OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Craniofacial/Pediatric

# Exploration of the Utility of the Generic ICHOM Standard Set Measures in Evaluating the Speech of Patients with Cleft Lip/Palate

Amanda C. Del Risco, BA*
Kristina Dunworth, BA*
Banafsheh Sharif-Askary, MD*
Arthur H. "Barron" Suárez, MS, CCC-SLP†
Jillian Nyswonger, MS, CCC-SLP†
Matthew Ford, MS, CCC-SLP‡
Jennifer Kern, MS, CCC-SLP†
Carlee Jones, MS, CCC-SLP§¶
Eileen Raynor, MD*†§
Alexander C. Allori, MD, MPH*§¶

**Background:** The International Consortium of Health Outcome Measurements (ICHOM) standard set for cleft care appraisal recommends clinicians assess articulation with percentage consonants correct (PCC) and velopharyngeal function with velopharyngeal competency rating (VPC-R). This study explores the utility and limitations of these generic measures in detecting cleft speech sound disorders by comparing them with two cleft-specific speech-rating systems, cleft audit protocol of speech–augmented Americleft modification (CAPS-A-AM) and Pittsburgh weighted speech scale (PWSS).

**Methods:** Consecutive children with repaired, nonsyndromic cleft lip/palate, aged 5 years or older (n = 27) underwent prospective speech evaluations conducted at a single academic institution. These evaluations were conducted, recorded, and evaluated by blinded speech-language pathologists experienced with all tools.

**Results:** When comparing measures of articulation, PCC scores correlated better with scores for relevant subcomponents of CAPS-A-AM than PWSS. When comparing measures of velopharyngeal function, VPC-R scores correlated well with relevant components of both scales. Using a "screening test versus diagnostic test" analogy, VPC-R ratings were 87.5% sensitive and 73.7% specific for detecting velopharyngeal dysfunction according to subcomponents of CAPS-A-AM, and 70.6% sensitive and 100% specific according to subcomponents of PWSS.

**Conclusions:** This exploratory study demonstrates that PCC and VPC-R perform moderately well in detecting articulatory and velopharyngeal dysfunction in patients with cleft lip/palate; however, these tools cannot describe nuances of cleft speech sound disorder. Thus, although PCC and VPC-R adequately track basic minimum outcomes, we encourage teams to consider extending the standard set by adopting a cleft-specific measurement system for further evaluation of the tools. *(Plast Reconstr Surg Glob Open 2024; 12:e5519; doi: 10.1097/GOX.0000000000005519; Published online 19 January 2024.)*

*From the *Department of Otolaryngology, Head and Neck Surgery & Communication Sciences, Duke University School of Medicine, Durham, N.C.; †Department of Otolaryngology, Head and Neck Surgery, & Communication Sciences, Duke University Health System, Durham, N.C.; ‡Division of Plastic, Maxillofacial, and Oral Surgery, Department of Surgery, Duke University Health System, Durham, N.C.; §Duke Cleft & Craniofacial Center, Duke Children's Hospital, Durham, N.C.; and ¶Division of Plastic, Maxillofacial, and Oral Surgery, Department of Surgery, Duke University Health System, Durham, N.C.*

*Received for publication June 15, 2023; accepted November 8, 2023.*

*Presented at NCSPS, October 9, Pinehurst, North Carolina, and at ASPO on May 5, 2023, Boston, Mass.*



DOI: 10.1097/GOX.0000000000005519

## INTRODUCTION

Children born with cleft palate with or without cleft lip are at high risk of difficulty with speech and communication. This risk is influenced by many factors, such as age, cleft phenotype and severity, timing and quality of palatal repair, presence of oronasal fistula, velopharyngeal function, hearing ability, primary language spoken, and also socioeconomic determinants of health.[1,2] The American Cleft Palate-Craniofacial Association recommends that children with cleft lip and/or palate (CL/P) receive coordinated care through a multidisciplinary team that includes speech-language pathologists.[3] Formal speech assessment by these clinicians may identify pathology that

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

can be addressed by speech therapy and/or operative intervention for structural pathology.

Formal perceptual assessment by a trained listener remains the gold standard for appraisal of speech in CL/P, but this remains a subjective endeavor. Multiple tools have been developed to help standardize this process of formal speech assessment and also to quantify some measurements for ease of comparison. Two very detailed measurement systems that are widely used in the United States are cleft audit protocol for speech–augmented (CAPS-A) and Americleft modification (CAPS-A-AM)[4,5] and Pittsburgh Weighted Values for Speech Symptoms Associated with Velopharyngeal Incompetence (PWSS).[6,7] The CAPS-A-AM and PWSS systems have each been validated[5,6] in the peer-reviewed literature and have demonstrated great utility when used by experienced speech-language pathologists.[4,7–12] Appropriate usage of these measurement systems requires formal training—for example, Americleft provides CAPS-A-AM training, usually lasting 2 days, which has resulted in high intra- and interrater reliability for trained raters after this training period.[5,13]

Although formal speech assessment tools such as CAPS-A-AM and PWSS have been well-received and adopted widely, no single measurement system has emerged as being universally accepted at this time.[14–16] Moreover, CAPS-A-AM and PWSS are specific to English (and CAPS-A-AM is even further specific to the American idiom, as opposed to CAPS-A for the British idiom) and thus cannot be applied to other languages for cross-linguistic cleft speech outcome studies. For these reasons, the speech-language pathologists in the working group for the International Consortium of Health Outcomes Measurement (ICHOM) chose not to include cleft-specific instruments in its 2017 Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care[17] but rather proposed generic tools that were available for many languages and might be more easily disseminated and implemented globally. The specific tools recommended by the ICHOM working group are the clinician-rated percentage consonants correct (PCC) imitative sentence scoring form[18] for articulation, the clinician-rated velopharyngeal closure or competence rating scale (VPC-R)[19] for velopharyngeal function, the parent/guardian-rated intelligibility in context scale (ICS)[20,21] for functional intelligibility, and the patient-reported speech function and speaking-related distress from CLEFT-Q.[17]

Because PCC and VPC-R are generic instruments, there is the possibility that they may be less accurate or descriptive than cleft-specific measurement systems such as CAPS-A-AM or PWSS. The limited scope of these modular tools (PCC to articulation, and VPC-R to velopharyngeal function) may underevaluate other aspects of cleft speech pathology that can be captured in the cleft-specific systems. Interestingly, to date, PCC and VPC-R have not been compared with CAPS-A-AM or PWSS. Admittedly, side-by-side comparison of scales is difficult; however, such a comparison is necessary to understand what aspects of speech pathology the generic measures can accurately quantify, what additional information they may provide, and what aspects of speech pathology may be underevaluated by them. Moreover, such a comparison is critically important if teams are to decide whether to stick with the generic instruments recommended in the ICHOM standard set or to adopt a cleft-specific measurement system in their own practice. To this end, the purpose of this study is to provide an initial exploration of these important questions, namely comparing and contrasting PCC and VPC-R with the corresponding components of CAPS-A-AM and PWSS. Our expectation is that PCC and VPC-R would serve well as generic instruments but may face certain limitations and circumstances that may warrant use of cleft-specific measurement systems.

### Takeaways

**Question:** What are the utilities and limitations of the ICHOM standard set measures for articulation and velopharyngeal function?

**Findings:** Our results demonstrate that the ICHOM standard set measures can grossly detect articulatory and velopharyngeal dysfunction in patients with cleft lip/palate; however, these tools cannot describe nuances of cleft speech sound disorder.

**Meaning:** The ICHOM standard set measures can track basic minimum outcomes as intended by ICHOM and can be adopted by cleft teams in addition to cleft-specific measures.

## METHODS

This study was a single-institution prospective exploratory study approved by the institutional review board (PRO00104806). Parental consent and child assent were obtained.

### Participants

Consecutive children aged 5–17 years old with repaired CL/P were invited to participate in this study during their routine clinic visits with the multidisciplinary cleft/craniofacial team. Subjects were included if they were English-speaking with primary language at home being English and excluded if they were non-English-speaking or if their primary language at home was not English. Children were also included if they had intellectual disability or neurocognitive delay. All cleft phenotypes were eligible for inclusion. Children with cleft lip only (who would be expected to have fewer speech errors when compared with other cleft phenotypes) were included to better characterize the ability of the scales to describe the full range of speech outcomes.

### Data Collection

All participants underwent a standardized speech evaluation in a private clinic room. These evaluations consisted of the third version of the Goldman-Fristoe Test of Articulation (GFTA-3)[22] as the collection of consonant sounds in English, and a 2-minute spontaneous speech sample based on the prompts from the Americleft Speech Handbook.[5] (**See figure, Supplemental Digital**

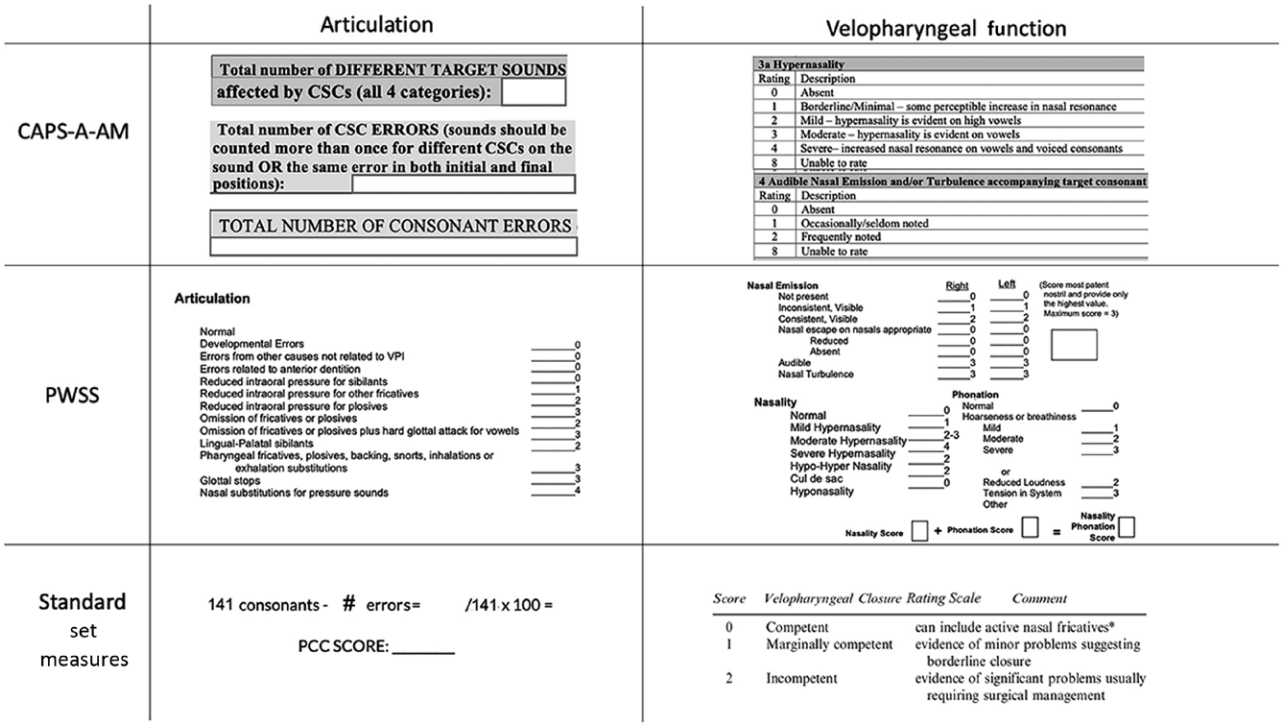


**Fig. 1.** Subcomponents from each speech assessment measure compared via domain. The first column describes the different instruments used in the study, with the remaining columns showing the subcomponents from each domain: articulation and velopharyngeal function.

**Content 1**, which shows spontaneous speech prompt, http://links.lww.com/PRSGO/D3.) These speech evaluations were conducted by trained clinical speech-language pathologists belonging to the cleft/craniofacial team. All evaluations were audio/video recorded using an Apogee MiC 96k studio-quality cardioid condenser microphone (Apogee Electronics Corp., Santa Monica, Calif.), connected to an iPad (Apple Inc., Cupertino, Calif.) and set to 24-bit analog-to-digital conversion with the highest level of gain. Videos were later edited by a researcher (B.B.S.-A.), standardized to contain only the speech assessment of each subject with an equalized volume and audio quality. Each video was de-identified and uploaded to a secure digital platform for later blinded appraisal.

Four speech-language pathologists were assigned to view each video and rate speech using one outcome measurement system: two raters (A.S. and J.N.) were assigned to use PCC and VPC-R, one rater (J.N.) was assigned to use CAPS-A-AM, and one rater (M.F.) was assigned to use PWSS. The raters using CAPS-A-AM and PWSS had received formal training in these systems and each has used them in clinical practice for more than 7 years. All raters were experienced with the PCC and VPC-R scales.

### Speech Outcome Measures

The speech outcome instruments are described in detail. For CAPS-A-AM and PWSS, subcomponents relevant to articulation, velopharyngeal function, and functional intelligibility needed to be identified for the intermeasure comparisons. Discussion between our speech-language pathologists and outside colleagues was held to discuss the individual subcomponents related to each of the intended outcomes. The specific subcomponents selected from each scale to be compared are depicted in Figure 1, and the full instruments are provided in **appendix, Supplemental Digital Content 2**, which shows the speech-rating instruments used in the study. (http://links.lww.com/PRSGO/D4.)

#### *ICHOM Standard Set Outcome Measures*

PCC is a measure of speech sound production, namely articulation.[17,18] It is a continuous scale (0%–100%) indicating the speech-language pathologist's perception of a participant's correct consonant production skills, at the single-word level for our study, with higher percentages on this scale indicating better articulation accuracy. For the speech sample used, a total of 141 consonants were presented to participants.

VPC-R, also included in the ICHOM standard set, is an overall measure of hypernasality, audible nasal air leakage, and weak articulation.[17,19] VPC-R is a three-point ordinal scale (0, 1, and 2) indicating the speech-language pathologist's overall auditory perceptual impression of a participant's velopharyngeal function after a clinical examination. A score of 0 on this scale indicates a competent velopharyngeal closure, whereas a score of 1 indicates "marginally competent" and 2 indicates "incompetent" velopharyngeal closure.

*Cleft Audit Protocol for Speech–Augmented Americleft Modification*

This measurement system was developed by the Americleft speech group as a modification of the UK cleft audit protocol for speech–augmented (CAPS-A) scale using American (rather than British) idiomatic English words and phrases (**Supplemental Digital Content 2**, **http://links.lww.com/PRSGO/D4**).[4,5,13] CAPS-A-AM consists of eight subsections, scored by a speech-language pathologist trained in the system. The subsections of the measure include cleft speech characteristics (CSCs), hyper- and hyponasality, audible nasal emissions, and speech acceptability. The subcomponents on the CAPS-A-AM deemed most relevant to articulation by our group are "total consonant errors" and "total CSC errors," the latter of which is more specific for cleft-related articulation errors. Each is scored on a continuous scale, with higher numbers on each of these scales indicating more articulatory dysfunction (more consonant errors and CSC errors, respectively). The CAPS-A-AM subcomponents related to velopharyngeal function are "hypernasality," which is scored on a five-point ordinal scale (0–4), and "audible nasal emission (and/or turbulence)," scored on a three-point ordinal scale (0–3).[23] Of note, audible nasal emissions can also be caused by unrepaired oronasal fistulas in addition to velopharyngeal incompetence, but the scale is agnostic to etiology.

*Pittsburgh Weighted Speech Scale*

The Pittsburgh Weighted Values for Speech Symptoms Associated with Velopharyngeal Incompetence (PWSS) system is a quantitative scale, validated in 1979, used to measure velopharyngeal insufficiency based on nasal air emission, facial grimace, resonance, voice quality, and articulation.[9,10,19,24] Each speech characteristic receives a separate tally score, and the sum of various characteristics is used to classify velopharyngeal function: a sum of 0 indicates competency; 1–2, borderline competency; 3–6, borderline incompetency; 7 or greater, incompetent valve. The PWSS outcomes most relevant to velopharyngeal function are total "nasality phonation score," total "nasal emission score," and "probable nature of valve," with each of these graded on a continuous scale. A higher "nasality phonation" score indicates a higher degree of hypernasality and/or a higher degree of breathiness or hoarseness of phonation. A higher "nasal emission score" indicates the presence of nasal turbulence and/or audible nasal emissions, which can be due to velopharyngeal incompetence or the presence of an unrepaired oronasal fistula. The "probable nature of valve" score reflects an overall estimation of velopharyngeal valve function. Although the PWSS is primarily designed for appraising function of the velopharyngeal valve, the instrument itself does also include a direct measurement of articulation.

**Graphical and Statistical Analysis**

Because the study is explorational, no formal hypothesis was tested, and this study is considered hypothesis-generating for further work. For each subdomain (articulation and velopharyngeal function), graphs of PCC and VPC-R were constructed and compared with the corresponding components identified from the CAPS-A-AM and PWSS scales. For articulation, PCC was plotted against CAPS-A-AM "total consonant errors" and "total CSC errors" and PWSS "articulation." PWSS "articulation" was then plotted against CAPS-A-AM "total consonant errors" and "CSC errors." These variables are continuous, so in each scatterplot, the line of best fit was drawn and Spearman correlation coefficients were calculated. The strength of Spearman correlation coefficients was judged using the predefined thresholds established by Chan[25] for interpretation of correlation coefficients in medicine. According to these guidelines, a correlation coefficient of less than 0.3 is poor, fair is 0.30–0.59, moderately strong is 0.6–0.79, and a very strong correlation is 0.8 and above.[25] The accuracy of PCC was judged to be adequate if its correlation to conceptually equivalent measures was at least fair. For velopharyngeal function, VPC-R was plotted against CAPS-A-AM "hypernasality" and "audible nasal emission" and against PWSS "nasality phonation score," "nasal emission score," and "probable nature of valve." As VPC-R is an ordinal variable and the others are continuous, these were box-and-whisker plots depicting median, interquartile range, and range. Correlation of VPC-R to other measures was quantified using a Kendall tau coefficient because of the ordinal nature of this scale. As no commonly accepted interpretation guidelines exist for Kendall tau, the same thresholds applied to Spearman correlation coefficients were applied here. Again, the accuracy of VPC-R was judged to be adequate if its correlation to conceptually equivalent measures was at least fair. Statistics and figures were generated using Python 3.8.12,[26] matplotlib 3.5.2,[27] SciPy 1.8.0,[28] pingouin 0.5.2,[29] and scikit-learn 1.01.[30]

**Interrater Reliability of PCC and VPC-R**

Two raters were assigned to PCC and VPC-R to allow calculation of interrater reliability, thus to better characterize the instruments included in the ICHOM standard set. (Interrater reliability of CAPS-A-AM and PWSS was not the subject of the present study, although these characteristics have been previously published.) Assessment of interrater reliability of PCC was determined by calculation of intraclass correlation coefficients based on a single-rater, consistency, two-way mixed-effects model using the pingouin 0.5.2[29] package in Python, with interpretation according to the guidelines published by Cicchetti.[31] Assessment of interrater reliability for VPC-R was based on Cohen Kappa calculated using sklearn.metrics[30] package in Python, with interpretation according to the guidelines published by Cohen.[32]

**Calculation of Sensitivity/Specificity of VPC-R as a "Screening Test"**

To assess whether the VPC-R instrument might best serve as a screening test, we calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for VPC-R. For this calculation, VPC-R was dichotomized, where VPC-R scores of 1 (marginally competent) and 2 (incompetent) translated to a "positive" screening test for velopharyngeal dysfunction, and a VPC-R score of 0 (competent) translated to a "negative" screening test. These dichotomized results were then compared against a composite

CAPS-A-AM score (based on "hypernasality" and "audible nasal emissions") and a composite PWSS score (based on "probable nature of valve," "nasality phonation score," and "nasal emission score"). To calculate the composite CAPS-A-AM score, a participant's velopharyngeal apparatus was considered to be in the functional state if the "hypernasality" score was 0, 1, or 2 and the "audible nasal emission" score was 0. Participants not meeting these requirements were classified to be in the pathological state. (Note that in clinical practice, some clinicians might consider a "hypernasality" score of 2 to warrant further investigation, regardless of the "audible nasal emission" score being zero.) To calculate the composite PWSS score, a participant's velopharyngeal apparatus was considered to be in the functional state if the "probable nature of valve" score was 0 or 1–2, the "nasality phonation" score was 0, 1, or 2, and the "nasal emission" score was 0. Participants not meeting these requirements were considered to be with the dysfunctional state.

## RESULTS

### Cohort Description

Recruited subjects totaled 27 children with a history of isolated nonsyndromic, repaired CL/P, without comorbid intellectual disability or neurocognitive delay. The average age of the cohort was 8.6 years (range 5–17 y). Most subjects (21 of 27) had both a cleft lip and palate, four of 27 had cleft palate only, and two of 27 had cleft lip only. Four subjects (14.8%) had an unrepaired oronasal fistula at the time of speech evaluation, with four different subjects (14.8%) having a persistent anterior nasolabial fistula (which is typically repaired at time of alveolar bone grafting). Twelve subjects (44.4%) had history of oronasal fistula that had been repaired before this study. Seven subjects (25.9%) had undergone prior operative intervention for velopharyngeal insufficiency (eg, pharyngoplasty). Seven subjects (25.9%) had been adopted as infants or very young children; however, all participants, regardless of adoption history, were primarily English-speaking at the time of evaluation. The table (**Supplemental Digital Content 3**, which shows distribution of instrument score data, **http://links.lww.com/PRSGO/D5**) displays the range of instrument score data on the respective scales for the overall cohort (also refer to Fig. 4, which is discussed further in the limitation section of this study).

### Interrater Reliability for PCC Was "Good" and for VPC-R Was "Moderate"

The two raters evaluating PCC demonstrated "good" intrarater reliability according to the guidelines published by Cicchetti,[31] with an intraclass correlation coefficient and 95% confidence interval of 0.73 (0.49, 0.87). Interrater reliability for VPC-R was "moderate" according to the guidelines published by Cohen,[32] with a Cohen Kappa of 0.59.

### PCC Scores for Articulation Correlated Moderately Well with CAPS-A-AM

The median PCC score was 87.5%, median PWSS "articulation" score was 0, median CAPS-A-AM "total consonant errors" score was 7, and median CAPS-A-AM "total CSC errors" was 6. Figure 2 depicts the relationship of four articulatory measurements: PCC; CAPS-A-AM "total consonant error"; CAPS-A-AM "total CSC errors"; and PWSS "articulation."

PCC, as a measure of general articulatory precision, demonstrated a moderately strong[25] inverse correlation to the generic measure of articulatory errors from CAPS-A-AM, called "total consonant errors," with a Spearman correlation coefficient (and *P* value) of −0.60 (0.001). PCC demonstrated a fair[25] inverse correlation to the CAPS-A-AM measure of cleft-specific articulatory errors, called "total CSC errors," with a Spearman correlation coefficient of −0.48 (*P* value 0.011).

### PCC Could Not Be Meaningfully Compared Against PWSS "Articulation"

PCC demonstrated a fair[25] inverse correlation to the PWSS measure called "articulation," with a Spearman correlation coefficient −0.44 (*P* value 0.022). PCC is a count of articulatory errors, whereas the PWSS does not quantify the number of errors but only describes what types

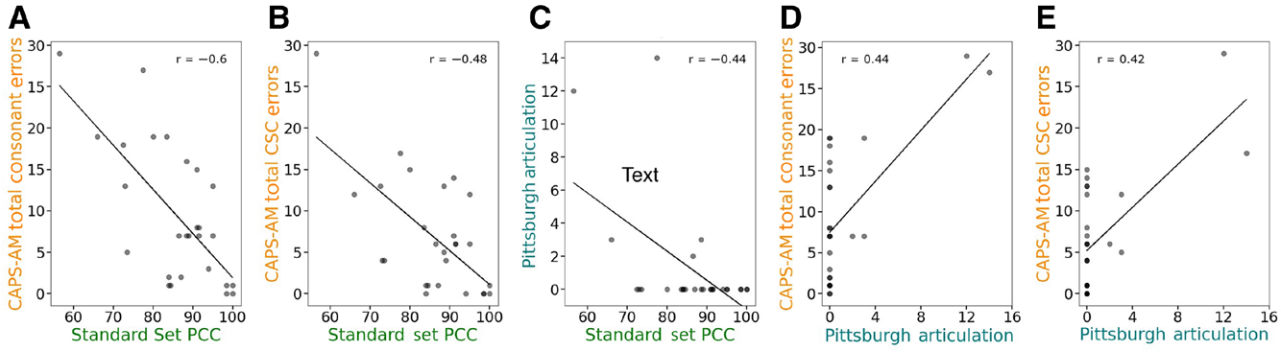


**Fig. 2.** Comparison of instruments related to articulation. The first three graphs (A–C) plot PCC scores along the horizontal axis with CAPS-A-AM "total consonant errors," CAPS-A-AM "total CSC errors," and PWSS "articulation" scores respectively plotted along the vertical axis. The fourth and fifth graphs (D–E) plot PWSS articulation scores along the horizontal axis with CAPS-A-AM "total consonant errors" and CAPS-A-AM "total CSC errors" respectively plotted along the vertical axis. All subcomponents on this graph are scored on a continuous scale, with the minimum score being 0 and maximum being 100.

of errors are observed in the sample. To double-check whether the poor correlation between PCC and PWSS "articulation" was due to the structure and purpose of PWSS, we also compared PWSS "articulation" against the aforementioned articulatory measures within CAPS-A-AM: the Spearman correlation coefficient of PWSS "articulation" versus CAPS-A-AM "total consonant errors" was 0.44 (*P* value 0.023), and PWSS "articulation" versus CAPS-A-AM "total CSC errors" was 0.42 (*P* value 0.029). In other words, the "articulation" component of PWSS is structured too differently to permit a meaningful comparison with PCC or the articulatory components of CAPS-A-AM.

**VPC-R Scores Demonstrated "Fair" Correlation with CAPS-A-AM and "Moderately Strong" Correlation with PWSS**

Figure 3 depicts the relationship of VPC-R scores compared with scores from related subcomponents of CAPS-A-AM and PWSS. The median VPC-R score was 0, the median CAPS-A-AM "hypernasality" score was 1, median CAPS-A-AM "audible nasal emissions" score was 0, median PWSS "probable nature of valve" score was 3–6 (borderline incompetent), median PWSS "nasality phonation" score was 0, and median PWSS "nasal emission" score was 3. VPC-R demonstrated a fair positive correlation with CAPS-A-AM "hypernasality" and "audible nasal emission," with Kendall tau coefficients of 0.599 (*P* value 0.0004) and 0.495 (*P* value 0.0059), respectively. VPC-R demonstrated a moderately strong positive correlation with PWSS's "probable nature of valve," "nasality phonation scores," and "nasal emission," with Kendall tau coefficients of 0.741 (*P* value 0.000025), 0.787 (*P* value 0.000007), and 0.628 (*P* value 0.0007), respectively. Figure 4 presents the distribution of all scores reported for each scale.

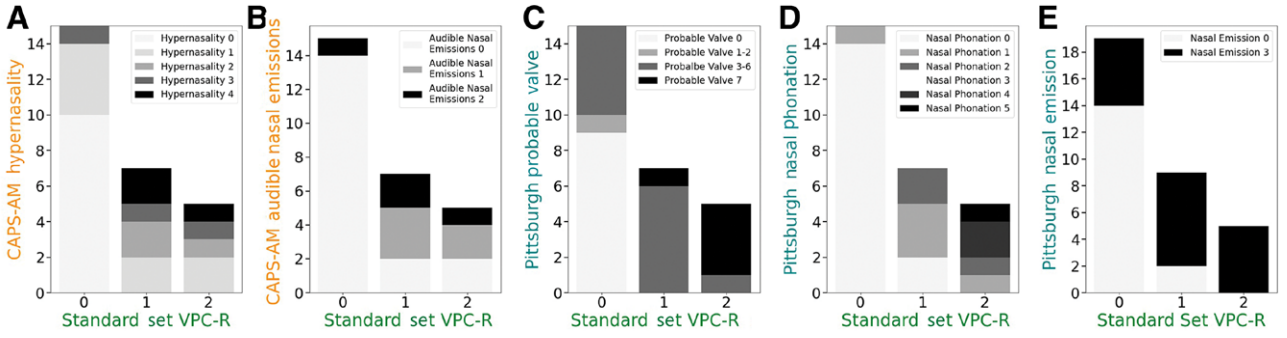


**Fig. 3.** Comparison of instruments related to velopharyngeal dysfunction. All five graphs (A–E) separate participants by VPC-R score A, (range 0–2) along the horizontal axis. The vertical axes of the graphs show the distribution of (A) CAPS-A-AM hypernasality scores (range 0–4), (B) CAPS-A-AM "audible nasal emission" scores (range 0–2), (C) PWSS's "probable nature of valve" scores (range 0–14), (D) PWSS's "nasality phonation scores" (range 0–14), and (E) PWSS's "nasal emission scores" (range 0–18) for each participant, respectively. More pathological subcomponent scores are shown in darker shades. All graphs show a general association between better (lower) VPC-R scores and better (lighter) CAPS-A-AM and PWSS scores.

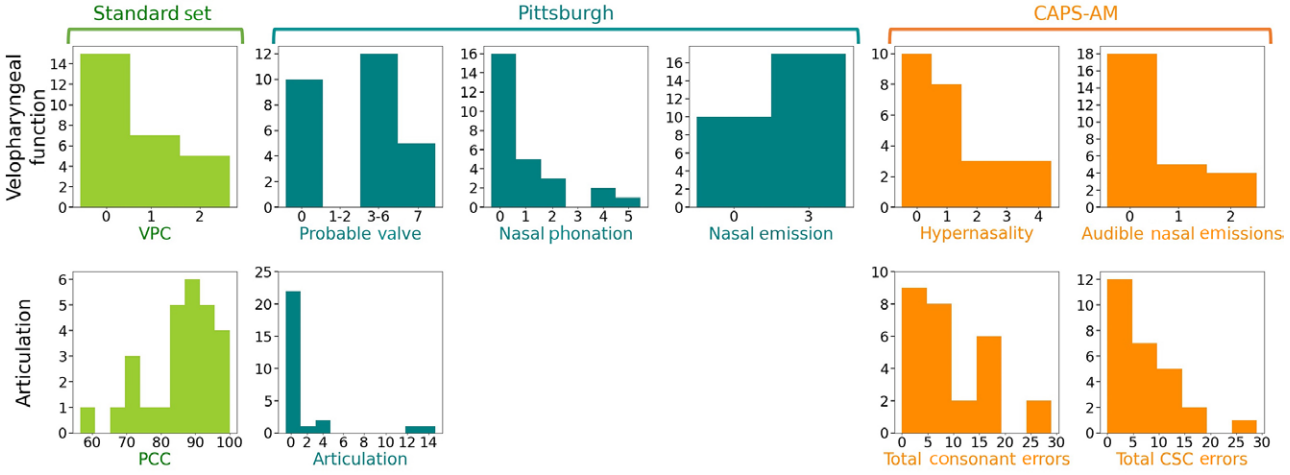


**Fig. 4.** Scoring distribution for all scales. Histograms showing distribution of scores reported for each scale. The top row shows scores that reflect velopharyngeal function. The bottom row shows scores that reflect articulation. The first column contains scores obtained during evaluation with the standard set instruments. The second, third, and fourth columns contain scores obtained during evaluation with the PWSS rating system. The fifth and sixth columns contain scores obtained during evaluation with the CAPS-A-AM rating system. Further details of score distribution are provided in **Supplemental Digital Content 3** (**http://links.lww.com/PRSGO/D5**).

### VPC-R Score May Be a Reasonable Generic "Screening Test" for the More Detailed CAPS-A-AM and PWSS

When using the composite CAPS-A-AM score as the diagnostic standard, the sensitivity and specificity of a "positive" (pathological) VPC-R score (1 or 2) were 87.5% and 73.7%. In this sample, the PPV and NPV were 58.3% and 93.3%, respectively. When using composite PWSS score as the diagnostic standard, the sensitivity of a "positive" VPC-R score was 70.6%, specificity was 100%, PPV was 100%, and NPV was 66.7%.

## DISCUSSION

It must be clearly stated that clinical speech assessment does not require the use of a scale. Many teams take excellent care of their patients using clinical perceptual assessments, without further qualifying or quantifying speech characteristics according to the rubrics of a scale. The utility of an outcome measure is two-fold: (1) standardizing an otherwise subjective process and (2) in order that comparisons may be made (eg, between raters, of the same patient at different times, or between patient groups, treatment types, or centers). Of course, measurements by any scale must be interpreted by a trained clinician to draw meaningful conclusions from the data. In many cases, the clinician may determine that further clinical investigation is warranted, beyond the scope of what is contained in the scales.

This study explored how PCC and VPC-R, the generic speech outcome measures included in the ICHOM standard set, compare with cleft-specific speech outcome measures, specifically the CAPS-A-AM and PWSS scales. The results of this study are critically important to inform future usage of the ICHOM standard set measures. If PCC and VPC-R scores were related poorly to relevant parts of their CAPS-A-AM and/or PWSS scores, the generic screening tools might be insufficient to accurately or completely measure the pathology of interest. In this case, the argument should be made for replacement (or supplementation) of these generic instruments by cleft-specific instruments. On the other hand, if this study reports sufficient correlation of PCC and VPC-R scores to the relevant parts of CAPS-A-AM and/or PWSS, this would provide reassurance for continuing routine use of these instruments for prospective outcome measurement. In either case, it is important to explore the strengths and weaknesses of the generic measures and how they relate to the cleft-specific scales.

### PCC Is an Adequate Measure of Overall Articulatory Function

PCC conceptually relates closely with CAPS-A-AM "total consonant errors," as both scales consider general articulatory imprecision regardless of etiology. For the younger participants sampled, this may include developmental errors of articulation in addition to those errors more specific their history of cleft. The moderately strong correlation between PCC and CAPS-A-AM "total consonant errors" found in this study is coherent with this conceptual similarity in scope. In contrast, CAPS-A-AM "total CSC errors" was designed to evaluate only cleft-related articulatory errors. Interestingly, however, the correlation between "total consonant errors" and "total CSC errors" in our sample was very strong (correlation coefficient 0.94, $P < 0.001$), suggesting that the majority of the consonant errors observed were cleft-specific errors rather than developmental errors. The fair correlation found between PCC and "total CSC errors" is coherent with the difference in scope between the two scales, made less apparent by fewer developmental errors. Thus, PCC performed as expected as an outcome measure of general articulatory precision but should be interpreted with an understanding of the patient's developmental stage.

PCC also relates conceptually to PWSS "articulation," although in practice, the different structure and purpose of the PWSS "articulation" scale results in poor correlation. That is, the way that the PWSS scale appraises articulatory function (more descriptive than quantifying) differs from the way that PCC and CAPS-A-AM do. Our ability to comment further is limited by extreme skew of PWSS "articulation" scores in our exploratory sample. Several other studies have also reported score distributions with a similar skew to our sample and lower mean scores for PWSS "articulation" measurement.[33,34] Most studies using the PWSS report total scores (not scores from subcomponents). Total scores typically demonstrate a normal distribution among cleft patients, whereas the individual components of the scale may not demonstrate a normal distribution.[9,35] The PWSS does not have a direct subcomponent category, and ranges for articulation scores are from 0 to 23 and may be less amenable to using subcomponents than the CAPS-A-AM scale. In summary, one may conclude that PCC is an adequate estimate of overall articulatory precision (as also measured by CAPS-A-AM "total consonants correct"), but it is an imperfect or incomplete representation of cleft-specific articulatory dysfunction (as may be better measured by CAPS-A-AM "total CSC errors").

### VPC-R Appropriately Measures Overall Velopharyngeal Competency and Some Individual Speech Pathology Constructs

VPC-R reflects the clinician's overall perception of the speaker's velopharyngeal function. In this sample, VPC-R scores correlated moderately with the PWSS "probable nature of valve." This may reflect the common scope of both these scales to encompass many signs of velopharyngeal dysfunction. VPC-R also correlated moderately with two smaller subcomponents from PWSS, "nasality phonation" and "nasal emission." Meanwhile, VPC-R correlated only fairly with CAPS-A-AM "hypernasality" and "audible nasal emission." As both the latter measures are intended to be conceptually distinct and specific, they are not expected to correlate perfectly with the more summative VPC-R score. In summary, these data suggest that VPC-R appropriately evaluates overall velopharyngeal function but may not be as useful in describing in detail the individual signs of dysfunction, such as hypernasality alone or nasal emissions alone.

### VPC-R May Be Useful as a "Screening Test" to Identify Patients Most at Need for Further Evaluation of Velopharyngeal Function

This study calculated the sensitivity and specificity of VPC-R in detecting pathology according to other

measures of velopharyngeal function—both to quantify the overlap between these scales and to explore alternative uses of this scale to cleft lip/palate teams. In clinical settings where time does not always allow for more thorough speech evaluation systems, VPC-R may be useful as an efficient "screening test" (ascertained by the speech-language pathologist after a spontaneous speech evaluation) that provides a rough estimate of what would be measured by a cleft-specific "diagnostic" system.

A VPC-R score of 1 or 2 is highly sensitive but less specific in detecting patients who will score poorly on CAPS-A-AM "hypernasality" and/or "audible nasal emissions." A VPC-R score of 1 or 2 is less sensitive but highly specific in detecting patients who will score poorly on PWSS "probable nature of valve." Considering its performance against both scales, VPC-R offers a simple, time-efficient screening test to identify patients that may need additional speech evaluation. Thus VPC-R can serve as an "entry point" to more thorough speech investigation when such a thorough evaluation is not possible for every routine cleft care appointment. If patients score very well on VPC-R, perhaps no further rating is needed, whereas if VPC-R is poor, then the team might consider more thorough characterization via CAPS-A-AM or PWSS and/or further workup via nasometry, nasoendoscopy, or videofluoroscopy.

**PWSS "Articulation" and "Nasality Phonation" Scores Were Highly Skewed**

No outcome measure is perfect. Every measurement requires a certain level of abstraction to create the model or conceptual framework for measurement. Scale development is arguably as much art as science, and myriad factors (such as accuracy, precision, completeness, feasibility, generalizability, etc.) must be balanced and vetted. Scale validation is outside the scope of this study; therefore, in this study, we accept each scale as it has previously been described and validated. Scale performance may vary based on the prevalence of clinical characteristics or confounders in the sample.

In our study, no scale had a normal distribution (Fig. 4), as all scores were skewed to some degree toward the pathological direction. This is likely related to the prevalence of speech dysfunction in this sample of participants. All scales utilized the full range of available scores except for PWSS "articulation" and "nasality phonation." These two subcomponents did not make full use of their theoretical range when applied to this sample of patients. For example, very few participants received a PWSS "articulation" score above 3, despite the maximum value being set at 23; thus, most of this scale remained unutilized. As a result, PWSS's "articulation" subcomponent was difficult to relate to either of the other measures of articulation available. This apparent underutilization of the full range in the scale may have been due to small sample size, sampling bias, or to a flaw in the scales themselves. A thorough critique of each scale from the perspective of scale design is beyond the scope of this article, but it remains an important characteristic to be considered more fully when adopting any outcome measure.

**Implication for the ICHOM Standard Set**

In this exploratory study, the generic measures, PCC and VPC-R, did not show exact correspondence with their counterparts in CAPS-A-AM or PWSS, and thus should not be considered exact substitutes for the cleft-specific speech assessments.[5,6] One might summarize that the standard set measures provide more of a snapshot of speech dysfunctions in articulation and velopharyngeal function. For more detail, as would be required for clinical care or detailed studies, further in-depth evaluation would be justified. All patients with concern for speech dysfunction would likely still benefit from more detailed, cleft-specific methods of describing and documenting their speech pathology.

Nonetheless, both PCC and VPC-R seem to be useful. PCC is reflective enough of cleft-specific articulatory ratings in CAPS-A-AM that it can be considered a practical outcome measure for dissemination and implementation across teams. Despite its simplicity, VPC-R seems to appropriately measure overall velopharyngeal function, if not specific characteristics of velopharyngeal dysfunction. VPC-R was found to function well as a "screening test" for the more detailed and specific CAPS-A-AM and PWSS evaluations. The fact that both PCC and VPC-R are available for many languages has the added benefit of permitting cross-linguistic comparisons—one of the major goals and design parameters for ICHOM. Recent research has found that English-speaking SLPs have limited potential in assessing cleft-specific speech characteristics in a foreign language.[36] PCC and VPC-R can be used for comparison between languages when the assessment is based on a wordlist designed for cross-linguistic comparison following international guidelines.[17,37] In this way, SLPs can uses these tools in their own languages to assess patients, and these outcomes can be compared. In contrast, CAPS-A, CAPS-A-AM, and PWSS are limited to use in English-speaking populations.[4–6,38]

This study provides reassurance to teams that have adopted the ICHOM standard set that PCC and VPC-R are useful and appropriate measures of articulation and velopharyngeal function, respectively. These instruments can be used as the basis for longitudinal and cross-sectional (eg, intercenter) comparisons. For teams in the United States, ICHOM might consider extending the standard set measures by way of adding CAPS-A-AM or PWSS. ICHOM includes a scientific advisory council that may propose such a recommendation, and its stewardship committee may determine whether this guideline remain optional or required for participating teams.

**Limitations**

There are several limitations to this study. All scales used in this study involve subjective assessments, and thus, their conceptual overlap is not perfectly defined. Our group was thoughtful in the selection of appropriate comparisons between the subcomponents of the scales; however, given that similar comparisons have not yet been made, different professionals may argue that other subcomponents of the individual assessment, or "composite scores" thereof, may have been better. Additionally,

although the PWSS scale assigns scores to individual signs of velopharyngeal function, the scale was developed to provide an overall assessment of the probable nature of a velopharyngeal valve.[6,38] Another limitation is that we only used one rater for the CAPS-A-AM and PWSS scales, whereas we had two for the ICHOM measures, so we did not also evaluate intrarater and interrater reliability of the CAPS-A-AM and PWSS systems. Additionally, a small sample size was utilized for this preliminary study. Despite a large range of phenotypes identified in the cohort, there was not a normal distribution of phenotypes, which may have affected overall results, and subanalysis of each phenotype could not be conducted due to small sample size. Three speech-language pathologists on the study team worked at the study institution and may have potentially seen subjects prior, which may have introduced bias to the study. A final limitation of this study was that it evaluated only English-speaking patients. If the formal speech assessments, such as CAPS-A-AM and PWSS could be translated, validation from studies focusing on different languages should be performed. Speech pathology detected by any outcome measures always warrants further investigation by a trained clinician using the speech evaluation tools they feel most comfortable with.

## CONCLUSIONS

In this exploratory study, the standard set measures of PCC and VPC-R performed moderately well when compared with their cleft-specific measurement counterparts. Based on this study, our group hypothesizes that these standard set measures would perform sufficiently as standardized outcome tools for the assessment of articulation and velopharyngeal dysfunction when implemented broadly within cleft and palate teams; however, larger randomized studies are needed to validate these results.

***Alexander C. Allori, MD, MPH***
Division of Plastic, Maxillofacial, and Oral Surgery
Department of Surgery
Duke University Health System
DUMC 3974, 40 Medicine Circle
Durham, NC 27705
E-mail: alexander.allori@duke.edu

## DISCLOSURES

*The authors have no financial interest to declare in relation to the content of this article. Del Risco was supported by a National Institutes of Health R25 grant (1R25DC020172-01).*

## ACKNOWLEDGMENT

*We would like to thank Kristen DeLuca, MS, CCC-SLP, for her courteous external review of the article.*

## REFERENCES

1. Scherer NJ, D'Antonio LL. Parent questionnaire for screening early language development in children with cleft palate. *Cleft Palate Craniofac J.* 1995;32:7–13.
2. Kuehn DP, Moller KT. Speech and language issues in the cleft palate population: the state of the art. *Cleft Palate Craniofac J.* 2000;37:1–35.
3. Parameters for evaluation and treatment of patients with cleft lip/palate or other craniofacial differences. *Cleft Palate Craniofac J.* 2018;55:137–156.
4. John A, Debbie S, Sweeney T, et al. The cleft audit protocol for speech-augmented: a validated and reliable measure for auditing cleft speech. *Cleft Palate Craniofac J.* 2006;43:272–288.
5. Chapman KL, Baylis A, Trost-Cardamone J, et al. The Americleft speech project: a training and reliability study. *Cleft Palate Craniofac J.* 2016;53:93–108.
6. McWilliams B, Philips B. Audioseminars in speech pathology, velopharyngeal incompetence. *Phila PA Saunders.* 1979;1:1–134.
7. Georgievska-Jancheska T, Gjorgova J, Popovska M. The role of the velopharyngeal sphincter in the speech of patients with cleft palate or cleft lip and palate using perceptual methods. *Open Access Maced J Med Sci.* 2016;4:674–679.
8. Nayar HS, Cray JJ, MacIsaac ZM, et al. Improving speech outcomes after failed palate repair: evaluating the safety and efficacy of conversion furlow palatoplasty. *J Craniofac Surg.* 2014;25:343–347.
9. Afrooz PN, MacIsaac Z, Rottgers SA, et al. A comparison of speech outcomes using radical intravelar veloplasty or furlow palatoplasty for the treatment of velopharyngeal insufficiency associated with occult submucous cleft palate. *Ann Plast Surg.* 2015;74:182–186.
10. Dudas JR, Deleyiannis FWB, Ford MD, et al. Diagnosis and treatment of velopharyngeal insufficiency: clinical utility of speech evaluation and videofluoroscopy. *Ann Plast Surg.* 2006;56:511–517; discussion 517.
11. Pereira VJ, Tuomainen J, Lee KYS, et al. A perceptual outcome measure of velopharyngeal function based on the cleft audit protocol for speech-augmented (CAPS-A VPC-Sum): validation through a speech osteotomy study. *Int J Lang Commun Disord.* 2021;56:754–767.
12. Bruneel L, Bettens K, Bodt MD, et al. Stages in the development and validation of a Belgian Dutch outcome tool for the perceptual evaluation of speech in patients with cleft palate. *Cleft Palate Craniofac J.* 2019.
13. Sell D, John A, Harding-Bell A, et al. Cleft audit protocol for speech (CAPS-A): a comprehensive training package for speech analysis. *Int J Lang Commun Disord.* 2009;44:529–548.
14. Henningsson G, Kuehn DP, Sell D, et al; Speech Parameters Group. Universal parameters for reporting speech outcomes in individuals with cleft palate. *Cleft Palate Craniofac J.* 2008;45:1–17.
15. Kummer AW, Hosseinabad HH, Redle E, et al. Protocols for reporting speech outcomes following palatoplasty or velopharyngeal surgery: a literature review. *Plast Reconstr Surg Glob Open.* 2019;7:e2151.
16. Kummer AW, Clark SL, Redle EE, et al. Current practice in assessing and reporting speech outcomes of cleft palate and velopharyngeal surgery: a survey of cleft palate/craniofacial professionals. *Cleft Palate Craniofac J.* 2012;49:146–152.
17. Allori AC, Kelley T, Meara JG, et al. A standard set of outcome measures for the comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2017;54:540–554.
18. Shriberg LD, Austin D, Lewis BA, et al. The percentage of consonants correct (PCC) metric: extensions and reliability data. *J Speech Lang Hear Res.* 1997;40:708–722.
19. Lohmander A, Hagberg E, Persson C, et al. Validity of auditory perceptual assessment of velopharyngeal function and dysfunction—the VPC-sum and the VPC-rate. *Clin Linguist Phon.* 2017;31:589–597.
20. McLeod S, Crowe K, Shahaeian A. Intelligibility in context scale: normative and validation data for English-speaking preschoolers. *Lang Speech Hear Serv Sch.* 2015;46:266–276.
21. McLeod S. Intelligibility in context scale: cross-linguistic use, validity, and reliability. *Speech Lang Hear.* 2020;23:9–16.

22. Goldman R, Fristoe M. Goldman-Fristoe test of articulation 3. Available at https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Developmental-Early-Childhood/Goldman-Fristoe-Test-of-Articulation-3/p/100001202.html. Accessed August 15, 2022.
23. Fitzpatrick B, Coad J, Sell D, et al. Assessing speech at three years of age in the cleft palate population: a scoping review of assessment practices. *Int J Lang Commun Disord.* 2020;55:165–187.
24. Gart M, Gosain A. Diagnosis and management of velopharyngeal insufficiency following cleft palate repair. *J Cleft Lip Palate Craniofacial Anom.* 2014;1:4–10.
25. Chan YH. Biostatistics 104: correlational analysis. *Singapore Med J.* 2003;44:614–619.
26. Van Rossum G, Drake FL. *Python 3 Reference Manual.* CreateSpace; 2009.
27. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9:90–95.
28. Virtanen P, Gommers R, Oliphant TE, et al; SciPy 1.0 Contributors. SciPy 10: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–272.
29. Vallat R. Pingouin: statistics in Python. *J Open Source Softw.* 2018;3:1026.
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–2830.
31. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6:284–290.
32. Cohen J. A coefficient of agreement for nominal scales. 1960. Available at https://journals.sagepub.com/doi/10.1177/001316446002000104. Accessed October 30, 2022.
33. Thongprayoon S, Liadprathom K, Chuangsuwanich A, et al. Speech outcome analysis after primary cleft palate repair: interim Siriraj hospital audit. *Siriraj Med J.* 2021;73:744–751.
34. Swanson JW, Mitchell BT, Cohen M, et al. The effect of furlow palatoplasty timing on speech outcomes in submucous cleft palate. *Ann Plast Surg.* 2017;79:156–161.
35. Losee JE, Smith DM. Acellular dermal matrix in palatoplasty. *Aesthet Surg J.* 2011;31(7_Suppl):108S–115S.
36. Stoehr JR, Park E, Reddy NK, et al. The feasibility of cross-linguistic speech evaluation in the care of international cleft palate patients. *J Craniofac Surg.* 2022;33:1413–1417.
37. Lohmander A, Willadsen E, Persson C, et al. Methodology for speech assessment in the Scandcleft project—an international randomized clinical trial on palatal surgery: experiences from a pilot study. *Cleft Palate Craniofac J.* 2009;46:347–362.
38. McWilliams BJ, Morris HL, Shelton RL. Cleft Palate Speech (2nd ed). Philadelphia: BC Decker; 1990.