked to read a text previously selected by the researchers. The duration of this task was an average of 20 minutes (SD = 2.07).

Each programming interface lasted approximately 1 hour.

Selection of evaluators

The number of evaluators was determined from the Kappa Index proposed by Fleiss[12] with statistical power of 80% and significance level of 5%. These calculations confirmed that 15 evaluators of 57 observations (voices) and 20 variables (R, B, A, S, and I parameters with degrees neutral, slight, moderate, and intense) were adequate to achieve meaningful data.

Fifteen nonexperienced evaluators were selected from a pool of Speech Therapy graduate students, male and female, aged 18 and 28 years. The same evaluators participated in both programming interfaces.

The following inclusion criteria were considered: absence of a hearing loss complaint, agreed to participate in the research accepting the informed consent, responded to the initial questionnaire, and participated in all stages of two programming interfaces.

The order of activities was randomized.

Statistical analysis

All data were archived for further analysis. Responses to the questionnaires, as well as the results of the sample analyzes of the three stages in each programming interface, were inserted into Microsoft Excel. Gwet's AC1 coefficient[12,13] was used to analyze intra- and inter-rater concordance, and to compare pre- and post-training evaluation and pre- and postinterval evaluation concordances. Concordances between the two programming interfaces were calculated via confidence interval (CI) and percentage variation of agreement. All statistical analyses were performed using R (v3.3.1). A significance level of 5% was considered for all analyses.

RESULTS

Twenty-two evaluators initially participated in the study. However, seven were excluded as they did not return to perform the second programming interface. In a brief questionnaire, all participants stated that they did not have previous experience or training in perceptual-auditory assessment of voice. All the participants declared to be undergraduate students of the speech-language pathology course between the first and third period, and had not received content on perceptual-auditory assessment of the voice. The participants' ages ranged from 17 to 28 years, with a mean of 20.06 years.

A mean duration of 58.33 minutes (SD = 13.41) was observed for completion of the auditory training activity, and 48.13 minutes (SD = 19.69) in the performance of control activity. In the pilot study, a mean duration of 21 minutes (SD = 4.02) for the training stage was estimated, and 20 minutes (SD = 2.07) for the interval stage. However, it was found that the evaluators had a mean duration of 12.87 minutes (SD = 3.99) in the training stage, and 9.73 minutes (SD = 6.21) in the interval stage. No difference was observed between inter-rater concordances for the control activity and auditory training activity (Table 1).

Concordance in the control activity decreased with regard to variation of inter-rater agreement in the pre- and post-training evaluation and pre and postinterval evaluation stages. Improved agreement was observed following the auditory training activity, with exception of asthenia, which decreased in both groups (Table 1 and Figure 2).

The order of activities was randomized, and in this way, nine evaluators completed Programming Interface A first, and six evaluators performed the analysis of Programming Interface B initially (Table 2).

DISCUSSION

Auditory-perceptual analysis is influenced by extrinsic and intrinsic factors. The main intrinsic factors identified in the literature are time of professional experience, type of professional training, and previous auditory training, as well as internal standards of the evaluators and their state of attention during analysis.[3,6,14,15] As for extrinsic factors, presentation of vocal stimuli, type of perceptive-auditory scale, parameters, and speech task have been described to alter auditory perceptual analysis.[3,6,14,15]

In the present study, we sought to control many of these factors to ensure that evaluators only experience training. As such, graduate students in speech therapy were selected as they had no previous experience in auditory-perceptual assessment of voice. Stimuli and differentiation with feedback were employed, which have been shown to facilitate learning.[9]

Anchor stimuli were also employed to control extrinsic factors.[4,5] Recent investigation[7] suggested that auditory anchors associated with auditory training yielded increased inter-rater agreement when compared with training without anchors or only with written anchors. The current study performed auditory training using auditory anchor stimuli in addition to written definitions of the parameters across severities. Previous work has shown that anchor stimuli potentially increased inter-rater agreement for the G, R, and B parameters.[7,10,16] In the present study, the G, R, B, A, S, and I parameters were evaluated using the GRBASI scale as a widely used scale in clinical and speech-language pathology research.[6,14,16] Sustained /a/ was also employed as this vowel is stable and consistent and has been associated with high inter-rater concordance.[17,18]

TABLE 1.
Inter-rater Agreement and Percentage Variation of Inter-rater Agreement Between Pre- and Postinterval Evaluation and Pre and Post-training Evaluation in Control and Auditory Training Activities

		CA Pre	CA Pos	PV%	ATA Pre	ATA Pos	PV%
R	AC1 (%)	13,53	12.76	−5.69	15.60	20.42	30.90
	CI 95% (%)	−2.46 to 29.53	−3.07 to 28.58		−0.41 to 31.62	2.49 to 33.48	
B	AC1 (%)	25.07	20.54	−18.07	18.13	20.42	12.63
	CI 95% (%)	10.07–40.08	12.76 to 35.99		2.74–33.52	5.45–20.42	
A	AC1 (%)	25.75	23.92	−7.11	24.32	20.31	−16.49
	CI 95% (%)	10.84–40.65	8.93–38.91		9.68–38.97	5.60–35.03	
S	AC1 (%)	30.37	29.30	−3.51	19.83	27.14	36.85
	CI 95% (%)	16.61–44.13	15.73–42.87		6.12–33.54	13.41–40.87	
I	AC1 (%)	23.99	23.60	−1.63	22.04	22.23	0.86
	CI 95% (%)	8.79–39.20	8.53–38.67		7.02–37.07	6.85–37.61	

Notes: For statistical analysis, AC1 (first-order agreement coefficient), confidence interval, and percentage variation (the difference between the percentage values of the pre- and postinterval and pre and post-training evaluation stages) were considered.
Abbreviations: CA Pre, preinterval evaluation of control activity; CA Pos, postinterval evaluation of control activity; PV, percentage variation; ATA Pre, pretraining evaluation of auditory training activity; ATA Pos, post-training evaluation of auditory training activity; R, roughness; B, breathiness; A, asthenia; S, tension; I, instability.

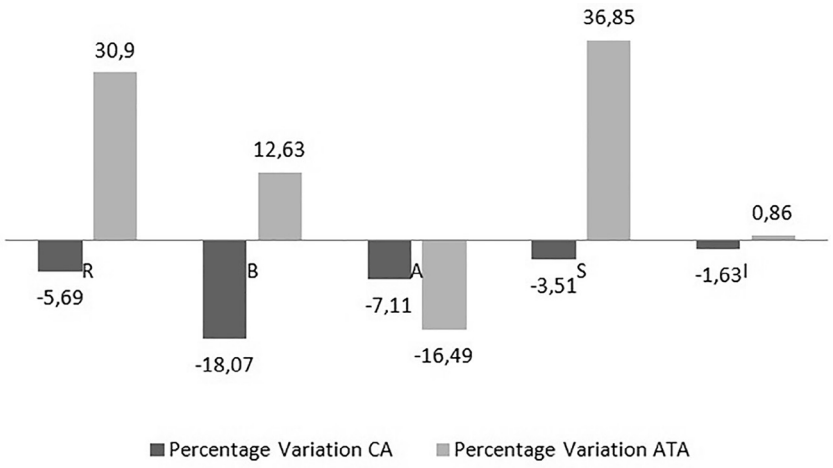

FIGURE 2. Comparison of the percentage variation of inter-rater agreement between pre- and postinterval and pre- and posttraining evaluation in control and auditory training activities. *Abbreviations:* R, roughness; B, breathiness; A, asthenia; S, tension; I, instability; CA, control activity; ATA, auditory training activity.

Decreased inter-rater agreement was observed, which was expected as the evaluators were inexperienced. These data agree with findings from a recent study that verified that experience of the evaluator impacted inter-rater agreement.[19] However, decreased agreement may also be related to the state of evaluator attention, as has been reported in the literature.[6] The duration of activities in the current study was approximately 1 hour. For future studies, the pretraining evaluation could be performed at a different sitting to attempt to avoid this confound.

Another factor that may have influenced inter-rater agreement was the number of training sessions. Studies have shown an increase in agreement after approximately 10 training sessions.[6,8] Additional training sessions are suggested for future studies, as well as a post-training evaluation at a different sitting from the training to analyze the duration of the training effect. No statistically significant differences were observed between the auditory training and control activities with regard to inter-rater agreement. However, upon *post hoc* review, decreased agreement was observed across all parameters in the postinterval stage. In the auditory training activity, an increase in agreement across all parameters was observed in the post-training stage, except for asthenia, for which there was no change in agreement (Figure 1). Based on these data, one may infer that the improved agreement following auditory training was the result of training. That is, training enhanced the acquisition of new internal references. However, this increased agreement was discreet. Clearly, further investigation is warranted regarding the relevant factors that interfere with auditory-perceptual evaluation to eventually standardize training, and consequently, increase reliability.

TABLE 2.
The Order of Activities by the Evaluators

First Activity	Evaluators	Total
Programming Interface A	Evaluator 1 Evaluator 2 Evaluator 4 Evaluator 6 Evaluator 9 Evaluator 10 Evaluator 12 Evaluator 13 Evaluator 14	9 Evaluators
Programming Interface B	Evaluator 3 Evaluator 5 Evaluator 7 Evaluator 8 Evaluator 11 Evaluator 15	6 Evaluators

Notes: The order of activities was randomized. The evaluators were identified by numbers.

Challenges regarding the evaluation of asthenia has been reported across studies.[14,16,20] This decreased agreement may be related to the lower number of stimuli for this parameter, due to the lower occurrence of asthenic voice in the bank used for research, yielding greater inconsistency for this classification. The multidimensional characteristics of voice and the nature of auditory processing make auditory-perceptual evaluation challenging. The evaluation of voice depends on several internal standards, which likely underlie substantial variability in the reliability of auditory-perceptual evaluation.[3,14,21] The fact that voices are not stable and are often characterized by more than one parameter, such as roughness and breathiness, breathiness and tension, or these three parameters combined, also influences the reliability of auditory-perceptual evaluation.[3,14,21] Synthesized stimuli have been employed to optimize auditory-perceptual evaluation as it allows for precise control of acoustic properties. Synthetic stimuli also allow for analysis of each vocal parameter separately, which is not always possible as voices are typically characterized by more than one parameter.[16,21] In this regard, synthesized anchored stimuli in a structure called "binary tree" likely hold promise.[16] Based on these data, a clear need exists for further investigation into factors that influence auditory-perceptual evaluation with the ultimate goal of standardized training to ensure increased intra- and inter-rater agreement and reliability.

The order of activities was randomized. Nine evaluators performed Programming Interface A first, and six evaluators performed Programming Interface B first (Table 2). It was not possible to estimate if there is a statistically significant difference of inter-rater agreement between the different orders of programming interface realization, since the number of evaluators does not generate degree of freedom for this type of analysis through the coefficients AC1 and kappa, losing the statistical significance of analysis.[13]

CONCLUSION

Training with natural voice anchors suggest an increased inter-rater agreement during perceptual voice analysis, potential indicating that new internal references were established.

REFERENCES

1. Behlau M. *Voz: o Livro do Especialista*. Rio de Janeiro: Revinter; 2001.
2. Oates J. Auditory-perceptual evaluation of disordered vocal quality—pros, cons and future directions. *Folia Phoniatr Logop*. 2009;61:49–56.
3. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Hear Res*. 2011;54:430–447.
4. Solomon NP, Helou LB, Stojadinovic A. Clinical versus laboratory ratings of voice using the CAPE-V. *J Voice*. 2011;25:e7–e14.
5. Sofranko JL, Prosek RA. The effect of the levels and types of experience on judgment of synthesized voice quality. *J Voice*. 2014;28:24–35.
6. Silva RSA, Simões-Zenari M, Nemr NK. Impacto de treinamento auditivo na avaliação perceptivo-auditiva da voz realizada por estudantes de Fonoaudiologia. *J Soc Bras Fonoaudiol*. 2012;24:19–25.
7. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. *J Voice*. 2009;23:341–352.
8. Kreiman J, Gerratt BR. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812.
9. Goldstone RL. Perceptual learning. *Ann Rev Psychol*. 1998;49:585–612.
10. Brinca L, Batista AP, Tavares AI, et al. The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *J Voice*. 2015;29:776, e7-e14.
11. Hirano M. *Clinical Examination of Voice*. New York: Springer Verlag; 1981:81–84.
12. Wongpakaran N, Wongpakaran T, Wedding D, et al. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:1.
13. Gwet KL. Testing the difference of correlated agreement coefficients for statistical significance. *Educ Psychol Meas*. 2016;76:609–637.
14. Chaves CR, Campbell M, Gama ACC. The influence of native language on auditory-perceptual evaluation of vocal samples completed by Brasilian and Canadian SLPs. *J Voice*. 2017;31:258.e1–258.e5.
15. Lopes LW, Cavalcante DP, Costa PO. Severity of voice disorders: integration of perceptual and acoustic data in dysphonic patients. *Codas*. 2014;26:382–388.
16. Vieira MN, Sansão JPH, Yehia HC. Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms. *Speech Commun*. 2014;61–62:17–32.
17. Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *J Soc Bras Fonoaudiol*. 2012;24:107–112.
18. Martins PC, Couto TE, Gama ACC. Avaliação perceptivo-auditiva do grau de desvio vocal: correlação entre a Escala Analógica Visual e escala numérica. *CoDAS*. 2015;27:279–284.
19. Byron S, Gama ACC, Chaves CR. Interferência do tempo de experiência na concordância da análise perceptivo-auditiva. *Distúrb Comum*. 2016;28:415–422.
20. Freitas SV, Pestana PM, Almeida V, et al. Audio-perceptual evaluation of Portuguese voice disorders—an inter and intra-judge reliability study. *J Voice*. 2014;28:210–215.
21. Englert M, Madazio G, Gielow I, et al. Perceptual error identification of human and synthesized voices. *J Voice*. 2016;30:639, e17-e23.