Stages in the development and validation of a Belgian Dutch outcome tool for the perceptual evaluation of speech in patients with cleft palate

Laura Bruneel, PhD[1,*], Kim Bettens, PhD[1], Marc De Bodt, PhD[1, 2], Evelien D'haeseleer, PhD[1], Zoë Thijs, MSc[1], Nathalie Roche, MD, PhD[3], Kristiane Van Lierde, PhD[1, 4]

[1]Department of Rehabilitation Sciences, Language, and Hearing Sciences, Ghent University, Ghent, Belgium

[2]Department of ENT, Head and Neck Surgery and Communication Disorders, Antwerp University, Wilrijk, Belgium

[3]Department of Plastic and Reconstructive Surgery, Ghent University Hospital, Ghent, Belgium

[4]Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa

*Corresponding Author:

Laura Bruneel, Department of Rehabilitation Sciences, Ghent University, C. Heymanslaan 10 2P1, 9000 Ghent, Belgium. Email: laubrune.bruneel@ugent.be

Abstract

Objective: To develop and validate a Belgian Dutch outcome tool for the perceptual evaluation of speech in patients with cleft palate.

Setting: Cleft palate team in a tertiary university hospital

Methods: The tool was based on the CAPS-A (John et al., 2006; Sell et al., 2009), with adaptations to some of the speech variables and the structured listening protocol. Following a preliminary listening experiment in phase 1, the tool was optimized. In a second phase, a listening experiment with four experienced listeners was set up to assess face validity, inter- and intra-rater reliability and criterion validity.

Results: Results of phase 1 indicated good to very good inter- and intra-rater reliability for the majority of the speech variables, good discriminant validity, and varying sensitivity and specificity based on a comparison with nasalance values and the NSI 2.0 (criterion validity). Results of phase 2 showed good to very good inter-rater reliability for five of the 14 variables and good intra-rater reliability in three of the four experienced listeners. Sensitivity and specificity were sufficient, except the specificity of the hypernasality judgments in comparison with the nasalance values of the oral text. Overall, listeners positively judged the face validity of the tool.

Conclusion: The two-phase evaluation indicated varying validity and reliability results. Future studies will aim to optimize validity and reliability of the developed tool based on adaptations to listening protocol, the addition of speech variables and the inclusion of a more elaborate training.

Key words: speech perception; articulation; nasality

Introduction

In the past decades, great attention has been paid to define outcomes and related measurements of health conditions. This is an important field of study, as outcome measurements allow for quality control, and consequently quality improvement, evidence-based research and value-based healthcare reform (Allori et al., 2017). As a minimum, outcome measurements should be valid, reliable and practical to implement (Sitzman et al., 2014). Specifically for patients with cleft palate, speech is a crucial outcome. Through several working groups, great effort has been made to standardize speech outcome assessment (Brondsted et al., 1994; Harding et al., 1997; Hutters and Henningsson, 2004; John et al., 2006; Henningsson et al., 2008; Lohmander et al., 2009; Sell et al., 2009; Chapman et al., 2016). These efforts focused

on the perceptual evaluation of speech, as yet no instrumental analysis can surpass the clinical relevance of these auditory (and visual) assessments (Kuehn and Moller, 2000).

Speech variables and measurements are key concepts of a speech outcome tool that should be clearly stated, defined, and evaluated in a valid and reliable manner (Lohmander and Olsson, 2004; Sell, 2005). Three recent tools and/or methodologies specifically designed for outcome studies attracted international attention: the Cleft Audit Protocol for Speech – Augmented (CAPS-A) (John et al., 2006; Sell et al., 2009), the methodology designed for speech assessment in the Scandcleft project (Lohmander et al., 2009; Lohmander et al., 2017a; Willadsen et al., 2017), and the Universal Parameters for Reporting Speech Outcomes in Individuals with Cleft Palate (UPS) (Henningsson et al., 2008). The CAPS-A has been used for both audit tool and research purposes in the U.K. (Britton et al., 2014; Sell et al., 2015; Sell et al., 2017) and has been adapted to American English (CAPS-A-AM) (Chapman et al., 2016). The methodology for speech assessment in the Scandcleft project was composed with the aim of cross-linguistic comparison of speech outcomes to evaluate different surgical techniques for primary palatal repair (Lohmander et al., 2017a; Willadsen et al., 2017), and is currently being used in the Timing Of Primary Surgery in cleft palate (TOPS) project (https://www.tops-trial.org.uk/). Henningsson et al. (2008) proposed the UPS, a universal system for speech outcome measurements, including guidelines for patient identification and speech sampling, definitions and descriptions of speech outcome parameters, and a conversion procedure to adapt current outcome tools to the UPS. However, the validity and reliability of this universal system have not been evaluated yet, and the proposed conversion procedure raised some concerns regarding validity, using previously collected data that might be unreliable (Lohmander, 2008). The eventual aim of standardized speech assessment is the comparison of outcomes within and between centers. However, such inter-center comparisons may require adjustments for language as a background variable (Brondsted et al., 1994; Henningsson and Hutters, 1997;

Hutters and Henningsson, 2004). Based on the efforts of the Scandcleft Speech Group, recommendations for the construction of speech samples for cross-linguistic comparisons were described (Hutters and Henningsson, 2004; Lohmander et al., 2009). The importance of phonetically identical sound inventories that control for high-pressure consonants and high vowels was emphasized, given the vulnerability of these sounds for speech disorders resulting from the cleft condition (Moon et al., 1994; Kuehn and Moon, 1998; Peterson-Falzone et al., 2001). Recently the multidisciplinary ICHOM working group proposed a minimum standard set of outcome measurements in patients with cleft palate (Allori et al., 2017). For speech outcomes, the group proposed the use of the Intelligibility in Context Scale (McLeod et al., 2012), articulatory proficiency expressed in percent consonants correct (Shriberg et al., 1984), perceptually evaluated velopharyngeal competence (Lohmander et al., 2009), and the Speech subscale of the CLEFT-Q (Wong Riff et al., 2017). Collection of a standardized speech sample as described by Hutters and Henningsson (2004) (https://clispi.org/) was highly recommended.

Specifically for Dutch, two tools have been developed over the years. Dutch is spoken in both the Netherlands and Belgium, resulting in two institutionalized versions of the standard language with mainly phonetic differences (Verhoeven, 2005). In 2006 the Screeningsinstrument Schisis Leuven (SISL) for clinical speech assessment in Belgian Dutch, often referred to as Flemish, was developed, partially based on the GOS.SP.ASS ’98 (Sell et al., 1999; Breuls et al., 2006). However, validity and/or reliability results have not been reported, neither when using this tool in outcome studies (Vander Poorten et al., 2006; Samoy et al., 2015). Recently the Netherlandic Dutch Cleft Speech Evaluation Test (DCSET) was presented, showing varying inter- and intra-rater reliability results (Spruijt et al., 2018). The analysis was based on 130 evaluated items per speech sample (e.g. separate ratings of hypernasality for each oral, nasal or oronasal sentence), which raises questions regarding the unreported validity and time efficiency. Moreover, definitions of the evaluated variables were

missing. An algorithm was proposed to transpose DCSET results to the UPS (Henningsson et al., 2008), but results on the validity and/or reliability of this conversion procedure were lacking.

Given the lack of a valid and reliable outcome tool for the perceptual evaluation of speech in Belgian Dutch-speaking patients with cleft palate, the purpose of this study was to develop and validate an outcome tool based on the CAPS-A (John et al., 2006; Sell et al., 2009). In a first phase, the tool was constructed and optimized based on a preliminary listening experiment. In a second phase, a listening experiment with experienced listeners was set up to evaluate the final version of the tool.

Methods

This research (phase 1 and phase 2) was approved by the Committee of Ethics of the Ghent University Hospital (2016/0338; 2016/1139).

Phase 1: construction and first evaluation of the outcome tool

The tool, including the speech sample, speech variables and listening protocol was constructed in consultation with the co-authors, who all have extensive experience in patients with cleft palate (six speech therapists and one plastic surgeon).

Construction of the Belgian Dutch speech sample

Following the CAPS-A, a speech sample consisting of spontaneous speech, automatic rote speech and sentences was composed (Sell et al., 2009). Automatic rote speech included counting from 1 to 10 and reciting the days of the week in patients younger than seven years old. Older patients were asked to count from 1 to 20 and 60 to 70, and to recite the days of the week. Sentences were constructed following the guidelines described by Henningsson et al. (2008). Thirteen sentences targeting the pressure consonants of the Belgian Dutch sound system

Table 1. Belgian Dutch Sentences Constructed for the Evaluation of Resonance, Nasal Airflow, and Consonant Production.

Target Consonant	Belgian Dutch Sentence	Phonetic Transcription	English Translation
[p]	Papa riep opa.	[**p**ɑ**p**ɑ ri'**p** o:**p**ɑ]	Daddy called grandpa.
[b]	Bel boer Robbe.	[**b**ɛl **b**u'r rə**b**ə]	Call farmer Robbe.
[t]	Tuur eet later.	[**t**y'r e:**t** la:**t**ər]	Tuur eats later.
[d]	Lode doet de deur toe.	[lo:**d**ə **d**u't **d**ə **d**ø:r tu']	Lode closes the door.
[k]	De cake rook lekker.	[də **k**e:**k** ro:**k** lɛ**k**ər]	The cake smelled good.
[f]	De toffe fee is lief.	[də tɔ**f**ə **f**e: ɪs li'**f**]	De nice fairy is sweet.
[v]	Eva viel voorover.	[e:**v**ɑ **v**il **v**o:ro:**v**ər]	Eva tripped.
[s]	Sara wil de losse jas.	[**s**a:rɑ wɪl də lɔ**s**ə jɑ**s**]	Sara wants the loose coat.
[z]	Ze ziet de roze zoo.	[**z**ə **z**it də ro:**z**ə **z**o:]	She sees the pink zoo.
[x]	Ik goochel graag.	[ɪk ɣo:**x**əl ɣra:**x**]	I like magic.
[ɣ]	De egel gilt.	[də e:**ɣ**əl **ɣ**ɪlt]	The hedgehog screams.
[ʃ]	Liesje showt de sjaal.	[li**ʃ**ə **ʃ**o:wt də **ʃ**a:l]	Liesje shows the scarf.
[ʒ]	De logé wou gelei.	[də lo:**ʒ**e: wɑu **ʒ**əlɛ^{j}]	The guest wants jelly.
Nasal sentences	Neem mama maar mee.	[**n**e:**m** **m**ɑ**m**ɑ **m**a:r **m**e:]	Take mommy with you.
	Oma nam een oranje mand.	[o:**m**ɑ **n**ɑ**m** ə**n** o:ra**ɲ**ə **m**ɑ**n**t]	Grandma took an orange basket.
	Noem namen van mannen.	[**n**u'**m** **n**a:**m**ə**n** vɑ**n** **m**ɑ**n**ə**n**]	Call names of men.
	Mama kamt de lange manen.	[**m**ɑ**m**ɑ kɑ**m**t də lɑ**ŋ**ə **m**a:**n**ə**n**]	Mommy combs the long mane.
Sentences without high-pressure consonants or high vowels	Wil jij haar rol?	[wɪl jɛ^{i} ha:r rəl]	Do you want her role?
	Hoor jij Lara?	[ho:r jɛ^{i} la:rɑ]	Do you hear Lara?
Sentence with s-clusters	Stella speelt liever de heks.	[**st**lɑ **sp**e:lt li'vər də hɛ**ks**]	Stella prefers to play the witch.

were constructed (table 1). Sentences controlling for pressure consonants facilitate the evaluation of cleft speech, given the vulnerability of these sounds for speech disorders resulting from the cleft condition (Watson et al., 2001; Peterson-Falzone, 2006; Henningsson et al., 2008). Following Henningsson et al. (2008), sentences should contain ten words with high vowels only. As great importance was given to the feasibility of reciting these sentences by using high frequency words, the thirteen sentences included eight words with high vowels only ([i], [ʏ:],[u]) and two words with a high vowel and a mid-vowel. In addition, four sentences loaded with nasal consonants and no pressure consonants (when possible) were composed to facilitate the evaluation of hyponasality (Henningsson et al., 2008). Lastly, two sentences without pressure consonants or high vowels, and one sentence with s-clusters were constructed in alignment with the GOS.SP.ASS sentences and additions for CAPS-A outcome purposes (Sell et al., 1999; Sell et al., 2009). High-quality audio and audiovisual recordings were made of each speech sample using a unidirectional condenser microphone (Samson C01U) and a Sony Handycam HDR-CQ280E (Gooch et al., 2001).

Construction of the speech outcome tool - speech variables and definitions

The majority of the CAPS-A variables and their accompanying scales and definitions (John et al., 2006; Sell et al., 2009) were translated to Belgian Dutch: 'hypernasality', 'hyponasality', 'nasal emission', 'nasal turbulence', the framework of cleft-related speech characteristics (CSCs), 'grimace', and the need for SLT intervention. The expert panel described above acknowledged the appropriateness of the scales and definitions for use in the Belgian Dutch tool. For the translation of the parameters and the definitions of the CAPS-A scalar points to Belgian Dutch, a forward-backward translation procedure was used to ensure a content equivalency, with special attention given to the wording (Wild et al., 2005). Two Belgian Dutch-speaking speech-language pathologists, with professional proficiency in English and experience in the field of cleft palate, independently conducted the forward translation from

English to Belgian Dutch. Subsequently, a Belgian Dutch consensus translation was agreed and was then translated back to English by another Belgian Dutch-speaking speech-language pathologist with professional English proficiency. After comparing the backward translation with the original definitions, a definitive translation was constructed.

Some minor adaptations to the CAPS-A variables were made. Regarding the speech variable 'voice', the CAPS-A makes a distinction between 'normal' and 'abnormal or distinctive voice quality'. To improve clarity, the current tool distinguishes between 'normal' and 'a dysphonic voice, possibly influencing the perception of resonance' (Kuehn and Moller, 2000; Sell et al., 2009). Similar to the CAPS-A-AM (Chapman et al., 2016), no distinction between passive and active nasal fricatives was made, as these speech disorders are difficult to distinguish perceptually (Harding and Grunwell, 1998) and no differential diagnosis using nose holding was included (John et al., 2006). The variables 'speech understandability' and 'speech acceptability' described by Henningsson et al. (2008) replaced the CAPS-A variable 'speech intelligibility/distinctiveness'. The latter variable describes a combined evaluation of the variables speech intelligibility and distinctiveness, although these are conceptually different (Whitehill, 2002; Henningsson et al., 2008). Furthermore, Castick et al. (2017) reported good validity and reliability of the UPS scales for speech understandability and speech acceptability.

Construction of the speech outcome tool - structured listening protocol

Some important changes to the structured listening protocol of the CAPS-A (Sell et al., 2009) were made. The sentences of the current tool were constructed to provide a balanced speech sample, specifically controlling for high-pressure consonants and high vowels (Henningsson et al., 2008). Therefore, hypernasality, hyponasality, nasal emission and nasal turbulence were evaluated based on the sentences alone, whereas in the CAPS-A these evaluations are based on automatic rote speech and sentence repetition (Sell et al., 2009). Furthermore, it can be difficult to elicit standardized automatic rote speech in younger children,

Table 2. Structured Listening Protocol.

	Speech Sample	Speech Variables Rated on Ordinal Scales		
1.	Audio: spontaneous speech	Speech understandability Consonant production	Voice	Speech acceptability Non-CP ± L-related SLT intervention CP ± L-related SLT intervention
2.	Audio: automatic rote speech	Consonant production		
3.	Audio: sentences	Consonant production (CSCs) Resonance: hypernasality and hyponasality Nasal airflow: nasal emission and nasal turbulence		
4.	Video: spontaneous speech, automatic rote speech	Revision of consonant production (visual aspects)	Grimace	
5.	Video: sentences	Revision consonant production (CSCs) (visual aspects) Revision resonance and nasal airflow (when necessary)		

Abbreviation: CSCs, cleft-related speech characteristics.

e.g. some children start singing songs, or are not able to produce the sequence independently. The repetition of sentences allows for a more structured and standardized speech sample, also controlling speech rate (Sell, 2005). In alignment with the CAPS-A, speech understandability was rated based on spontaneous speech. Consonant production was transcribed phonetically based on spontaneous speech, automatic sequences and the sentences. Transcription of consonant errors based on spontaneous speech and automatic rote speech provided the listener an overall insight in the child's consonant production. Only the transcription based on the sentences was taken into account for the categorization of the CSCs. Speech samples were evaluated in a specific order (table 2) as the interaction between speech variables can influence judgements (Kent, 1996; Sell et al., 2009).

Preliminary evaluation of the Belgian Dutch speech outcome tool - listening experiment

The first author (L.B.), a speech pathologist with four years of experience in patients with cleft palate, and a master's student in speech language pathology (Z.T.) performed a first evaluation of the tool. This evaluation included an assessment of practical aspects, such as the composition of the rating form, and an evaluation of the inter- and intra-rater reliability, discriminant validity and criterion validity. Before the listening experiment, the first author set up a training of approximately two hours. During this training, the speech variables and structured listening protocol described above were presented, supported by edited speech samples to illustrate each speech variable and their scalar points. Additionally, one and a half hours was spent on consensus listening during which complete speech samples of five cases were evaluated.

To determine inter-rater reliability, 20 speech samples of patients with an isolated cleft palate with or without cleft lip (CP±L) (mean (*M*) age = 6.5 years (y), standard deviation (*SD*) = 2.40) were evaluated. Additionally, ten samples of controls without CP±L (*M* age = 5.9 y, *SD* = 1.73) or any other craniofacial anomaly, matched for age and gender with ten of the patients

with CP±L, were evaluated to determine discriminant validity. All children were between 3 and 10 years old, had Belgian Dutch as their native language, no moderate or severe hearing loss, and no cognitive impairment or neurological deficit. Speech samples of the patients with CP±L represented a range of severity. Ten speech samples of patients with CP±L were rerated in a different randomized order after one week to determine intra-rater reliability. All evaluations were conducted independently using over-ear headphones (Sennheiser EH 150 and Sennheiser Momentum).

To determine criterion validity, perceptual ratings of the speech variables hypernasality and hyponasality were compared with nasalance values and the NSI 2.0. Nasalance values of an oral, oronasal and nasal text (Van de Weijer and Slis, 1991) of each participant were determined using the Nasometer^{TM} II model 6450. Additionally, the Nasality Severity Index (NSI) 2.0 (Bettens et al., 2016), a multiparameter index for hypernasality, was calculated. First, a sustained production of the vowel [i:] of minimum two seconds was recorded with a unidirectional condenser microphone (Samson C01U) using Praat-software, version 6.0.14 (Boersma and Weenink). Based on this sample, the "voice low tone to high tone ratio" (VLHR) was determined. Subsequently, the NSI 2.0 was calculated using the NSI 2.0 Praat script, based on the VLHR and the nasalance values of [u:] and the oral text.

Phase 2: evaluation of the developed speech outcome tool – listening experiment with experienced listeners

Listening experiment

Four SLPs with a minimum of six years of experience in the evaluation of speech in patients with cleft palate were included. They were familiar with ordinal scaling and already had a global understanding of the CAPS-A. Three of the four listeners work in the same cleft team. Before conducting the listening experiment, the tool was presented to the SLPs including

the description and explanation of the definitions, the rating scales and the structured listening protocol (table 2). Edited speech samples were used to illustrate the speech variables and their scalar points. Subsequently, consensus listening of four cases was performed. In total, approximately four hours were spent on the presentation and consensus listening.

Thereafter, listeners were asked to independently rate ten speech samples (9 CP±L and 1 without CP±L (*M* age = 6.0 y, *SD* = 2.00)) in a blind and randomized order while listening through over-ear headphones (Sennheiser EH 150 and Sennheiser Momentum). These samples were part of the 30 speech samples used in phase 1. Ratings were administered on the forms that were optimized following phase 1. For the evaluation of speech understandability, the sample of spontaneous speech was edited using Praat software version 6.0.14 (Boersma and Weenink), resulting in speech samples with a length of approximately 60 syllables, in accordance with the smallest speech sample collected, and without utterances of the conversation partner.

All samples could be replayed as much as needed, with the exception of the sample of spontaneous speech for the evaluation of speech understandability. Listeners were allowed to consult the handouts of the presentation and an overview of the structured listening protocol. Speech samples were rerated in a different randomized order one month after the first evaluation to determine intra-rater reliability. By means of a questionnaire consisting of open questions, the listeners were requested to judge the acceptability of the tool for speech outcome assessment and to provide feedback on the presentation, the training session, the rating form, and the content of the tool (speech variables, rating scales, listening protocol). The sensitivity and specificity of the outcome tool were determined by comparing the perceptual ratings of hypernasality and hyponasality with the subject's nasalance values for an oral, oronasal and nasal text (Van de Weijer and Slis, 1991), and the NSI 2.0 (Bettens et al., 2016).

Statistical analysis phase 1 and 2

Statistical analyses were performed using SPSS software version 25 (SPSS Inc., Chicago, Illinois). Inter- and intra-rater reliability were determined by means of two-way mixed intra-class single measures correlation coefficients (ICC) with consistency agreement, and interpreted following the classification from Landis and Koch (1977) adapted by Altman (1990). As limited variance might cloud the interpretation of ICC measurements, the percentage of absolute agreement was calculated additionally (Hallgren, 2012). For inter-rater reliability, the mean percentage of agreement was calculated based on the mean level of agreement across all pairs of listeners (Fleiss, 1971). The Fisher's exact test ($\alpha = 0.05$) was used to assess the tool's discriminant validity. Perceptual ratings of hypernasality and hyponasality, and the nasalance values and the NSI 2.0 were recoded to determine the criterion validity. For hypernasality, ratings on the ordinal scale of 0 ('absent hypernasality') or 1 ('borderline hypernasality') were coded as absent hypernasality. Ratings for hypernasality of 2 or more were coded as present hypernasality. Regarding hyponasality, a rating of 0 was coded as absent hyponasality whereas ratings of 1 and 2 were coded as present hyponasality. Nasalance values were considered hypernasal or hyponasal when they crossed the 95% confidence interval of the normative values (Bettens et al., 2017). An NSI 2.0 value can be negative or positive, with a negative value indicating the presence of hypernasality. Hence, NSI 2.0 values were coded as either hypernasal or normal. Thereafter the sensitivity, specificity and the percentage of absolute agreement of the perceptual evaluations were calculated for each listener individually. As a summary, the median results for sensitivity, specificity and percentage of absolute agreement are presented.

Table 3. Interrater Reliability of the 2 Listeners for the Evaluation in Phase 1.

	Single Measures ICC Type Consistency	95% CI ICC	Interpretation ICC[a]	% Absolute Agreement
Speech understandability	0.87	0.70 to 0.95	Very good	65
Anterior oral CSCs	0.10	–0.35 to 0.51	Poor	40
Posterior oral CSCs	0.70	0.39 to 0.87	Good	70
Nonoral CSCs	0.73	0.44 to 0.88	Good	75
Passive CSCs	0.64	0.28 to 0.84	Good	60
Hypernasality	0.75	0.47 to 0.89	Good	55
Hyponasality	0.49	0.07 to 0.76	Moderate	90
Nasal emission	0.51	0.09 to 0.77	Moderate	60
Nasal turbulence	0.90	0.76 to 0.96	Very good	85
Voice	0.66	0.31 to 0.85	Good	95
Grimace	0.00	–0.43 to 0.43	Poor	90
Speech acceptability	0.82	0.59 to 0.92	Very good	45
Need for SLT intervention	0.58	0.19 to 0.81	Moderate	80

Abbreviations: CI, confidence interval; CSCs, cleft-related speech characteristics; ICC, intraclass single measures correlation coefficients.
[a]Based on Altman (1990): ICC < 0.20: poor, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80: good, 0.81-1.00: very good.

Table 4. Intrarater Reliability of the 2 Listeners for the Evaluation in Phase 1.

	Listener 1				Listener 2			
	ICC	95% CI	Interpretation[a]	% Agreement	ICC	95% CI	Interpretation[a]	% Agreement
Speech understandability	0.81	0.41-0.95	Very good	60	0.94	0.78-0.99	Very good	80
Anterior oral CSCs	0.89	0.63-0.97	Very good	90	0.62	0.03-0.89	Good	70
Posterior oral CSCs	0.92	0.70-0.98	Very good	90	1.00	1.00-1.00	Very good	100
Nonoral CSCs	1.00	1.00-1.00	Very good	100	0.84	0.49-0.96	Very good	80
Passive CSCs	0.74	0.24-0.93	Good	70	1.00	1.00-1.00	Very good	100
Hypernasality	0.83	0.46-0.96	Very good	60	0.87	0.57-0.97	Very good	70
Hyponasality	1.00	1.00-1.00	Very good	100	1.00	1.00-1.00	Very good	100
Nasal emission	0.83	0.45-0.95	Very good	70	0.82	0.44-0.95	Very good	80
Nasal turbulence	0.77	0.31-0.94	Good	70	0.84	0.49-0.96	Very good	80
Voice	1.00	1.00-1.00	Very good	100	1.00	1.00-1.00	Very good	100
Grimace	–	–	–[b]	100	1.00	1.00-1.00	Very good	100
Speech acceptability	0.70	0.18-0.92	Good	50	0.66	0.10-0.90	Good	50
Need for SLT intervention	1.00	1.00-1.00	Very good	100	0.76	0.28-0.93	Good	90

Abbreviations: CI, confidence interval; CSCs, cleft-related speech characteristics; ICC, intraclass single measures correlation coefficients.
[a]Based on Altman (1990): ICC < 0.20: poor, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80: good, 0.81-1.00: very good.
[b]ICC impossible to calculate due to zero variance or negative covariance.

Results

Phase 1: Preliminary evaluation of the Belgian Dutch speech outcome tool - listening experiment

Inter-rater reliability was good to very good for the majority of the parameters (8/13) (table 3), except for the variables 'anterior oral CSCs', 'hyponasality', 'nasal emission', 'grimace' and the need for SLT intervention. Intra-rater reliability ranged from good to very good for all variables (table 4). Furthermore, definition and ratings of the speech variables clearly discriminated between patients with and without CP±L, except for the variables 'hyponasality', 'voice' and 'grimace' (table 5). Comparison of the perceptual evaluations with nasalance values and the NSI 2.0 is presented in table 6. Specificity was optimal, with the exception of hypernasality judgments in comparison with the oronasal text (0.56). Varying results were found for sensitivity, ranging from 0.67 for the oral and nasal text, to 0.83 for the oronasal text and 0.90 for the NSI 2.0.

Table 5. Results of the Fisher Exact Test for the Evaluation of the Discriminant Validity of the Speech Outcome Tool (Phase 1).

	P
Speech understandability	.016[a]
Anterior oral CSCs	.009[a]
Posterior oral CSCs	.033[a]
Non-oral CSCs	.033[a]
Passive CSCs	<.001[a]
Hypernasality	.001[a]
Hyponasality	.582
Nasal emission	.011[a]
Nasal turbulence	.011[a]
Voice	.087
Grimace	_[b]
Acceptability	<.001[a]
Need for SLT intervention	<.001[a]

Abbreviations: CSCs, cleft-related speech characteristics.
[a]Statistically significant, $P < .05$.
[b]Unable to compute as all ratings in the 2 groups were the same.

Table 6. Preliminary Evaluation of the Construct Validity of the Speech Outcome Tool (Phase 1).

Parameters	Sensitivity	Specificity	% Absolute Agreement
Hypernasality and nasometry oral text	0.67	1.00	75.00
Hypernasality and nasometry oronasal text	0.83	0.56	66.67
Hypernasality and NSI 2.0	0.90	1.00	93.33
Hyponasality and nasometry nasal text	0.67	1.00	93.33

Abbreviation: NSI, Nasality Severity Index.

Following phase 1, we considered that variables influencing the evaluation of speech understandability should be controlled for, such as the length of the speech sample and context (Yorkston and Beukelman, 1980). As a result, the structured listening protocol evaluated in phase 2 consisted of two samples of spontaneous speech: one sample without utterances of the conversation partner for the evaluation of speech understandability, and one sample of spontaneous speech with utterances of the conversation partner for the evaluation of consonant production. Furthermore, all samples could be replayed unlimitedly with the exception of the sample for the evaluation of understandability, which could only be played once to avoid the influence of familiarity on the ratings (Yorkston and Beukelman, 1980). Second, it was decided to report the age of each participant to the assessor. As such, when determining the need for SLT intervention, age appropriate utterances could be taken into account. Third, a distinction between the need for CP±L-related and non-CP±L-related SLT intervention was made. Additionally, the rating form was optimized by presenting the scales in the same order as the listening protocol and by adding a phonetic transcription of the target sentences.

Phase 2: evaluation of the speech outcome tool – listening experiment with experienced listeners

Face validity and feedback from the listeners

Generally, the four experienced listeners judged that the tool had face validity, meaning that the tool measured what was supposed to measure. The time spent on the presentation and the training was appreciated. However, two listeners felt the need for a longer training period with more time dedicated to consensus listening. Although the listeners argued the evaluation to be rather time consuming, the benefit of a narrow phonetic transcription was acknowledged. One listener proposed to eliminate the evaluation of consonant production based on spontaneous speech and automatic rote speech from the structured listening protocol, as these phonetic transcriptions were not taken into account when determining the presence of CSCs.

Inter- and intra-rater reliability

Only good to very good inter-rater reliability based on ICC values could by achieved for five variables, namely ‘speech understandability’, ‘non-oral CSCs’, ‘hypernasality’, ‘hyponasality’ and ‘speech acceptability’ (table 7). Inter-rater reliability of the variables ‘anterior oral CSCs’, ‘passive CSCs’, ‘non-CP±L-related SLT intervention’, ‘nasal emission’, ‘voice’ and ‘grimace’ was poor to fair. Twenty-four of the 56 (4 listeners – 14 variables) ICCs for intra-rater reliability ranged from good to very good (table 8). In general, good intra-rater reliability was found in three of the four experienced listeners.

Criterion validity

Overall, sufficient sensitivity and specificity of the tool when comparing with the nasalance values was found (table 9). This was not the case for the oral text, with a specificity of 0.64 and the poorest percentage of agreement (62.50%). Additionally, very good results were found when comparing the perceptual evaluation with the NSI 2.0.

Table 7. Interrater Reliability of the 4 Experienced Listeners in Phase 2.

	Single Measures ICC Type Consistency	95% CI ICC	Interpretation ICC[a]	Mean % Absolute Agreement
Speech understandability	0.77	0.52 to 0.93	Good	51.67
Anterior oral CSCs	0.35	0.05 to 0.72	Fair	45.00
Posterior oral CSCs	0.56	0.25 to 0.84	Moderate	55.00
Nonoral CSCs	0.73	0.46 to 0.91	Good	63.33
Passive CSCs	0.45	0.13 to 0.78	Fair	43.33
Non-CP ± L-related SLT intervention	0.17	−0.09 to 0.58	Poor	60.00
Hypernasality	0.69	0.40 to 0.89	Good	56.67
Hyponasality	0.85	0.66 to 0.95	Very good	88.33
Nasal emission	0.36	0.06 to 0.73	Poor	63.33
Nasal turbulence	0.58	0.27 to 0.85	Moderate	60.00
Voice	0.21	0.06 to 0.62	Fair	88.33
Grimace	0.33	0.03 to 0.71	Fair	93.33
Speech acceptability	0.76	0.52 to 0.93	Good	50.00
CP ± L-related SLT intervention	0.50	0.18 to 0.81	Moderate	80.00

Abbreviations: CI, confidence interval; CSCs, cleft-related speech characteristics; ICC, intraclass single measures correlation coefficients.
[a]Based on Altman (1990): ICC < 0.20: poor, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80: good, 0.81-1.00: very good.

Table 8. Intrarater Reliability of the 4 Experienced Listeners in Phase 2

	Listener 1				Listener 2			
	ICC	95% CI	Interpretation ICC[a]	% Agreement	ICC	95% CI	Interpretation ICC[a]	% Absolute Agreement
Speech understandability	0.82	0.42 to 0.95	Very good	70	0.86	0.52 to 0.96	Very good	70
Anterior oral CSCs	0.67	0.12 to 0.91	Good	70	0.86	0.54 to 0.96	Very good	80
Posterior oral CSCs	0.73	0.24 to 0.93	Good	70	0.75	0.27 to 0.93	Good	70
Nonoral CSCs	0.93	0.74 to 0.98	Very good	90	0.94	0.78 to 0.99	Very good	90
Passive CSCs	0.82	0.42 to 0.95	Very good	60	0.83	0.46 to 0.96	Very good	60
Non-CP ± L-related SLT intervention	1.00	1.00 to 1.00	Very good	100	0.00	–0.60 to 0.60	Poor	90
Hypernasality	0.59	–0.03 to 0.88	Moderate	70	0.65	0.07 to 0.90	Good	60
Hyponasality	0.84	0.49 to 0.96	Very good	90	1.00	1.00 to 1.00	Very good	100
Nasal emission	0.65	0.08 to 0.90	Good	80	0.71	0.20 to 0.92	Good	60
Nasal turbulence	1.00	1.00 to 1.00	Very good	100	0.72	0.20 to 0.92	Good	80
Voice	–	–	–[b]	80	0.64	0.06 to 0.90	Good	90
Grimace	1.00	1.00 to 1.00	Very good	100	1.00	1.00 to 1.00	Very good	100
Speech acceptability	0.91	0.67 to 0.98	Very good	90	0.86	0.54 to 0.96	Very good	70
CP ± L-related SLT intervention	1.00	1.00 to 1.00	Very good	100	1.00	1.00 to 1.00	Very good	100
	Listener 3				Listener 4			
Speech understandability	0.89	0.62 to 0.97	Very good	60	0.95	0.81 to 0.99	Very good	80
Anterior oral CSCs	0.53	–0.10 to 0.86	Moderate	70	0.93	0.76 to 0.98	Very good	90
Posterior oral CSCs	0.45	–0.21 to 0.83	Moderate	70	0.64	0.06 to 0.90	Good	70
Nonoral CSCs	0.53	–0.10 to 0.86	Moderate	70	0.60	0.00 to 0.88	Moderate	60
Passive CSCs	0.70	0.17 to 0.92	Good	60	0.88	0.60 to 0.97	Very good	90
Non-CP ± L-related SLT intervention	0.60	–0.00 to 0.88	Moderate	80	0.67	0.11 to 0.91	Good	80
Hypernasality	0.83	0.46 to 0.96	Very good	70	0.52	–0.13 to 0.85	Moderate	50
Hyponasality	0.80	0.83 to 0.95	Good	90	0.88	0.60 to 0.97	Very good	90
Nasal emission	–	–	–[b]	50	0.82	0.42 to 0.95	Very good	80
Nasal turbulence	0.66	0.10 to 0.90	Good	50	0.87	0.55 to 0.97	Very good	80
Voice	0.00	–0.60 to 0.60	Poor	90	1.00	1.00 to 1.00	Very good	100
Grimace	0.00	–0.60 to 0.60	Poor	70	–	–	–[b]	60
Speech acceptability	0.73	0.23 to 0.93	Good	70	0.89	0.62 to 0.97	Very good	70
CP ± L-related SLT intervention	0.52	–0.11 to 0.86	Moderate	80	1.00	1.00 to 1.00	Very good	100

Abbreviation: ICC, intraclass single measures correlation coefficients.
[a]Based on Altman (1990): ICC < 0.20: poor, 0.21-0.40: fair, 0.41-0.60: moderate, 0.61-0.80: good, 0.81-1.00: very good.
[b]ICC impossible to calculate due to zero variance or negative covariance.

Table 9. Evaluation of the Construct Validity of the Speech Outcome Tool (Phase 2).

Parameters	Sensitivity	Specificity	% Absolute Agreement
Hypernasality and nasometry oral text	1.00	0.64	62.50
Hypernasality and nasometry oronasal text	0.83	1.00	92.86
Hypernasality and NSI 2.0	0.88	1.00	91.67
Hyponasality and nasometry nasal text	1.00	1.00	92.86

Abbreviation: NSI, Nasality Severity Index.

Discussion

This paper presents the development and two-phase validation of an outcome tool for perceptual speech assessment in Belgian Dutch-speaking patients with CP±L. Evaluation of the tool in phase 1 showed good to very good inter- and intra-rater reliability for the majority of the variables, good discriminant validity, and varying results regarding the criterion validity. Following phase 1, some adaptations to the tool were made. In a second phase, less straightforward results regarding reliability were obtained with good to very good inter-reliability for five of the 14 variables, and good intra-rater reliability in three of the four experienced listeners. Sensitivity and specificity were sufficient, with the exception of the specificity of the hypernasality rating in comparison with the nasalance values of the oral text. Overall, listeners positively judged the face validity of the tool.

In both phases, inter-rater reliability was poorer than intra-rater reliability, similar to findings by Chapman et al. (2016). More specifically, inter-rater reliability results were better in phase 1 than phase 2. This can probably be explained by the fact that the listeners of phase 1 also collected the evaluated speech samples and thus were familiar with the patients (Sell, 2005). Similar to CAPS-A and CAPS-A-AM studies, the variables ‘anterior oral CSCs’, ‘non-

CP±L-related SLT intervention' and 'voice' were susceptible for poorer reliability (John et al., 2006; Sell et al., 2009; Chapman et al., 2016). Sell et al. (2009) listed explanations, including the limited focus spent on these variables in training, the possible overlap between developmental and cleft-related speech errors, and the subjective nature of the ratings. The third listener of phase 2 showed poor intra-rater reliability. Additionally, inter-rater reliability for passive CSCs was only fair, whereas good to very good intra-rater reliability for this variable was achieved for all listeners. To optimize reliability, a more elaborate training to delineate and maintain the listeners' internal standard over time should be considered in future studies (Gooch et al., 2001; Brunnegard et al., 2009; Sell et al., 2009). Intra-rater reliability of the CSCs showed the greatest variability among the listeners. This can be the result of the narrow phonetic transcription, which might be beyond our perceptual capabilities (Shriberg and Lof, 1991).

In the future, the inclusion of a greater number of listeners and speech samples is needed to provide a clearer insight in variables that are susceptible for poorer reliability. Moreover, inclusion of more speech samples might increase the variability of the speech variables, and hence improve ICCs (Hallgren, 2012). The influence of the limited variability of some variables in this study, such as voice and grimace, is illustrated by the discrepancy between ICC measures and the percentage of absolute agreement. Specifically single measures ICCs were calculated, as these are interchangeable with quadratic weighted kappas for ordinal scales and can be used in case of more than two listeners (Norman and Streiner, 2008). However, interpretation of these measures is generally poorer in comparison to average measures ICCs, used for example by Sell et al. (2009), because of the different nature of these measurements (Hallgren, 2012). The number of speech samples was limited due to time constraints and practical considerations, especially in phase 2, which also prevented a re-evaluation of the discriminant validity. In phase 1, the poorer discriminant validity results variables 'hyponasality', 'voice' and 'grimace' can most likely be explained by the limited variability as well.

Regarding the listening protocol, some opportunities to optimize the tool's validity (and reliability) were identified. The majority of the parameters was evaluated at sentence level, which provides a general understanding of the child's speech performance (Van Demark, 1964; Klintö et al., 2011). The speech sample at sentence level mainly included sentences devoid of nasal consonants (Henningsson et al., 2008). Such phonetic content is not representative of conversational speech, possibly hampering the speech sample's face validity (Sweeney and Sell, 2008). Additionally, evaluation of consonant production based on the samples of spontaneous speech and automatic sequences might be too time-consuming. Hence, evaluation of the reliability and validity of the tool when eliminating these steps from the structured listening protocol, and when adding the level of automatic sequences to evaluate resonance and nasal airflow is subject for further research. Furthermore, the validity of the composed speech samples for the different age groups of participants should be evaluated. The criterion validity clearly improved in phase 2. However, the specificity of the hypernasality rating in comparison with the nasalance values of the oral text remained poor. Varying correlations between nasality judgements and nasalance values have been reported (Watterson et al., 1993; Keuning et al., 2004; Sweeney and Sell, 2008; Brancamp et al., 2010; Brunnegard et al., 2012). Several aspects, including the presence of nasal airflow, the speech sample and the acoustic restrictions of the Nasometer, may influence the relationship between perceptual ratings and nasalance values (Sweeney and Sell, 2008). Specifically for this study, both speech samples were devoid of nasal consonants. However, the difference in the number of high vowels between both samples might be an explanation for the poorer specificity in comparison with the nasalance values of the oral text (Hutters and Henningsson, 2004). In addition to these analyses, nasendoscopy and/or videofluoroscopy results could have substantiated the evaluation of criterion validity.

To date, discussion is ongoing about the most appropriate type of rating scale when evaluating resonance, nasal airflow and by extent understandability and acceptability (Castick

et al., 2017). The validity of partition measures, such as ordinal and equal appearing interval (EAI) scales, and magnitude measures, such as visual-analogue scaling (VAS) and direct magnitude estimation (DME), is based on the observation whether a perceptual phenomenon is either metathetic or prothetic in nature (Stevens, 1975). Growing evidence in the literature indicates the prothetic nature of perceptual phenomena such as hypernasality, hyponasality, nasal emission and nasal turbulence, understandability and acceptability (Baylis et al., 2015; Castick et al., 2017; Bettens et al., 2018), and thus the use of VAS (Stevens, 1975). Nevertheless, ordinal scaling was used for this outcome tool in accordance to internationally accepted protocols (John et al., 2006; Henningsson et al., 2008; Chapman et al., 2016) and because ordinal scaling allows for clear communication and straightforward comparisons between studies (Brancamp et al., 2010). A solution to compromise for the prothetic nature of the ratings on one hand, and the need for straightforward communication on the other hand might be the use of a graphic rating scale or color-coding applied to a VAS scale (Castick et al., 2017). Similarly, Yamashita et al. (2018) described excellent reliability when using the Borg centiMax (CM) scale (Borg and Borg, 2001), which combines partition and magnitude measures by placing verbal anchors on a ratio scale.

Another aspect to consider is the inclusion of additional variables, such as a summary score to indicate velopharyngeal function. Examples of such measures are the VPC-SUM (Lohmander et al., 2009; Lohmander et al., 2017b), the adaptation of the VPC-SUM to CAPS-A ratings by Pereira et al. (2013), and the structural outcome score described by Sell et al. (2017). As a major part of the definitions and parameters of the current outcome tool were derived from the CAPS-A, the scores proposed by Pereira et al. (2013) or Sell et al. (2017) might be favorable for the current tool. Subsequent to the current study, application of the former measure to the Belgian Dutch outcome tool showed very good inter- and intra-rater reliability results (Bruneel et al., 2019b). Regarding consonant production, additional

measurements should be considered as well. The current tool focusses on error analysis, with a categorization of anterior oral, posterior oral, non-oral and passive CSCs (Sell et al., 2009). An additional measure of consonant proficiency should be considered. A measure often used in studies assessing speech in patients with CP±L is 'percent consonants correct' (PCC) (Shriberg and Kwiatkowski, 1982; Shriberg et al., 1997). Several adaptations of the PCC measure have been described, such as percentage correct manner (Morris and Ozanne, 2003; Chapman et al., 2008; Lohmander and Persson, 2008; Klintö et al., 2011), percentage correct place (Lohmander and Persson, 2008; Klintö et al., 2011), PCC corrected for age (Klintö et al., 2014), and percentage correct oral consonants (Malmborn et al., 2018). Another possibility is to measure the size of the consonant inventory as was done by Chapman et al. (2008). Regardless of the specific choice, the validity and reliability of additional measures or adaptations to the tool should be evaluated. In addition to the perceptual evaluation, several authors advocated including the patient's (and the caregiver's perception) regarding speech and speech-related quality of life as part of speech outcome measures (Mossey et al., 2009; Eckstein et al., 2011; Klassen et al., 2012). A questionnaire comprising a parent report and a youth report has been validated in Belgian Dutch-speaking patients with cleft palate (Bruneel et al., 2017; Bruneel et al., 2019a), and several speech variables of the tool described in this manuscript showed a significant correlation with these health-related quality of life scores (Bruneel et al., 2019b).

Conclusion

This study showed varying results regarding inter- and intra-rater reliability and criterion validity of the developed Belgian Dutch outcome tool. The discriminant validity and face validity were considered sufficient. Adaptations to the current tool such as changes to the listening protocol, addition of speech variables and inclusion of a more elaborate training will be evaluated in future studies to optimize validity and reliability.

References

Allori, A. C., Kelley, T., Meara, J. G., Albert, A., Bonanthaya, K., Chapman, K., et al. (2017). A standard set of outcome measures for the comprehensive appraisal of cleft care. *The Cleft palate-craniofacial journal, 54*(5), 540-554.

Altman, D. G. (1990). *Practical statistics for medical research*: CRC press.

Baylis, A., Chapman, K., Whitehill, T. L., Group, T. A. (2015). Validity and Reliability of Visual Analog Scaling for Assessment of Hypernasality and Audible Nasal Emission in Children With Repaired Cleft Palate. *The Cleft palate-craniofacial journal, 52*(6), 660-670.

Bettens, K., Bruneel, L., Maryn, Y., De Bodt, M., Luyten, A., Van Lierde, K. M. (2018). Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores. *Journal of Communication Disorders, 76*, 11-20.

Bettens, K., Van Lierde, K. M., Corthals, P., Luyten, A., Wuyts, F. L. (2016). The Nasality Severity Index 2.0: revision of an objective multiparametric approach to hypernasality. *The Cleft palate-craniofacial journal, 53*(3), 60-70.

Bettens, K., Wuyts, F. L., Jonckheere, L., Platbrood, S., Van Lierde, K. (2017). Influence of gender and age on the Nasality Severity Index 2.0 in Dutch-speaking Flemish children and adults. *Logopedics Phoniatrics Vocology, 42*(3), 133-140.

Boersma, W., Weenink, D. PRAAT: doing phonetics by computer

Borg, G., Borg, E. (2001). A new generation of scaling methods: Level-anchored ratio scaling. *Psychologica, 28*(1), 15-45.

Brancamp, T. U., Lewis, K. E., Watterson, T. (2010). The relationship between nasalance scores and nasality ratings obtained with equal appearing interval and direct magnitude estimation scaling methods. *The Cleft palate-craniofacial journal, 47*(6), 631-637.

Breuls, M., Sell, D., Manders, E., Boulet, E., Poorten, V. (2006). SISL (ScreeningsInstrument Schisis Leuven): assessment of cleft palate speech, resonance and myofunction. *B ENT*, 71.

Britton, L., Albery, L., Bowden, M., Harding-Bell, A., Phippen, G., Sell, D. (2014). A Cross-Sectional Cohort Study of Speech in Five-Year-Olds With Cleft Palate±Lip to Support Development of National Audit Standards: Benchmarking Speech Standards in the United Kingdom. *The Cleft palate-craniofacial journal, 51*(4), 431-451.

Brondsted, K., Grunwell, P., Henningsson, G., Jansonius, K., Karling, J., Meijer, M., et al. (1994). A phonetic framework for the cross-linguistic analysis of cleft palate speech. *Clinical linguistics & phonetics, 8*(2), 109-125.

Bruneel, L., Alighieri, C., De Smet, S., Bettens, K., De Bodt, M., Van Lierde, K. (2019a). Health-related quality of life in patients with cleft palate: reproducibility, responsiveness and construct validity of the Dutch version of the VELO questionnaire. *Int. Journal of Pediatric Otorhinolaryngology, 119*, 141-146.

Bruneel, L., Bettens, K., Van Lierde, K. (2019b). The relationship between health-related quality of life and speech in patients with cleft palate. *International Journal of Pediatric Otorhinolaryngology, 120*, 112-117.

Bruneel, L., Van Lierde, K., Bettens, K., Corthals, P., Van Poel, E., De Groote, E., et al. (2017). Health-related quality of life in patients with cleft palate: Validity and reliability of the VPI Effects on Life Outcomes (VELO) questionnaire translated to Dutch. *International Journal of Pediatric Otorhinolaryngology, 98*, 91-96.

Brunnegard, K., Lohmander, A., van Doorn, J. (2009). Untrained listeners' ratings of speech disorders in a group with cleft palate: a comparison with speech and language pathologists' ratings. *Int J Lang Commun Disord, 44*(5), 656-674.

Brunnegard, K., Lohmander, A., van Doorn, J. (2012). Comparison between perceptual assessments of nasality and nasalance scores. *Int J Lang Commun Disord, 47*(5), 556-566.

Castick, S., Knight, R.-A., Sell, D. (2017). Perceptual judgments of resonance, nasal airflow, understandability, and acceptability in speakers with cleft palate: ordinal versus visual analogue scaling. *The Cleft palate-craniofacial journal, 54*(1), 19-31.

Chapman, K. L., Baylis, A., Trost-Cardamone, J., Cordero, K. N., Dixon, A., Dobbelsteyn, C., et al. (2016). The Americleft Speech Project: A Training and Reliability Study. *The Cleft palate-craniofacial journal, 53*(1), 93-108.

Chapman, K. L., Hardin-Jones, M. A., Goldstein, J. A., Halter, K. A., Havlik, R. J., Schulte, J. (2008). Timing of palatal surgery and speech outcome. *The Cleft palate-craniofacial journal, 45*(3), 297-308.

Eckstein, D. A., Wu, R. L., Akinbiyi, T., Silver, L., Taub, P. J. (2011). Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. *Plastic and Reconstructive Surgery, 128*(5), 518e-526e.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin, 76*(5), 378.

Gooch, J. L., Hardin-Jones, M., Chapman, K. L., Trost-Cardamone, J. E., Sussman, J. (2001). Reliability of listener transcriptions of compensatory articulations. *The Cleft palate-craniofacial journal, 38*(1), 59-67.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23.

Harding, A., Grunwell, P. (1998). Active versus passive cleft-type speech characteristics. *International Journal of Language & Communication Disorders, 33*(3), 329-352.

Harding, A., Harland, K., Razzell, R. (1997). Cleft audit protocol for speech (CAPS). *Broomfield, Chelmsford, Essex: St. Andrew's Plastic Surgery Centre*.

Henningsson, G., Hutters, B. (1997). *Perceptual assessment of cleft palate speech, with special reference to minimum standards for inter-centre comparisons of speech outcome.* Paper presented at the Transactions 8th International Congress on Cleft Palate and Related Anomalies.

Henningsson, G., Kuehn, D. P., Sell, D., Sweeney, T., Trost-Cardamone, J. E., Whitehill, T. L. (2008). Universal parameters for reporting speech outcomes in individuals with cleft palate. *The Cleft palate-craniofacial journal, 45*(1), 1-17.

Hutters, B., Henningsson, G. (2004). Speech outcome following treatment in cross-linguistic cleft palate studies: methodological implications. *The Cleft palate-craniofacial journal, 41*(5), 544-549.

John, A., Sell, D., Sweeney, T., Harding-Bell, A., Williams, A. (2006). The cleft audit protocol for speech-augmented: a validated and reliable measure for auditing cleft speech. *The Cleft palate-craniofacial journal, 43*(3), 272-288.

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.

Keuning, K. H., Wieneke, G. H., Dejonckere, P. H. (2004). Correlation between the perceptual rating of speech in Dutch patients with velopharyngeal insufficiency and composite measures derived from mean nasalance scores. *Folia Phoniatrica et Logopaedica, 56*(3), 157-164.

Klassen, A. F., Tsangaris, E., Forrest, C. R., Wong, K. W., Pusic, A. L., Cano, S. J., et al. (2012). Quality of life of children treated for cleft lip and/or palate: a systematic review. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 65*(5), 547-557.

Klintö, K., Salameh, E.-K., Svensson, H., Lohmander, A. (2011). The impact of speech material on speech judgement in children with and without cleft palate. *International Journal of Language & Communication Disorders, 46*, 348–360.

Klintö, K., Svensson, H., Elander, A., Lohmander, A. (2014). Speech and phonology in Swedish-speaking 3-year-olds with unilateral complete cleft lip and palate following different methods for primary palatal surgery. *The Cleft palate-craniofacial journal, 51*(3), 274-282.

Kuehn, D. P., Moller, K. T. (2000). Speech and language issues in the cleft palate population: the state of the art. *The Cleft palate-craniofacial journal, 37*(4), 348-348.

Kuehn, D. P., Moon, J. B. (1998). Velopharyngeal closure force and levator veli palatini activation levels in varying phonetic contexts. *Journal of Speech, Language, and Hearing Research, 41*(1), 51-62.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lohmander. (2008). Universal parameters for reporting speech outcomes in individuals with cleft palate. [Letter to the editor]. *The Cleft palate-craniofacial journal, 45*(4), 452-453.

Lohmander, Persson, C., Willadsen, E., Lundeborg, I., Alaluusua, S., Aukner, R., et al. (2017a). Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 4. Speech outcomes in 5-year-olds-velopharyngeal competency and hypernasality. *Journal of plastic surgery and hand surgery, 51*(1), 27-37.

Lohmander, A., Hagberg, E., Persson, C., Willadsen, E., Lundeborg, I., Davies, J., et al. (2017b). Validity of auditory perceptual assessment of velopharyngeal function and dysfunction–the VPC-Sum and the VPC-Rate. *Clinical linguistics & phonetics, 31*(7-9), 589-597.

Lohmander, A., Olsson, M. (2004). Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature. *The Cleft palate-craniofacial journal, 41*(1), 64-70.

Lohmander, A., Persson, C. (2008). A longitudinal study of speech production in Swedish children with unilateral cleft lip and palate and two-stage palatal repair. *The Cleft palate-craniofacial journal, 45*(1), 32-41.

Lohmander, A., Willadsen, E., Persson, C., Henningsson, G., Bowden, M., Hutters, B. (2009). Methodology for speech assessment in the Scandcleft project--an international randomized clinical trial on palatal surgery: experiences from a pilot study. *The Cleft palate-craniofacial journal, 46*(4), 347-362.

Malmborn, J.-O., Becker, M., Klintö, K. (2018). Problems With Reliability of Speech Variables for Use in Quality Registries for Cleft Lip and Palate—Experiences From the Swedish Cleft Lip and Palate Registry. *The Cleft palate-craniofacial journal*, 1051 - 1059.

McLeod, S., Harrison, L. J., McCormack, J. (2012). The intelligibility in context scale: Validity and reliability of a subjective rating measure. *Journal of Speech, Language, and Hearing Research, 55*(2), 648-656.

Moon, J. B., Kuehn, D. P., Huisman, J. J. (1994). Measurement of velopharyngeal closure force during vowel production. *The Cleft palate-craniofacial journal, 31*(5), 356-363.

Morris, H., Ozanne, A. (2003). Phonetic, phonological, and language skills of children with a cleft palate. *The Cleft palate-craniofacial journal, 40*(5), 460-470.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., Shaw, W. C. (2009). Cleft lip and palate. *The Lancet, 374*(9703), 1773-1785.

Norman, G., Streiner, D. (2008). Elements of statistical inference. *Biostatistics: The Bare Essentials*, 46-62.

Pereira, V. J., Sell, D., Tuomainen, J. (2013). Effect of maxillary osteotomy on speech in cleft lip and palate: perceptual outcomes of velopharyngeal function. *International Journal of Language and Communication Disorders, 48*(6), 640-650.

Peterson-Falzone. (2006). *The clinician's guide to treating cleft palate speech* (Vol. 1): Mosby.

Peterson-Falzone, S. J., Hardin-Jones, M. A., Karnell, M. P., McWilliams, B. J. (2001). *Cleft palate speech*. St. Louis: Mosby.

Samoy, K., Hens, G., Verdonck, A., Schoenaers, J., Dormaar, T., Breuls, M., et al. (2015). Surgery for velopharyngeal insufficiency: The outcomes of the University Hospitals Leuven. *International Journal of Pediatric Otorhinolaryngology, 79*(12), 2213-2220.

Sell. (2005). Issues in perceptual speech analysis in cleft palate and related disorders: a review. *International Journal of Language & Communication Disorders, 40*(2), 103-121.

Sell, Harding, A., Grunwell, P. (1999). GOS. SP. ASS.'98: an assessment for speech disorders associated with cleft palate and/or velopharyngeal dysfunction (revised). *International Journal of Language & Communication Disorders, 34*(1), 17-33.

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

Sell, Mildinhall, S., Albery, L., Wills, A., Sandy, J., Ness, A. (2015). The Cleft Care UK study. Part 4: perceptual speech outcomes. *Orthodontics & craniofacial research, 18*, 36-46.

Sell, Southby, L., Wren, Y., Wills, A., Hall, A., Mahmoud, O., et al. (2017). Centre-level variation in speech outcome and interventions, and factors associated with poor speech outcomes in 5-year-old children with non-syndromic unilateral cleft lip and palate: The Cleft Care UK study. Part 4. *Orthodontics & craniofacial research, 20*, 27-39.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research, 40*(4), 708-722.

Shriberg, L. D., Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47*(3), 256-270.

Shriberg, L. D., Kwiatkowski, J., Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech, Language, and Hearing Research, 27*(3), 456-465.

Shriberg, L. D., Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical linguistics & phonetics, 5*(3), 225-279.

Sitzman, T. J., Allori, A. C., Thorburn, G. (2014). Measuring outcomes in cleft lip and palate treatment. *Clin Plast Surg, 41*(2), 311-319.

Spruijt, N. E., Beenakker, M., Verbeek, M., Heinze, Z. C., Breugem, C. C., van der Molen, A. B. M. (2018). Reliability of the Dutch Cleft Speech Evaluation Test and Conversion to the Proposed Universal Scale. *Journal of Craniofacial Surgery, 29*(2), 390-395.

Stevens, S. (1975). Laws That Govern Behavior.(Book Reviews: Psychophysics. Introduction to Its Perceptual, Neural, and Social Prospects). *Science, 188*, 827-829.

Sweeney, T., Sell, D. (2008). Relationship between perceptual ratings of nasality and nasometry in children/adolescents with cleft palate and/or velopharyngeal dysfunction. *International Journal of Language & Communication Disorders, 43*(3), 265-282.

Van de Weijer, J., Slis, I. (1991). Nasaliteitsmeting met de nasometer. *Logopedie en Foniatrie, 63*, 97-101.

Van Demark, D. R. (1964). Misarticulations and listener judgements of the speech of individuals with cleft palates. *The Cleft palate journal, 1*(2), 232-245.

Vander Poorten, V., Ostyn, F., Van Kerckhoven, W., Wellens, W., Breuls, M., Verdonck, A., et al. (2006). The Leuven staged supraperiosteal retropositioning repair: long-term velopharyngeal function in non-syndromic cleft palate. *B-ENT, 2*(Suppl 4), 35-43.

Verhoeven, J. (2005). Belgian standard dutch. *Journal of the International Phonetic Association, 35*(2), 243-247.

Watson, A., Sell, D., Grunwell, P. (2001). *Management of cleft lip and palate*: John Wiley & Sons Incorporated.

Watterson, T., McFarlane, S. C., Wright, D. S. (1993). The relationship between nasalance and nasality in children with cleft palate. *Journal of Communication Disorders, 26*(1), 13-28.

Whitehill, T. L. (2002). Assessing intelligibility in speakers with cleft palate: a critical review of the literature. *The Cleft palate-craniofacial journal, 39*(1), 50-58.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in health, 8*(2), 94-104.

Willadsen, E., Lohmander, A., Persson, C., Lundeborg, I., Alaluusua, S., Aukner, R., et al. (2017). Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 5. Speech outcomes in 5-year-olds-consonant proficiency and errors. *Journal of plastic surgery and hand surgery, 51*(1), 38-51.

Wong Riff, K. W., Tsangaris, E., Goodacre, T., Forrest, C. R., Pusic, A. L., Cano, S. J., et al. (2017). International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). *BMJ open, 7*(1), e015467.

Yamashita, R. P., Borg, E., Granqvist, S., Lohmander, A. (2018). Reliability of Hypernasality Rating: Comparison of 3 Different Methods for Perceptual Assessment. *The Cleft palate-craniofacial journal*.

Yorkston, K. M., Beukelman, D. R. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders, 13*(1), 33-41.