# Reliability of Outcome Measures to Assess Consonant Proficiency Following Cleft Palate Speech Intervention: The Percentage of Consonants Correct Metric and the Probe Scoring System

Cassandra Alighieri,[a] Kim Bettens,[a] Laura Bruneel,[a] Evelien D'haeseleer,[a]
Ellen Van Gaever,[a] and Kristiane Van Lierde[a,b]

[a]Department of Rehabilitation Sciences, Ghent University, Belgium
[b]Department of Speech-Language Pathology and Audiology, University of Pretoria, South Africa

* Correspondence to Cassandra Alighieri: Cassandra.Alighieri@UGent.be

## Abstract

**Purpose:** This study compared the inter- and intrarater reliability of the percentage of consonants correct (PCC) metrics and the probe scoring system between an experienced and a less experienced rater and between two experienced raters. In addition, these outcome measures' ability to reflect changes following speech intervention was measured.

**Method:** During Phase 1, two raters (Rater 1 with 5 years of experience in cleft-related speech disorders and Rater 2 with limited experience in cleft-related speech disorders) independently assessed 134 speech samples at the word and sentence levels, which were collected on different data points before, during, and following a cleft palate speech intervention. During Phase 2, a third rater (with 8 years of experience) analyzed 34 speech samples. The percentage of consonants correct–revised, the percentage of correct places and manners, and probe scores at the word and sentence levels were measured.

**Results:** Poor-to-moderate interreliability between Raters 1 and 2 was found due to differences in error classification. Interrater reliability between Raters 1 and 3 was very good for both the PCC metrics and the probe scores. The interrater reliability for the amount of targets elicited was lower compared to the interrater reliability for the amount of targets correct. The probe scoring system demonstrated a greater ability to detect changes toward the correct production of the target consonant compared to the PCC metrics.

**Conclusions:** Having an experience with the assessment of cleft-related speech disorders is a crucial factor to gain reliable results. The interrater reliability for the PCC metrics and the probe scoring system between two experienced raters did not differ, suggesting that both outcome measures can be used in cleft palate speech intervention studies. Despite the ability of the probe scoring system to detect changes, further research should provide insight in the benefits of this system both for research and clinical purposes.

Speech assessment in children with a cleft of the palate with or without a cleft of the lip (CP ± L) is challenging given the complex character of speech disorders in this population (Sell, 2005). The analysis and transcription of such complex speech disorders are often associated with poor inter- and intrarater reliability (Sell, 2005; Sitzman et al., 2014; Vallino et al., 2008). There are several factors known to contribute to the reliability of transcriptions, such as the severity of the speech disorder, the phonetic and phonological background of the transcribers, the used speech samples, the methods of calculation, the criteria of agreement, and the time between the original transcription and the retranscription (Chapman et al., 2016; Henningsson et al., 2008; Johnson et al., 2004; Klintö et al., 2011). A crucial element that influences the reliability of speech assessments is the characteristics of the rater (Chapman et al., 2016; Cordes, 1994). Raters may have different internal standards affecting their perceptual ratings (Kreiman et al., 1993). This internal standard may change across time, for example, when the rater's experience increases (Kreiman et al., 1993). Both inter- and intrarater reliability might thus be affected by the level of experience (Kreiman et al., 1993). Chapman et al. (2016) reported that there are variable findings across studies that examined the impact of listener experience when assessing the speech of individuals with a CP ± L. While in some cleft palate speech studies better reliability was found for experienced listeners compared to inexperienced listeners (Gooch et al., 2001; Keuning et al., 1999; Lewis et al., 2003), other studies did not (Brunnegård et al., 2009; Tönz et al., 2002). Based on these divergent findings, Chapman et al. (2016) argued that the effect of listener experience required further research.

To date, the primary focus of the cleft literature has been on the assessment of speech outcomes following different surgical procedures (Sell & Sweeney, 2020). With regard to cleft palate speech, the International Consortium for Health Outcomes Measurement (ICHOM) identified three different outcome domains that should be considered when assessing speech in a child with a CP ± L: (a) speech intelligibility, (b) velopharyngeal competence, and (c) articulatory proficiency (Allori et al., 2017). Intelligibility is measured using the family-reported Intelligibility in Context Scale (McLeod et al., 2012). With regard to velopharyngeal competence, the group proposed the use of a clinician-reported three-tier rating scale developed by Lohmander et al. (2009). Concerning articulatory proficiency, the ICHOM group reported that this outcome should be assessed with a standardized speech sample controlling for high-pressure consonants and high vowels, which are vulnerable for cleft palate speech disorders (Kuehn & Moon, 1998; Peterson-Falzone et al., 2001). They suggested the use of the percentage of consonants correct (PCC) score to assess articulatory proficiency or consonant proficiency (Shriberg & Kwiatkowski, 1982). Originally, the PCC score was calculated by dividing the number of correctly produced consonants (numerator) by the total number of consonants in a conversational speech sample (denominator), multiplied by 100. In other words, “the PCC score reflects the total percentage of correct consonants, with each consonant's contribution to this total weighted by its frequency of intended occurrence in conversational speech” (Shriberg, 1993).

Over the years, these PCC metrics have been criticized for several reasons (Kent et al., 1994; Shriberg et al., 1997). Originally, the PCC score aimed to express the percentage of correctly produced consonants in a conversational speech sample. Conversational speech, however, lacks standardization because the phonetic content cannot be controlled. This hampers a comparison between and within individuals and between and within centers (Shriberg et al., 1997). Consequently, numerous variants of the original PCC metric were reported using various nonconversational speech tasks (e.g., calculating the percentage of correct consonants on a picture-naming task). Nonconversational speech tasks, in turn, may be less representative of the child's spontaneous speech (James et al., 2016). Beside these issues regarding the speech sample, the PCC score may obscure important differences associated with only certain sounds or only certain subgroups of sounds because the 24 English consonants are treated as one response class (Shriberg et al., 1997). Additionally, the concern was raised that the PCC score made no distinction between different types of speech-sound errors (i.e., distortions, omissions, and substitutions; Shriberg et al., 1997). To respond to these concerns, several PCC extensions were developed, for example, the percentage of consonants correct–adjusted (PCC-A) and the percentage of consonants correct–revised (PCC-R; Shriberg et al., 1997). The PCC-A score allows for common clinical distortions (i.e., labialized or velarized /l/ or /r/, lateralized or dentalized voiced or voiceless fricatives or affricates, and derhotacized /r/), whereas the PCC-R score allows for both common clinical distortions and uncommon clinical distortions (i.e., weak or imprecise consonants, nasal emissions, [de]nasalized consonants, and failure to maintain appropriate voicing; Shriberg, 1993; Shriberg et al., 1997). Despite their efforts to address these concerns, it seemed that the different PCC scores were inappropriate to detect changes in a child's speech following speech intervention (Hall et al., 1998). This might not seem very surprising considering that the PCC procedure was originally developed as a measure for severity rather than a measure for change (Shriberg & Kwiatkowski, 1982).

Despite these concerns, the PCC scores are frequently used to measure changes in a child's consonant proficiency in cleft palate speech studies. Scherer et al. (2008) were among the first to use the PCC-R metric to assess the effect of early intervention in 10 children with a cleft of the palate. The interrater reliability between two transcribers yielded a weighted kappa of .68 (i.e., good agreement). Unfortunately, the study provided no information about the background and experience of the transcribers. Dobbelsteyn et al. (2014) assessed the effectiveness of a corrective babbling speech program in 12 children with a cleft palate or velopharyngeal dysfunction. The authors used the PCC metric and included all consonants from all targets words, with the exception of a few consonants that could vary with the dialects used by the included children. Unfortunately, the study did not report any results with regard to the reliability of the used PCC metric. Sell and Sweeney (2020) responded to this need by describing the challenges they experienced in terms of achieving good inter- and intrarater reliability for the PCC measure in a speech intervention study in children with a CP ± L (Sell & Sweeney, 2020; Sweeney et al., 2020). Two trained and experienced raters each analyzed 119 recordings, which were randomly selected from five data points before, during, and following speech intervention. The analysis included the calculation of a modified PCC score, which meant that consonants produced with a correct place, manner, and voice but with accompanying nasal emission/turbulence, weak/nasalized realizations, and dental/interdental quality were categorized as correct. The modified PCC scores at the word and sentence levels were calculated by dividing the number of correctly produced consonants (numerator) by the total number of consonants elicited (denominator), multiplied by 100.

Initially, the results revealed poor reliability for the amount of targets elicited for the words (intraclass correlation coefficient [ICC] = .07) and sentences (ICC = .42). Differences between the two raters in classification of errors as glottal stops and consonant deletions were mentioned as factors accounting for this poor reliability. The two raters followed additional training after which a second reliability study was performed. This second study demonstrated improvements in the reliability of the number of targets elicited. Following the training, the study results revealed an improvement in the reliability of the number of targets elicited in words (ICC = .85) and sentences (ICC = .94). There was very good interrater reliability for the PCC score for the word data set (ICC = .90) and the sentence data set (ICC = .88). The authors advocated that, although the modified PCC was a reliable outcome measure, this metric demonstrated significant limitations when used in a speech intervention study. Specifically in cleft palate speech interventions, a child often struggles with the correct production of a specific target consonant but has made some progress during therapy. For example, a child who produces an active nasal fricative for /ʃ, tʃ, ʤ/ preintervention and uses an orally produced /s, tj, dj/ postintervention has made some progress (Sell & Sweeney, 2020). This change is not detected using the modified PCC score because the target consonants are not yet produced correctly (Sell & Sweeney, 2020). Lohmander and Persson (2008) attempted to detect these positive changes toward the correct production of a target by calculating two PCC modifications, namely, the percentage of correct places (PCP) and the percentage of correct manners (PCM). The PCP score determines the number of correct articulatory places, whereas the PCM score assesses the total number of correct articulatory manners. Unfortunately, the authors did not report any reliability measures on these PCP and PCM scores. Beside these PCP and PCM scores, percentage of nonoral errors, percentage of oral consonants corrects (POCC), and percentage of oral errors have also been used in studies investigating speech outcomes in individuals with a CP ± L (Klintö et al., 2018; Lohmander et al., 2009; Malmborn et al., 2018; Willadsen et al., 2017). Although previous Swedish studies have found good reliability for the percentage of nonoral errors (Brunnegård et al., 2020; Malmborn et al., 2018), it should be noted that Swedish is a language without nonoral consonants (Brunnegård et al., 2020). Thereby, few children in the Swedish outcome studies displayed nonoral errors (Malmborn et al., 2018). If more children with nonoral errors had been included, the reliability results might have been lower (Malmborn et al., 2018). POCC is defined as a scoring system in which phonological and articulatory errors and errors related to velopharyngeal competence (e.g., audible nasal air leakage or weak articulation) are scored as incorrect (Klintö et al., 2018). As suggested by Sell and Sweeney (2020), it appears that the definition of the POCC score is similar to the original description of the PCC score. Malmborn et al. (2018) raised a concern about the use of the POCC score. The authors suggested that an overall measure of articulatory and phonological ability, in which passive cleft speech characteristics are not scored as incorrect, should be used because speech errors related to the place and manner of articulation may be more stigmatizing than soft signs of velopharyngeal dysfunction (Nyberg & Havstam, 2016). The stigmatizing impact of errors related to the place and manner of articulation highlights the need to further investigate the reliability of the PCP and PCM scores. Furthermore, the different studies have used different definitions of the PCC metric, which hampered a valid comparison of study results. As can be concluded from the limited literature, there is an urgent need to investigate the reliability of the different PCC metrics used in cleft palate speech intervention studies. A reliable measurement of the speech outcomes following speech intervention is essential in order to enhance evidence-based practice, to determine treatment efficacy, and to plan and monitor future interventions (Bessell et al., 2013; McCauley & Strand, 2008; Sell & Sweeney, 2020; World Health Organization, 2002).

In the literature, an alternative measurement tool has been described to circumvent the limitations of the PCC metric: the probe scoring system (Hall et al., 1998; Sell & Sweeney, 2020). Where the PCC metric uses a binary scoring system (i.e., “correct” or “incorrect” consonant), the probe score detects changes toward the correct production of the target consonant. A child's production of a phoneme is negatively scored (“−1”) according to the degree of mismatch from the correct target production with regard to the place, manner, and voice. Correct production of the target consonant is scored “0.” The different scores are summed up, which provides a total negative mismatch score. An intervention study performed by Hesketh et al. (2000) compared the effect of two treatment approaches (i.e., an articulation-based approach and a metaphonologically based approach) in 61 children with developmental phonological disorders but without a CP ± L using both the PCC score and the probe score as outcome measures. The children in the articulation-based approach and metaphonologically based approach groups only differed from each other on the probe score. This finding suggested that there might be some considerable differences in the sensitivity of both outcome measures. Unfortunately, Hall et al. (1998) and Hesketh et al. did not report information on the inter- and intrarater reliability of the probe scoring system. Despite the lack of reliability reports, Sell and Sweeney (2020) advised that speech analysis using a probe score might be an interesting way forward to capture changes following cleft palate speech intervention.

The current study evaluated and compared the reliability of different outcome measures (i.e., PCC-R score, PCP score, PCM score, and probe score) to assess consonant proficiency following cleft palate speech intervention. The following questions were addressed: The results of this study might aid to promote discussions on reliable outcome measure options in cleft palate speech intervention studies while taking clinical and practical aspects into consideration.

1. Is there a difference in the inter- and intrarater reliability of the PCC metrics (i.e., PCC-R score, PCP score, and PCM score) and the probe scoring system by an experienced rater and a less experienced rater?
2. Is there a difference in the interrater reliability of the PCC metrics and the probe scoring system between two experienced raters?
3. Is there a difference in the ability of the PCC metrics and the probe scoring system in terms of reflecting changes following cleft palate speech intervention?

## Method

This study was performed as part of a randomized controlled trial (RCT) that compared the effect of a motor-phonetic intervention (MPI) versus a linguistic-metaphonological intervention (LPI) in children with a CP ± L (Alighieri et al., 2020). This RCT was approved by the Ethics Committee of the (Ghent) University Hospital (2018/1218, Registration Number B670201837572).

### Data Acquisition

The data for this study were collected in the context of an RCT comparing the effect of MPI and LPI in children with a CP ± L (Alighieri et al., 2020). For this RCT, (Dutch-speaking)-speaking children with a CP ± L aged between 4 and 12 years were recruited from the multidisciplinary craniofacial team at (Ghent) Hospital. Inclusion criteria for these patients were (a) presence of a repaired CP ± L, (b) presence of at least one active or compensatory speech error based on the perceptual assessment performed by a speech-language pathologist (SLP), and (c) speaking (Dutch) as native language. Included patients were not allowed to receive any other type of speech and/or language intervention during the study period. In addition, patients were excluded if there was a presence of (a) cognitive and related learning disabilities or syndromes based on the patients' files and questioning the parents; (b) an oronasal fistula based on oral examination performed by an SLP; (c) velopharyngeal insufficiency based on videofluoroscopic examination, which was performed as part of the clinical evaluation; and (d) hearing disabilities based on pure-tone audiometry (> 25 dB HL) or the five screening questions that were developed by the ICHOM (Allori et al., 2017). In total, 14 children with a CP ± L (seven boys and seven girls), with a mean age of 7.71 years (*SD* = 2.281, range: 5.33–10.48 years), were included in the intervention study. Irrespective of the speech approach, each child received 10 one-hour treatment sessions divided over 10 consecutive working days. Treatment target consonants differed between the patients. A detailed description of the target consonants per child is provided in the previous study of our research unit, which compared the effects of MPI and LPI (Alighieri et al., 2020).

A speech sample at the word level and at the sentence level was collected at eight different data collection points during the study period (i.e., three pretreatment data points with a 1-week interval between each data point, one data point after 1 week of intervention, three posttreatment data points with a 1-week interval between each data point, and a 3-month follow-up data point; Bruneel et al., 2020). The speech sample at the word level included the 13 target high-pressure consonants of the (Dutch) language in all possible positions (i.e., /p/, /b/, /t/, /d/, /s/, /z/, /ʃ/, /ʒ/, /f/, /v/, /k/, /x/, /ɣ/). Since voiced high-pressure consonants are devoiced in word-final position in the (Dutch) language, voiced high-pressure consonants were only targeted in word-initial and word-medial position (Grijzenhout & Krämer, 2000). The sample at the sentence level included the same target consonants supplemented by a sentence targeting s-clusters (i.e., /st/, /sp/, /sk/). To allow for a more structured and standardized speech sample, the patients were asked to repeat the sentences produced by the SLP. At the word level, patients were asked to name pictures. If the patient was unable to produce the word spontaneously, a semantic cue was provided. If semantic cueing was found to be insufficient, phonological cueing was used. If the child was still not able to produce the word after the provision of these cues, the child was asked to repeat the word after modeling of the SLP.

The different speech samples were simultaneously audio- and video-recorded in a quiet room using a unidirectional condenser microphone (Samson CO1U) and a Sony Handycam HDR-CQ280E with a high-quality built-in microphone. In accordance with the recommendations of Lohmander et al. (2009), the examiner sat opposite the child with the camera operator just behind so that the camera lens was directly facing the child. All speech samples were anonymized and randomized. Data were collected by three different SLPs (E.V.G., L.B., and C.A.).

## Raters

During a first phase of the speech analyses, two raters were involved in the assessment. Rater 1 was an SLP with 5 years of scientific and clinical experience in assessing cleft-related speech disorders. She was officially trained in using the Cleft Audit Protocol for Speech–Augmented tool (John et al., 2006; Sell et al., 2009). Rater 1 transcribed 100% of the speech samples (*n* = 112) over a 3-month period. In addition, she reassessed 20% of the speech samples (*n* = 22) within a 1-month period in order to calculate intrarater reliability. The analysis of 134 speech samples (*n* = 112 original samples and *n* = 22 samples for intrarater reliability) was time-consuming considering that the rater spend approximately 1.5 hr per sample, which meant that she needed about 201 hr to analyze all the samples. The problem of time efficiency when conducting outcome studies has been reported before (Ahl & Harding-Bell, 2018; Sell & Sweeney, 2020; Shriberg & Lof, 1991). A student SLP, who was doing her master's thesis on cleft palate speech intervention, was involved in the project as second rater after following a training course. Rater 2 had limited experience in assessing cleft-related speech disorders. In addition to the 5-year master program in speech-language pathology, which included a master class targeting the perceptual assessment of speech disorders in patients with a CP ± L, she attended an additional 3-hr training course on the assessment of cleft-related speech disorders. Following this course, five randomly selected speech samples were analyzed and discussed between Rater 2 and the main author (C. A.) to verify Rater 2's compliance with the assessment protocol. Rater 2 also analyzed 134 speech samples (112 original samples and 22 samples for intrarater reliability). During a second phase of the speech analyses, a third rater with more experience was involved in the project. Rater 3 had 8 years of scientific and clinical experience in assessing cleft-related speech disorders. She was also officially trained in the use of the Cleft Audit Protocol for Speech–Augmented tool (John et al., 2006; Sell et al., 2009). Considering time restrictions, the third rater analyzed 30% of the speech samples, which were randomly selected (*n* = 34).

The three raters did not provide speech therapy to any of the included patients and were blinded to the treatment allocation. Each rater transcribed the speech samples at the word and sentence levels using narrow phonetic transcription based on the International Phonetic Alphabet (International Phonetic Association, 1999) and the International Phonetic Alphabet extensions as well as additional symbols developed to describe cleft-related articulation errors (Peterson-Falzone et al., 2016). The raters performed the assessments independently using overear headphones (Sennheiser EH 150 and Sennheiser Momentum). Guidelines with regard to the assessment procedure were provided to the raters. These guidelines included the definition of (in)correct targets, the transcription of the words and sentences, and the completion of the listener form (see Appendix A).

### Data Analysis: The PCC-R Metrics and Probe Scoring System

Based on the narrow phonetic transcriptions, the PCC-R scores at the word and sentence levels were calculated (Klintö et al., 2011; Lohmander & Persson, 2008; Shriberg et al., 1997). The PCC-R score was calculated by dividing the number of correctly produced consonants (numerator) by the total number of consonants elicited (denominator), multiplied by 100. In accordance with previous cleft palate speech intervention studies, consonants produced with a correct place, manner, and voice but with an (inter)dental quality or weak realization or accompanying nasal emission/nasal turbulence were considered correct (Sell & Sweeney, 2020). If cluster reduction or deletion of (final) consonants were present, the omitted consonant(s) were not counted within the total amount of target consonants elicited. PCP and PCM scores were calculated in accordance with the PCC-R score following the guidelines described by Klintö et al. (2011).

Additionally, probe scores were also calculated at the word and sentence levels. The child's production of each different target consonant was scored according to the degree of mismatch from the correct production of the target consonant based on the features “place of articulation,” “manner of articulation,” “voice,” and “nasal release” (Hall et al., 1998; Hesketh et al., 2000; Sell & Sweeney, 2020). As suggested by Sell and Sweeney (2020), the original probe scoring system was adapted for cleft palate speech, including the feature “nasal release.” A score of “−1” was given for each feature that did not match the correct production. If the child omitted the target, a score of “−4” was given. If the target consonant was produced correctly, a score of “0” was provided. The different scores were summed, providing a total negative mismatch score. For example, if the child produced an apico-alveolar active nasal fricative for the /s/, a total score of “−2” was given (i.e., “0” for the correct place of articulation, “−1” for an error in the manner of articulation, “0” for correct voicing, and “−1” for an error in oral release”). Appendix B provides an example of the PCC-R, PCP, and PCM scoring and the probe scoring system.

### Statistical Analysis

SPSS Version 26 (SPSS Corp.) was used for the statistical analysis of the data. Analyses were conducted at α = .05.

Two-way mixed ICCs (single measures) were calculated to assess the inter- and intrarater reliability of the PCC metrics and probe scores at the word and sentence levels. Point-to-point percentage agreement was not calculated, since it is easy to achieve high agreement by chance when there are only two scoring options (i.e., “correct” and “incorrect” for the PCC-R scores; Sell & Sweeney, 2020). ICCs were interpreted following the classification of Altman (1990; ICC < .20: poor, ICC = 0.21–0.40: fair, ICC = 0.41–0.60: moderate, ICC = 0.61–0.80: good, ICC = 0.81–1.00: very good).

The ability of the PCC metrics and the probe scores to reflect changes following speech intervention was measured using linear mixed models with the restricted maximum likelihood estimation and the Toeplitz covariance structure. This covariance structure was chosen based on comparison of the Akaike's information criterion values. Time, group (i.e., MPI or LPI), and Time × Group effects were specified as fixed factors. In light of the purpose of this study, the preintervention values were compared with the immediate postintervention values. A comparison of time within the two groups was determined using pairwise comparisons with Bonferroni corrections at $p < .025$ (.05/2). With regard to these preintervention values, comparison of the three baseline data points between the MPI and LPI groups revealed no statistically significant differences indicating that there was no baseline difference between these two groups ($p > .025$). The mean value of the three baseline measures was therefore calculated for each outcome variable and used for further analyses (Alighieri et al., 2020). Both unstandardized (i.e., simple) and standardized effect sizes were calculated (Baguley, 2009; Wilkinson & Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999). Unstandardized effect sizes were measured by providing the estimated mean (EM) differences and 95% confidence intervals (Baguley, 2009). Standardized effect sizes were calculated for the Time × Group interactions using Cohen's *d*'s dividing the EM difference by the standard deviation of a linear null model on the baseline data (Feingold, 2013).

## Results

### Inter- and Intrarater Reliability by an Experienced (Rater 1) and a Less Experienced (Rater 2) Rater

The results for the inter- and intrarater reliability for Raters 1 and 2 with regard to the PCC-R, PCP, and PCM scores and the probe scores are presented in Tables 1, 2, and 3. As summarized in Table 1, a fair interrater reliability was found between Raters 1 and 2 for the PCC-R, PCP, and PCM metrics at the word and sentence levels. The PCC-R, PCP, and PCM scores are determined by dividing the number of correctly produced consonants/the number of correctly produced consonants in terms of place of articulation/the number of correctly produced consonants in terms of manner of articulation (numerator) by the total number of consonants elicited (denominator), multiplied by 100.

The interrater reliability for the amount of “targets elicited” was fair at the word level and poor at the sentence level. Fair to moderate interrater reliability was observed for the amount of “targets correct” at the word and sentence levels, respectively. For the amount of “targets corrects in terms of place of articulation,” the interrater reliability was found to be fair at both the word and sentence levels. Interrater reliability for the amount of “targets correct in terms of manner of articulation” was fair at the word level and moderate at the sentence level. Concerning the probe score at the word and sentence levels, moderate interrater reliability was observed between Raters 1 and 2.

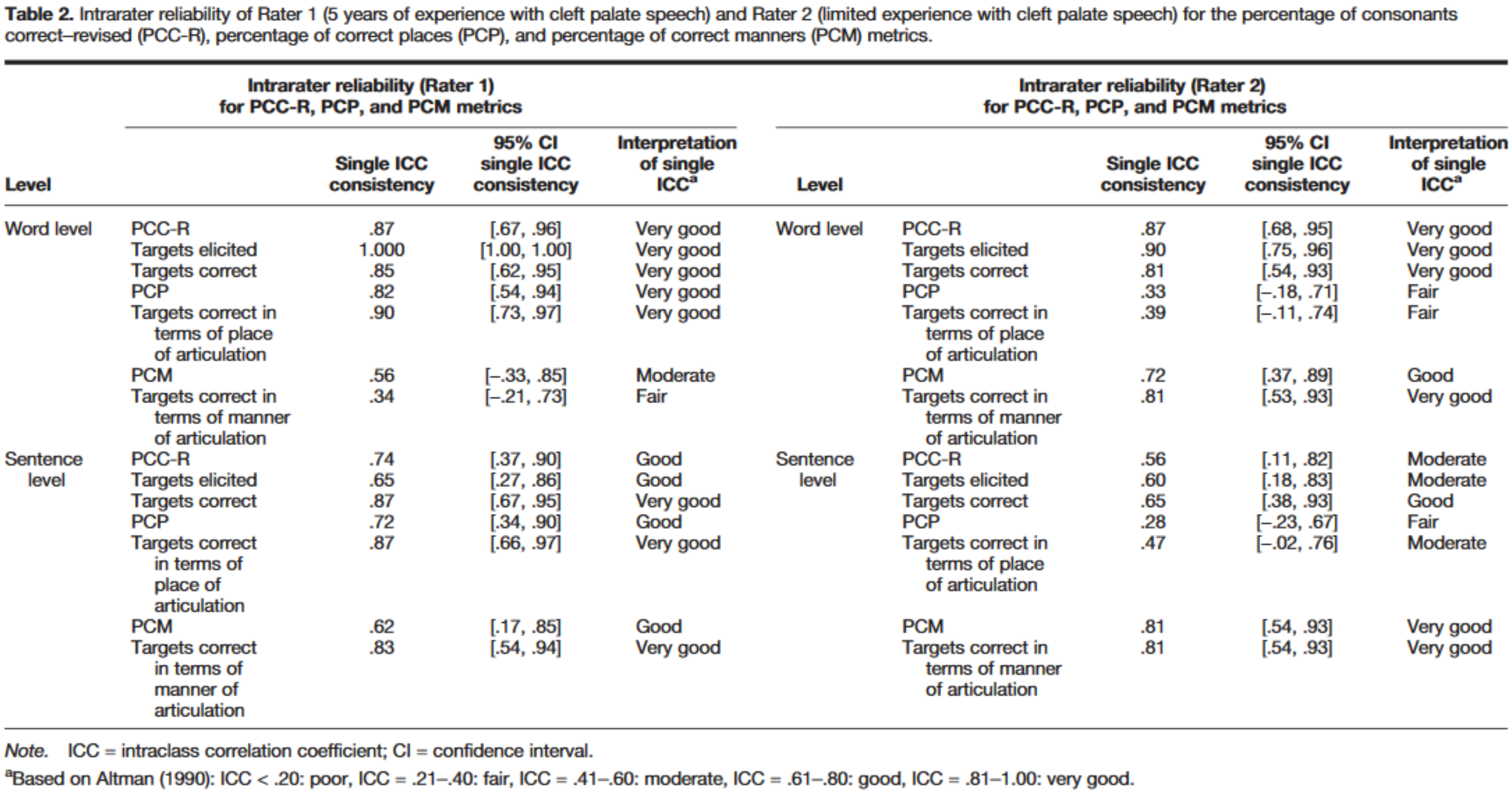
**Table 1.** Interrater reliability between Rater 1 (5 years of experience with cleft palate speech) and Rater 2 (limited experience with cleft palate speech) for the percentage of consonants correct–revised (PCC-R), percentage of correct places (PCP), and percentage of correct manners (PCM) metrics and the probe score.

| | Interrater reliability (Raters 1 and 2) for the PCC-R, PCP, and PCM metrics | | | | Interrater reliability (Raters 1 and 2) for the probe score | | | |
|---|---|---|---|---|---|---|---|---|
| **Level** | | **Single ICC consistency** | **95% CI single ICC consistency** | **Interpretation of single ICC[a]** | | **Single ICC consistency** | **95% CI single ICC consistency** | **Interpretation of single ICC[a]** |
| Word level | PCC-R | .36 | [–.02, .72] | Fair | Probe score on the word level | .60 | [.30, .80] | Moderate |
| | Targets elicited | .37 | [–.08, .70] | Fair | | | | |
| | Targets correct | .38 | [–.14, .69] | Fair | | | | |
| | PCP | .21 | [–.25, .59] | Fair | | | | |
| | Targets correct in terms of place of articulation | .22 | [–.23, .60] | Fair | | | | |
| | PCM | .39 | [–.05, .70] | Fair | | | | |
| | Targets correct in terms of manner of articulation | .35 | [–.10, .68] | Fair | | | | |
| Sentence level | PCC-R | .40 | [.11, .78] | Fair | Probe score on the sentence level | .59 | [.50, .77] | Moderate |
| | Targets elicited | .06 | [–.39, .48] | Poor | | | | |
| | Targets correct | .52 | [–.08, .72] | Moderate | | | | |
| | PCP | .34 | [–.11, .67] | Fair | | | | |
| | Targets correct in terms of place of articulation | .22 | [–.24, .60] | Fair | | | | |
| | PCM | .37 | [–.80, .69] | Fair | | | | |
| | Targets correct in terms of manner of articulation | .59 | [.19, .81] | Moderate | | | | |

*Note.* ICC = intraclass correlation coefficient; CI = confidence interval.

[a] Based on Altman (1990): ICC < .20: poor, ICC = .21–.40: fair, ICC = .41–.60: moderate, ICC = .61–.80: good, ICC = .81–1.00: very good.

**Table 2.** Intrarater reliability of Rater 1 (5 years of experience with cleft palate speech) and Rater 2 (limited experience with cleft palate speech) for the percentage of consonants correct–revised (PCC-R), percentage of correct places (PCP), and percentage of correct manners (PCM) metrics.



|  | Intrarater reliability (Rater 1) for PCC-R, PCP, and PCM metrics |  |  |  |  | Intrarater reliability (Rater 2) for PCC-R, PCP, and PCM metrics |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Level |  | Single ICC consistency | 95% CI single ICC consistency | Interpretation of single ICC[a] | Level |  | Single ICC consistency | 95% CI single ICC consistency | Interpretation of single ICC[a] |
| Word level | PCC-R | .87 | [.67, .96] | Very good | Word level | PCC-R | .87 | [.68, .95] | Very good |
|  | Targets elicited | 1.000 | [1.00, 1.00] | Very good |  | Targets elicited | .90 | [.75, .96] | Very good |
|  | Targets correct | .85 | [.62, .95] | Very good |  | Targets correct | .81 | [.54, .93] | Very good |
|  | PCP | .82 | [.54, .94] | Very good |  | PCP | .33 | [–.18, .71] | Fair |
|  | Targets correct in terms of place of articulation | .90 | [.73, .97] | Very good |  | Targets correct in terms of place of articulation | .39 | [–.11, .74] | Fair |
|  | PCM | .56 | [–.33, .85] | Moderate |  | PCM | .72 | [.37, .89] | Good |
|  | Targets correct in terms of manner of articulation | .34 | [–.21, .73] | Fair |  | Targets correct in terms of manner of articulation | .81 | [.53, .93] | Very good |
| Sentence level | PCC-R | .74 | [.37, .90] | Good | Sentence level | PCC-R | .56 | [.11, .82] | Moderate |
|  | Targets elicited | .65 | [.27, .86] | Good |  | Targets elicited | .60 | [.18, .83] | Moderate |
|  | Targets correct | .87 | [.67, .95] | Very good |  | Targets correct | .65 | [.38, .93] | Good |
|  | PCP | .72 | [.34, .90] | Good |  | PCP | .28 | [–.23, .67] | Fair |
|  | Targets correct in terms of place of articulation | .87 | [.66, .97] | Very good |  | Targets correct in terms of place of articulation | .47 | [–.02, .76] | Moderate |
|  | PCM | .62 | [.17, .85] | Good |  | PCM | .81 | [.54, .93] | Very good |
|  | Targets correct in terms of manner of articulation | .83 | [.54, .94] | Very good |  | Targets correct in terms of manner of articulation | .81 | [.54, .93] | Very good |

*Note.* ICC = intraclass correlation coefficient; CI = confidence interval.

[a]Based on Altman (1990): ICC < .20: poor, ICC = .21–.40: fair, ICC = .41–.60: moderate, ICC = .61–.80: good, ICC = .81–1.00: very good.

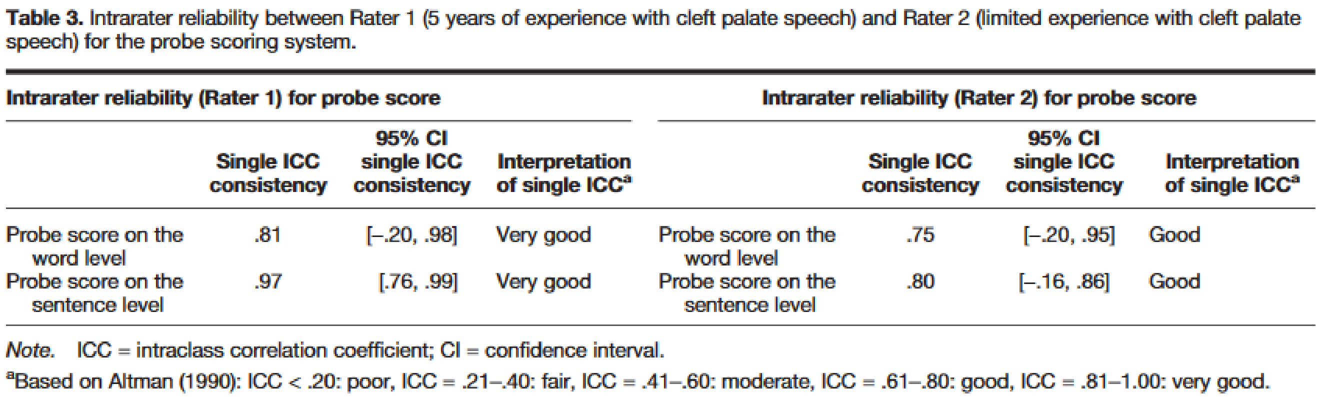


**Table 3.** Intrarater reliability between Rater 1 (5 years of experience with cleft palate speech) and Rater 2 (limited experience with cleft palate speech) for the probe scoring system.

| Intrarater reliability (Rater 1) for probe score | | | | Intrarater reliability (Rater 2) for probe score | | | |
|---|---|---|---|---|---|---|---|
| | Single ICC consistency | 95% CI single ICC consistency | Interpretation of single ICC[a] | | Single ICC consistency | 95% CI single ICC consistency | Interpretation of single ICC[a] |
| Probe score on the word level | .81 | [–.20, .98] | Very good | Probe score on the word level | .75 | [–.20, .95] | Good |
| Probe score on the sentence level | .97 | [.76, .99] | Very good | Probe score on the sentence level | .80 | [–.16, .86] | Good |

*Note.* ICC = intraclass correlation coefficient; CI = confidence interval.

[a] Based on Altman (1990): ICC < .20: poor, ICC = .21–.40: fair, ICC = .41–.60: moderate, ICC = .61–.80: good, ICC = .81–1.00: very good.

For Rater 1, intrarater reliability was good to very good for the PCC-R and PCP metrics at both the word and sentence levels (see Table 2). For the PCM metric, however, intrarater reliability was moderate at the word level and good at the sentence level. For the probe scores at the word and sentence levels, Rater 1 obtained very good intrarater reliability (see Table 3). For Rater 2, good to very good intrarater reliability was observed for the variables "PCC-R at the word level" and "PCM at the word and sentence level." For the PCP score at the word and sentence levels, intrarater reliability was fair. For the probe scores at the word and sentence levels, good intrarater reliability was observed for Rater 2 (see Table 3).

## Interrater Reliability Between Two Experienced Raters (Raters 1 and 3)

The results for the interrater reliability between Raters 1 and 3 with regard to the PCC-R metrics and the probe scores are presented in Table 4. For the PCC-R, PCP, and PCM metrics at the word and sentence levels, very good interrater reliability was found between Raters 1 and 3. At both the word and sentence levels, good reliability was observed for the variable "targets elicited," whereas very good reliability was found for the variable "targets correct." Interrater reliability was very good at both the word and sentence levels for the amount of targets correct in terms of place of articulation. For the variable "targets correct in terms of manner of articulation," good to very good interrater reliability was observed at the word and sentence levels, respectively. For the probe scores at the word and sentence levels, very good interrater reliability was observed between Raters 1 and 3.

## Reflection of Changes Following Speech Intervention

The results for the linear mixed models measuring the changes of the PCC metrics and the probe scores pre- and postintervention are presented in Tables 5 and 6. These results are based on the ratings performed by Rater 1.

Significant Time × Group interactions were revealed for the changes in the PCC-R scores at the word ($p$ = .010, $d$ = 1.73) and sentence ($p$ = .012, $d$ = 1.42) levels and the PCP scores at the word ($p$ = .009, $d$ = 0.97) and sentence ($p$ = .014, $d$ = .81) levels when comparing the preintervention data point with the immediate postintervention data point. The Time × Group interactions are marked by large effect sizes ($d$ > 0.8). Comparison of time within each group (i.e., preintervention vs. immediate postintervention) revealed that only those children who received LPI showed significantly larger PCC-R and PCP scores at the word and sentence levels following the intervention. Despite the observed increase in PCC-R scores at the word (EM = +15.33, $p$ = .037) and sentence (EM = +13.64, $p$ = .028) levels immediately after the intervention in the MPI group, this progress in the PCC-R scores was not statistically significant.

Considering the results for the probe scoring system, significant Time × Group interactions were revealed for the changes in the probe scores at the word ($p$ = .030, $d$ = 1.07) and sentence ($p$ = .016, $d$ = 1.49) levels. Comparison of time within each group revealed that the probe scores of both the children in the MPI and LPI groups were significantly higher (i.e., the scores improved) following speech intervention (see Table 6). In accordance with the PCC metrics, the Time × Group interactions for the probe scores are marked by large effect sizes ($d$ > 0.8).

**Table 4.** Interrater reliability between Rater 1 (5 years of experience with cleft palate speech) and Rater 3 (8 years of experience with cleft palate speech) for the percentage of consonants correct–revised (PCC-R), percentage of correct places (PCP), and percentage of correct manners (PCM) metrics and the probe score.

| | | Interrater reliability (Raters 1 and 3) for the PCC-R, PCP, and PCM metrics | | | Interrater reliability (Raters 1 and 3) for the probe score | | | |
|---|---|---|---|---|---|---|---|---|
| **Level** | | **Single ICC consistency** | **95% CI single ICC consistency** | **Interpretation of single ICC[a]** | | **Single ICC consistency** | **95% CI single ICC consistency** | **Interpretation of single ICC[a]** |
| Word level | PCC-R | .81 | [.53, .93] | Very good | Probe score on the word level | .95 | [.89, .98] | Very good |
| | Targets elicited | .63 | [.24, .85] | Good | | | | |
| | Targets correct | .96 | [.87, .99] | Very good | | | | |
| | PCP | .88 | [.67, .96] | Very good | | | | |
| | Targets correct in terms of place of articulation | .92 | [.75, .98] | Very good | | | | |
| | PCM | .72 | [.34, .90] | Good | | | | |
| | Targets correct in terms of manner of articulation | .76 | [.46, .91] | Good | | | | |
| Sentence level | PCC-R | .82 | [.54, .93] | Very good | Probe score on the sentence level | .94 | [.79, .97] | Very good |
| | Targets elicited | .65 | [.27, .86] | Good | | | | |
| | Targets correct | .86 | [.63, .95] | Very good | | | | |
| | PCP | .87 | [.66, .96] | Very good | | | | |
| | Targets correct in terms of place of articulation | .92 | [.79, .97] | Very good | | | | |
| | PCM | .90 | [.75, .97] | Very good | | | | |
| | Targets correct in terms of manner of articulation | .90 | [.73, .96] | Very good | | | | |

*Note.* ICC = intraclass correlation coefficient; CI = confidence interval.

[a] Based on Altman (1990): ICC < .20: poor, ICC = .21–.40: fair, ICC = .41–.60: moderate, ICC = .61–.80: good, ICC = .81–1.00: very good.

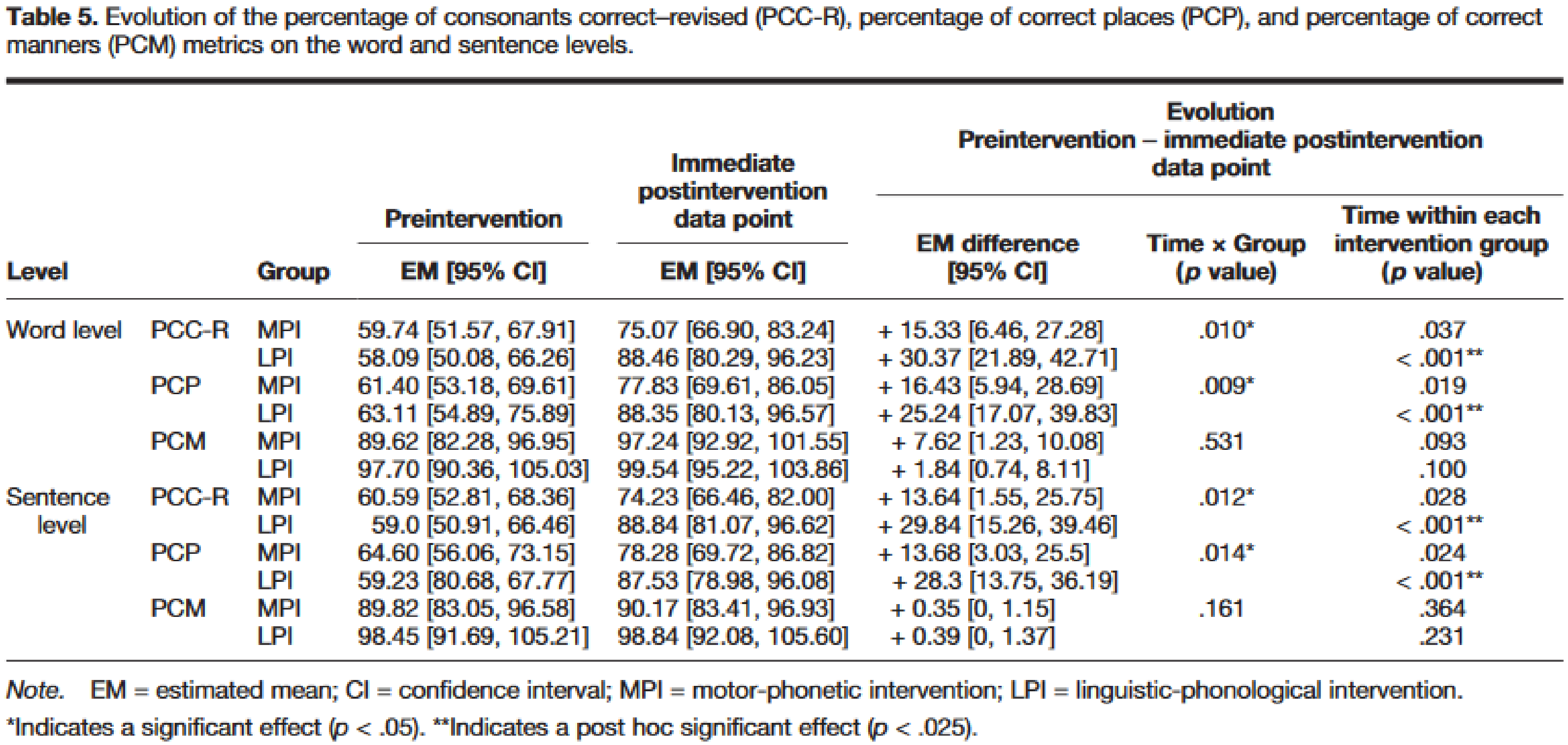


**Table 5.** Evolution of the percentage of consonants correct–revised (PCC-R), percentage of correct places (PCP), and percentage of correct manners (PCM) metrics on the word and sentence levels.

| | | | | | Evolution<br>Preintervention – immediate postintervention data point | | |
|---|---|---|---|---|---|---|---|
| | | | Preintervention | Immediate postintervention data point | | | |
| Level | | Group | EM [95% CI] | EM [95% CI] | EM difference [95% CI] | Time × Group ($p$ value) | Time within each intervention group ($p$ value) |
| Word level | PCC-R | MPI | 59.74 [51.57, 67.91] | 75.07 [66.90, 83.24] | + 15.33 [6.46, 27.28] | .010* | .037 |
| | | LPI | 58.09 [50.08, 66.26] | 88.46 [80.29, 96.23] | + 30.37 [21.89, 42.71] | | < .001** |
| | PCP | MPI | 61.40 [53.18, 69.61] | 77.83 [69.61, 86.05] | + 16.43 [5.94, 28.69] | .009* | .019 |
| | | LPI | 63.11 [54.89, 75.89] | 88.35 [80.13, 96.57] | + 25.24 [17.07, 39.83] | | < .001** |
| | PCM | MPI | 89.62 [82.28, 96.95] | 97.24 [92.92, 101.55] | + 7.62 [1.23, 10.08] | .531 | .093 |
| | | LPI | 97.70 [90.36, 105.03] | 99.54 [95.22, 103.86] | + 1.84 [0.74, 8.11] | | .100 |
| Sentence level | PCC-R | MPI | 60.59 [52.81, 68.36] | 74.23 [66.46, 82.00] | + 13.64 [1.55, 25.75] | .012* | .028 |
| | | LPI | 59.0 [50.91, 66.46] | 88.84 [81.07, 96.62] | + 29.84 [15.26, 39.46] | | < .001** |
| | PCP | MPI | 64.60 [56.06, 73.15] | 78.28 [69.72, 86.82] | + 13.68 [3.03, 25.5] | .014* | .024 |
| | | LPI | 59.23 [80.68, 67.77] | 87.53 [78.98, 96.08] | + 28.3 [13.75, 36.19] | | < .001** |
| | PCM | MPI | 89.82 [83.05, 96.58] | 90.17 [83.41, 96.93] | + 0.35 [0, 1.15] | .161 | .364 |
| | | LPI | 98.45 [91.69, 105.21] | 98.84 [92.08, 105.60] | + 0.39 [0, 1.37] | | .231 |

*Note.* EM = estimated mean; CI = confidence interval; MPI = motor-phonetic intervention; LPI = linguistic-phonological intervention.
*Indicates a significant effect ($p < .05$). **Indicates a post hoc significant effect ($p < .025$).

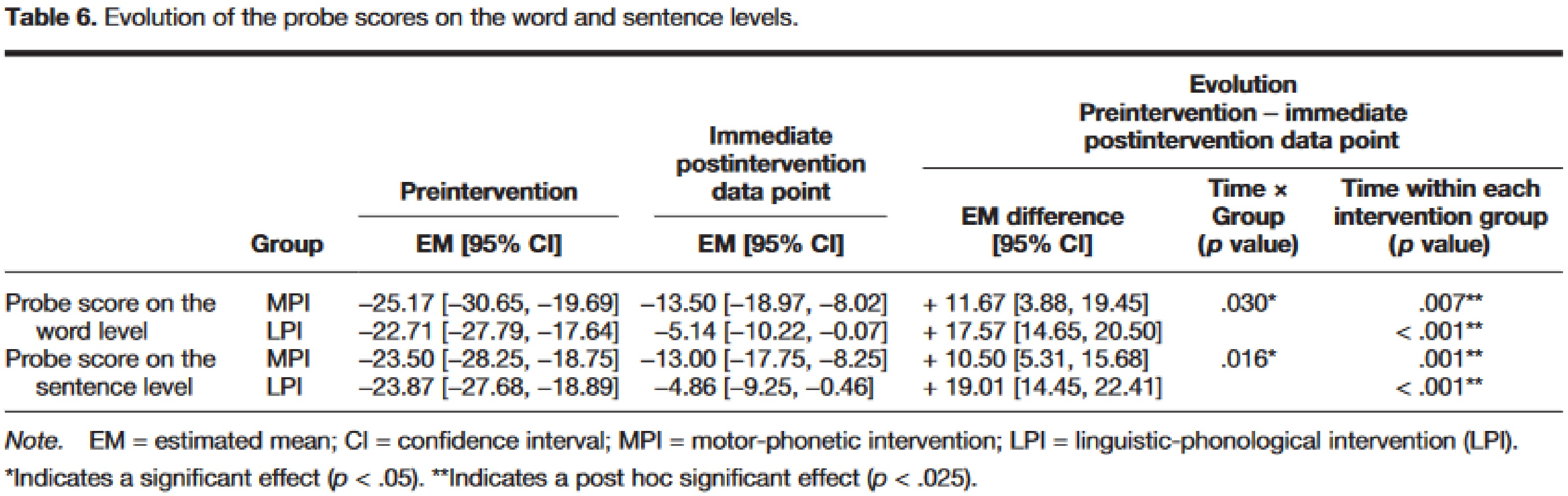


**Table 6.** Evolution of the probe scores on the word and sentence levels.

| | | | | Evolution<br>Preintervention – immediate postintervention data point | | |
|---|---|---|---|---|---|---|
| | | Preintervention | Immediate postintervention data point | EM difference [95% CI] | Time × Group ($p$ value) | Time within each intervention group ($p$ value) |
| | Group | EM [95% CI] | EM [95% CI] | | | |
| Probe score on the word level | MPI | –25.17 [–30.65, –19.69] | –13.50 [–18.97, –8.02] | + 11.67 [3.88, 19.45] | .030* | .007** |
| | LPI | –22.71 [–27.79, –17.64] | –5.14 [–10.22, –0.07] | + 17.57 [14.65, 20.50] | | < .001** |
| Probe score on the sentence level | MPI | –23.50 [–28.25, –18.75] | –13.00 [–17.75, –8.25] | + 10.50 [5.31, 15.68] | .016* | .001** |
| | LPI | –23.87 [–27.68, –18.89] | –4.86 [–9.25, –0.46] | + 19.01 [14.45, 22.41] | | < .001** |

*Note.* EM = estimated mean; CI = confidence interval; MPI = motor-phonetic intervention; LPI = linguistic-phonological intervention (LPI). *Indicates a significant effect ($p < .05$). **Indicates a post hoc significant effect ($p < .025$).

## Discussion

The PCC metrics have been criticized over the years, especially in terms of their usefulness in cleft palate speech intervention studies (Sell & Sweeney, 2020). To circumvent the limitations of the PCC metrics, several authors suggested the use of a new scoring system: the probe score (Hall et al., 1998; Hesketh et al., 2000; Sell & Sweeney, 2020; Sweeney et al., 2020). Unfortunately, no study yet reported on the reliability of this probe scoring system. The current study evaluated and compared the reliability of different outcome measures (i.e., PCC-R score, PCP score, PCM score, and probe score) to assess consonant proficiency following cleft palate speech intervention. In addition, the outcome measures' ability to reflect changes following speech intervention was determined.

### Inter- and Intrarater Reliability by an Experienced and a Less Experienced Rater

During Phase 1 of the speech sample analyses, two raters were involved in the assessment: an experienced rater (Rater 1) and an unexperienced rater (Rater 2). In accordance with the recommendations of Sell and Sweeney (2020), we calculated the interrater reliability for the amount of targets elicited and the amount of targets correct. For these variables, the interrater reliability ranged from poor to fair. Interestingly, the interrater reliability for the amount of targets correct at the sentence level was notably higher than the reliability for the amount of targets elicited (see Table 1). As suggested by Sell and Sweeney, this finding might perhaps be explained by "chance agreement." One might argue that there was a high risk for agreement by chance for the amount of targets correct since there were only two scoring options per target (i.e., "correct" or "incorrect"). To date, only one study reported reliability results on the amount of targets elicited and the amount of targets correct (Sell & Sweeney, 2020).

In general, it must be noted that Rater 2, even though she followed a training course before the onset of the study, had limited experience with the assessment of cleft-related speech disorders. Several authors demonstrated the positive effect of training on interrater reliability, although interrater reliability did not increase for several cleft-related speech errors (Chapman et al., 2016; Gooch et al., 2001; John et al., 2006; Lohmander et al., 2009; Sell et al., 2009; Willadsen et al., 2017). It was argued that limited time was spent on certain speech errors during training. In addition, the raters' experience with these speech errors prior to the training (i.e., more or less experience) was mentioned as a variable affecting interrater reliability. These findings highlight that training packages must pay more attention to assessing cleft-related speech errors. In addition, training programs should involve consensus listening and should provide raters with reference samples of the different cleft-related speech errors (Chapman et al., 2016; John et al., 2006; Sell et al., 2009).

Rater 2 almost consistently transcribed the production of active nasal fricatives as consonants produced with a correct place, manner, and voice but with accompanying nasal turbulence. Thereby, these consonants were scored as correct instead of incorrect. Nyberg and Havstam (2016) argued that speech errors related to the place and manner of articulation (e.g., the production of active nasal fricatives affecting the manner of articulation) are more stigmatizing than soft signs of velopharyngeal dysfunction (e.g., accompanying nasal turbulence). The production of active nasal fricatives is an example of an active or compensatory speech error, whereas accompanying nasal turbulence on consonant productions is often a sign of incomplete velopharyngeal closure (Harding & Grunwell, 1998; Zajac, 2015). It is important to make a distinction between these two types of errors. To eliminate active or compensatory speech errors, speech therapy is required to correct wrong place or manner of articulation (Harding & Grunwell, 1998; Kummer, 2011b). In contrast, further investigating of velopharyngeal functioning is required in cases where accompanying nasal turbulence on consonant productions is observed (Kummer, 2011a). Thereby, confusion between an active nasal fricative and accompanying nasal turbulence can have some serious implications in terms of planning appropriate intervention. The poor interrater reliability for the variable "targets elicited" was mainly influenced by differences in the transcription of glottal stops. Rater 1 transcribed such errors as glottal stops, whereas Rater 2 transcribed these errors as a deletion of the target consonant. This issue has been reported before by several authors (Gooch et al., 2001; Sell & Sweeney, 2020; Willadsen et al., 2017). When a child produces a glottal stop, the airflow is obstructed in the glottis. Because of this airflow obstruction, the glottal vibration either stops or becomes irregular with a sudden drop in vibration intensity (Umeda, 1978). Especially for less experienced raters, the absence of glottal vibration can be confused with a consonant omission. Interestingly, different studies have different viewpoints on the transcription of glottal stops as consonant omissions. Willadsen et al. (2017), for example, assessed consonant proficiency in 391 five-year-old children with a unilateral cleft lip and palate. A picture-naming test was phonetically transcribed by two SLPs. Training of phonetic transcription was carried out during the project. Despite that the two raters followed common training, lower than acceptable levels of inter- and intrarater agreement were observed. To increase the reliability, disagreements that were considered minor were counted as agreements in the inter- and intrarater agreement scores. Interestingly, a difference between a glottal stop and an omission of a target consonant was considered minor and was thus accepted. Sell and Sweeney (2020), in contrast, argued that glottal stops are cleft speech characteristics that are considered serious errors of articulation. In cleft palate speech intervention, it will be much more challenging to eliminate glottal stops compared to omissions of a target consonant (Sell & Sweeney, 2020). In our intervention study, these errors (i.e., transcription of glottal stops as consonant omissions) also decreased reliability. In eight of the 14 included children in the intervention study (57.2%), elimination of glottal stops was the treatment goal. Thereby, we considered this difference in transcription as a serious problem influencing the outcomes of the intervention. In our opinion, researchers investigating the effect of speech intervention in children with a CP ± L must carefully pay attention to the transcription of these types of errors when considering the reliability results.

Despite the reported reliability issues, it should be noted that Rater 2 obtained good to very good intrarater reliability for the PCC-R and PCM scores at the word and sentence levels. Interestingly, considerably lower reliability results were observed for the PCP score at the word and sentence levels (see Table 2). Especially with regard to the assessment of the amount of targets correct in terms of place of articulation, there seemed to be an effect of listener's experience. While Rater 1 (rater with more experience) showed very good intrarater reliability for the amount of targets correct in terms of place of articulation at the word and sentence level, Rater 2 (with less experience) showed fair to moderate intrarater reliability for this variable. Both clinicians and researchers should pay attention to this issue especially in the context of cleft palate speech intervention aiming to correct wrong articulatory place or manner (Harding & Grunwell, 1998; Kummer, 2011b). Despite that our study was one of the first to report reliability results on the PCP metric, these findings regarding the assessment of the place of articulation seem to confirm previous research. Gooch et al. (2001) and Santelmann et al. (1999), for example, argued that active or compensatory speech errors (which affect place or manner of articulation) may lead to low transcriber agreement.

### Interrater Reliability Between Two Experienced Raters

With regard to the PCC-R metric, interrater reliability between Raters 1 and 3 ranged from good to very good (see Table 4). In accordance, very good interrater reliability was observed for both the probe scores at the word and sentence levels (see Table 4). Based on a qualitative comparison of these ICC values, this study found no differences in the interrater reliability for the different PCC metrics and the probe scoring system between two experienced raters. This finding suggests that both outcome measures are suitable for application in cleft palate speech intervention studies. Raters 1 and 3 had been working together for 5 years in the same clinical and research setting, and thus, they might have had comparable internal standards, which might explain the high interrater reliability. As with Raters 1 and 2, divergent reliability results for the amount of targets correct and the amount of targets elicited were observed between Raters 1 and 3. However, it should be noted that this difference was not as pronounced (i.e., "good" reliability for the amount of targets elicited and "very good" reliability for the amount of targets correct), as found by Sell and Sweeney (2020). Latter study found unacceptably poor interrater reliability between two expert raters on the number of targets elicited at both the word (ICC = 0.07, 95% CI [−0.17, 0.30]) and sentence (ICC = 0.42, 95% CI [0.20, 0.59]) levels. Several factors might account for these divergent results. Even though Sell and Sweeney also used experienced raters who had additional training, the different raters' previous experience with and amount of training on the assessed parameters prior to this study might have played a role in the observed difference (Chapman et al., 2016). In addition, a valid comparison between our results and the results reported by Sell and Sweeney is hampered because they did not report which type of ICC measure they used. Since average-measure ICCs tend to be higher than single-measure ICCs (Hallgren, 2012), it is very important to describe the used methodology especially in the context of reliability studies. Despite the methodological discrepancies, it should be acknowledged that differences in the reliability between the amount of targets correct and the amount of targets elicited are issues (Sell & Sweeney, 2020). Our findings are in line with the reports by Sell and Sweeney: Even though two expert raters have high interrater reliability with regard to the PCC-R score, this score could be calculated on a different number of elicited consonants, which may result in an under- or overestimation of the actual PCC-R score. This issue raises some concerns especially in cleft palate speech intervention studies that lack reliability reports on the number of elicited target consonants. Reporting the reliability of the amount of targets elicited and the amount of targets correct for all the different PCC measures, including the PCC-R, PCP, and PCM metrics, is necessary in order to adequately interpret these scores. In contrast, if such reports are absent, one must be cautious when interpreting the reliability of the PCC metric (Sell & Sweeney, 2020).

### Reflection of Changes Following Speech Intervention

This study was the first to compare two outcome measures for consonant proficiency (i.e., the PCC metrics and the probe scoring system) in terms of reflecting changes following cleft palate speech intervention. Using the PCC-R and PCP metrics, only the LPI group was observed to present with statistically significant higher scores following speech intervention (see Table 5). Using the probe scoring system, in contrast, both the children in the MPI and LPI groups had statistically significant higher scores following the intervention (see Table 6). This finding suggests that the probe scoring system demonstrates a greater ability to detect changes toward the correct production of target consonants compared to the PCC metrics (more specifically, the PCC-R and PCP scores). Unfortunately, comparison of our results with previous studies is difficult given that no study yet compared the ability of these two scores to capture changes following intervention. Nevertheless, Sell and Sweeney (2020) hypothesized that the probe score would be able to show (smaller) degrees of progress toward the correct production of a target in contrast to the PCC metrics. The PCC metrics were reported to be less sensitive for positive changes toward a specific target because of the use of a binary scoring system (i.e., correct/incorrect; Hall et al., 1998; Hesketh et al., 2000; Sell & Sweeney, 2020). This study seems to confirm this hypothesis. In addition, the results of our study also confirm that the PCP and PCM scores might still be too restricted despite that these scores were developed in an effort to detect positive changes in terms of place and manner of articulation (Lohmander & Persson, 2008). Nevertheless, in the absence of probe scores, the PCP and PCM metrics provide important and useful information that can support the interpretation of the PCC-R metrics with regard to the correctness of the place and manner of articulation of the target consonants.

In summary, this study found good to very good interrater reliability between two expert raters for both the PCC-R metrics and the probe scoring system. One might question which outcome measure is the most appropriate for the use in both clinical environments and research settings. In the authors' opinion, the most important strength of the PCC-R metric is that percentages are easy to grasp which facilitates straightforward communication not only between members of the cleft team but also with the patients and their family. The probe scoring system, on the contrary, may be less convenient for communication in a clinical setting since a score of "0" reflects a correct production of the target consonant. It must be noted, however, that from a clinical point of view, it might also be interesting to calculate a child's probe score per target consonant in order to capture changes following speech intervention in more detail (Rvachew & Nowak, 2001). For research purposes, this approach would probably be too time-consuming. It should also be noted that cluster reduction and deletion of (final) consonants are assessed differently when using the PCC-R metric or the probe scoring system. Using the PCC-R metric, an omitted consonant is not counted within the total amount of target consonants elicited. One can thus argue that the PCC score somewhat ignores these type of errors. According to the probe scoring system, a score of "−4" should be given when a consonant was omitted. In our opinion, one must determine the importance of consonant deletions in light of the specific research purpose or clinical context before choosing either the PCC-R metric or the probe scoring system as an outcome measure. For example, a study investigating phonological disorders in children with a CP ± L might attach great importance to consonant omissions. In this case, the use of the PCC-R metric seems less appropriate. The decision to opt for one or another outcome measure should be made in light of the benefits and drawbacks of each scoring system.

### Limitations and Suggestions for Further Research

A common issue when conducting outcome studies is the problem of time efficiency (Ahl & Harding-Bell, 2018; Sell & Sweeney, 2020; Shriberg & Lof, 1991). Especially in intervention studies with multiple pre- and postintervention data collection points, a large amount of speech samples is collected, which need to be phonetically transcribed. We experienced this problem when conducting the intervention study that included eight different data collection points per participant (Alighieri et al., 2020). During the first phase of speech analyses, two raters were included (i.e., Rater 1 with experience and Rater 2 with less experience). During a second phase of the speech analyses, a third rater with more experience was involved in the project. Considering time restrictions, Rater 3 only analyzed 30% of the speech samples. Rater 1 spend approximately 90 min per sample, which is considerably longer than the report by Sell and Sweeney (2020), who noted that the speech analyses took between 20 and 40 min per sample. Because of the lack of data on time, it is not clear whether the analysis of the probe score resulted in this difference. Considering the importance of time-related aspects when conducting outcome studies, it would be interesting to include data on time related to the use of the PCC metrics and the probe scoring system. This information could help inform also a time-efficient outcome measure in outcome studies. Another limitation of this study is the lack of reports on the intrarater reliability of Rater 3. In addition, the speech samples assessed by Rater 3 were randomly selected. Unfortunately, we did not account for the severity of the speech disorders. Future studies should respond to this limitation by ensuring that a wide range of severity of speech is included. Despite that the findings of this study suggested that the probe scoring system demonstrates a greater ability to detect changes toward the correct production of target consonants compared to the PCC metrics, we highlight that further research on this matter is necessary in order to confirm these results.

This study measured the ability of the PCC metrics and the probe scoring system to reflect changes following cleft palate speech intervention. Nevertheless, to be able to investigate the effect of the intervention, future studies can compare the postintervention PCC scores with normative data.

## Conclusions

This innovative study was the first to compare the inter- and intrarater reliability of different PCC metrics (i.e., PCC-R, PCP, and PCM score) and the probe scoring system as outcome measures of consonant proficiency in a cleft palate speech intervention study. Having experience with the assessment of cleft-related speech disorders was observed to be a crucial factor to gain reliable results. This study found no differences in the interrater reliability for the different PCC metrics and the probe scoring system between two experienced raters. This finding suggests that both outcome measures are suitable for application in cleft palate speech intervention studies. Despite that the findings of this study suggested that the probe scoring system demonstrates a greater ability to detect changes toward the correct production of target consonants compared to the PCC metrics, further research on the probe score is necessary to lead to more understanding of how this system is beneficial in both research studies and clinical work. In the meantime, the decision to opt for one or another outcome measure should be made in light of the specific study purposes and the benefits and drawbacks of each scoring system.

## Acknowledgments

The first author was funded by a Ph.D. grant of the Research Fund Flanders (1120919N).

## References

Ahl, R., & Harding-Bell, A. (2018). Comparing methodologies in a series of speech outcome studies: Challenges and lessons learned. *The Cleft Palate–Craniofacial Journal*, 55(1), 35–44. https://doi.org/10.1177/1055665617718546

Alighieri, C., Bettens, K., Bruneel, L., D’haeseleer, E., Van Gaever, E., & Van Lierde, K. (2020). Effectiveness of speech intervention in patients with a cleft palate: Comparison of motor-phonetic versus linguistic-phonological speech approaches. *Journal of Speech,* *Language, and Hearing Research*, 63(12), 3909–3933. https://doi.org/10.1044/2020_JSLHR-20-00129

Allori, A. C., Kelley, T., Meara, J. G., Albert, A., Bonanthaya, K., Chapman, K., Cunningham, M., Daskalogiannakis, J., de Gier, H., Heggie, A. A., Hernandez, C., Jackson, O., Jones, Y., Kangesu, L., Koudstaal, M. J., Kuchhal, R., Lohmander, A., Long, R. E., Jr., Magee, L., . . . Wong, K. W. (2017). A standard set of outcome measures for the comprehensive appraisal of cleft care. *The Cleft Palate–Craniofacial Journal*, 54(5), 540–554. https://doi.org/10.1597/15-292

Altman, D. G. (1990). *Practical statistics for medical research*. CRC Press. https://doi.org/10.1201/9780429258589

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology,* 100(Pt. 3), 603–617. https://doi.org/10.1348/000712608x377117

Bessell, A., Sell, D., Whiting, P., Roulstone, S., Albery, L., Persson, M., Verhoeven, A., Burke, M., & Ness, A. (2013). Speech and language therapy interventions for children with cleft palate: A systematic review. *The Cleft Palate–Craniofacial Journal*, 50(1), 1–17. https://doi.org/10.1597/11-202

Bruneel, L., Bettens, K., De Bodt, M., D'haeseleer, E., Thijs, Z., Roche, N., & Van Lierde, K. (2020). Stages in the development and validation of a Belgian Dutch outcome tool for the perceptual evaluation of speech in patients with cleft palate. *The Cleft Palate–Craniofacial Journal,* 57(1), 43–54. https://doi.org/10.1177/1055665619862726

Brunnegård, K., Hagberg, E., Havstam, C., Okhiria, Å., & Klintö, K. (2020). Reliability of speech variables and speech-related quality indicators in the Swedish cleft lip and palate registry. *The Cleft Palate–Craniofacial Journal*, 57(6), 715–722. https://doi.org/10.1177/1055665619894497

Brunnegård, 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. *International Journal of Language & Communication Disorders,* 44(5), 656–674. https://doi.org/10.1080/13682820802295203

Chapman, K. L., Baylis, A., Trost-Cardamone, J., Cordero, K. N., Dixon, A., Dobbelsteyn, C., Thurmes, A., Wilson, K., Harding-Bell, A., Sweeney, T., Stoddard, G., & Sell, D. (2016). The Americleft speech project: A training and reliability study. *The Cleft Palate–Craniofacial Journal*, 53(1), 93–108. https://doi.org/10.1597/14-027

Cordes, A. K. (1994). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech and Hearing Research*, 37(2), 264–278. https://doi.org/10.1044/jshr.3702.264

Dobbelsteyn, C., Kay-Raining Bird, E., Parker, J., Griffiths, C., Budden, A., Flood, K., & Stilson, A. (2014). Effectiveness of the corrective babbling speech treatment program for children with a history of cleft palate or velopharyngeal dysfunction. *The Cleft Palate–Craniofacial Journal*, 51(2), 129–144. https://doi.org/10.1597/12-188

Feingold, A. (2013). A regression framework for effect size assessments in longitudinal modeling of group differences. *Review of General Psychology*, 17(1), 111–121. https://doi.org/10.1037/a0030048

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. https://doi.org/10.1597/1545-1569_2001_038_0059_roltoc_2.0.co_2

Grijzenhout, J., & Krämer, M. (2000). Final devoicing and voicing assimilation in Dutch derivation and cliticization. In B. Stiebels & D. Wunderlich (Eds.), *Lexicon in focus* (pp. 55–82). Akademie Verlag. https://doi.org/10.1515/9783050073712-004

Hall, R., Adams, C., Hesketh, A., & Nightingale, K. (1998). The measurement of intervention effects in developmental phonological disorders. *International Journal of Language & Communication Disorders*, 33(S1), 445–450. https://doi.org/10.3109/13682829809179466

Hallgren, K. A. (2012). Computing interrater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Harding, A., & Grunwell, P. (1998). Active versus passive cleft-type speech characteristics. *International Journal of Language & Communication Disorders*, 33(3), 329–352. https://doi.org/10.1080/136828298247776

Henningsson, G., Kuehn, D., 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. https://doi.org/10.1597/06-086.1

Hesketh, A., Adams, C., Nightingale, R., & Hall, A. (2000). Phonological awareness therapy and articulatory training approaches for children with phonological disorders: A comparative outcome study. *International Journal of Language & Communication Disorders*, 35(3), 337–354. https://doi.org/10.1080/136828200410618

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

James, D. G., Ferguson, W. A., & Butcher, A. (2016). Assessing children's speech using picture-naming: The influence of differing phonological variables on some speech outcomes. *International Journal of Speech-Language Pathology,* 18(4), 364–377. https://doi.org/10.3109/17549507.2015.1101159

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. https://doi.org/10.1597/04-141.1

Johnson, C. A., Weston, A. D., & Bain, B. A. (2004). An objective and time-efficient method for determining severity of childhood speech delay. *American Journal of Speech-Language Pathology*, 13(1), 55–65. https://doi.org/10.1044/1058-0360(2004/007)

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech. *American Journal of Speech-Language Pathology*, 3(2), 81–95. https://doi.org/10.1044/1058-0360.0302.81

Keuning, K. H., Wieneke, G. H., & Dejonckere, P. H. (1999). The intrajudge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: The effect of judges and speech samples. *The Cleft Palate–Craniofacial Journal*, 36(4), 328–333. https://doi.org/10.1597/1545-1569_1999_036_0328_tirotp_2.3.co_2

Klintö, K., Falk, E., Wilhelmsson, S., Schönmeyr, B., & Becker, M. (2018). Speech in 5-year-olds with cleft palate with or without cleft lip treated with primary palatal surgery with muscle reconstruction according to Sommerlad. *The Cleft Palate–Craniofacial Journal*, 55(10), 1399–1408. https://doi.org/10.1177/1055665618768541

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(3), 348–360. https://doi.org/10.3109/13682822.2010.507615

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

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. https://doi.org/10.1044/jslhr.4101.51

Kummer, A. W. (2011a). Disorders of resonance and airflow secondary to cleft palate and/or velopharyngeal dysfunction. *Seminars in Speech and Language*, 32(2), 141–149. https://doi.org/10.1055/s-0031-1277716

Kummer, A. W. (2011b). Speech therapy for errors secondary to cleft palate and velopharyngeal dysfunction. *Seminars in Speech and Language*, 32(2), 191–198. https://doi.org/10.1055/s-0031-1277721

Lewis, K. E., Watterson, T. L., & Houghton, S. M. (2003). The influence of listener experience and academic training on ratings of nasality. *Journal of Communication Disorders,* 36(1), 49–58. https://doi.org/10.1016/S0021-9924(02)00134-X

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. https://doi.org/10.1597/06-123.1

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. https://doi.org/10.1597/08-039.1

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*, 55(8), 1051–1059. https://doi.org/10.1177/1055665618765777

McCauley, R. J., & Strand, E. A. (2008). A review of standardized tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology*, 17(1), 81–91. https://doi.org/10.1044/1058-0360(2008/007)

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. https://doi.org/10.1044/1092-4388(2011/10-0130)

Nyberg, J., & Havstam, C. (2016). Speech in 10-year-olds born with cleft lip and palate: What do peers say? *The Cleft Palate–Craniofacial Journal*, 53(5), 516–526. https://doi.org/10.1597/15-140

Peterson-Falzone, S., Hardin-Jones, M., & Karnell, M. (2001). *Cleft palate speech*. Mosby.

Peterson-Falzone, S., Trost-Cardamone, J., Karnell, M., & Hardin-Jones, M. (2016). *The clinician's guide to treating cleft palate speech*. Elsevier.

Rvachew, S., & Nowak, M. (2001). The effect of target-selection strategy on phonological learning. *Journal of Speech, Language, and Hearing Research*, 44(3), 610–623. https://doi.org/10.1044/1092-4388(2001/050)

Santelmann, L., Sussman, J., & Chapman, K. (1999). Perception of middorsum palatal stops from the speech of three children with repaired cleft palate. *The Cleft Palate–Craniofacial Journal*, 36(3), 233–242. https://doi.org/10.1597/1545-1569_1999_036_0233_pompsf_2.3.co_2

Scherer, N. J., D'Antonio, L. L., & McGahey, H. (2008). Early intervention for speech impairment in children with cleft palate. *The Cleft Palate–Craniofacial Journal*, 45(1), 18–31. https://doi.org/10.1597/06-085.1

Sell, D. (2005). Issues in perceptual speech analysis in cleft palate and related disorders: A review. *International Journal of Language & Communication Disorders*, 40(2), 103–121. https://doi.org/10.1080/13682820400016522

Sell, D., 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 & Communication Disorders*, 44(4), 529–548. https://doi.org/10.1080/13682820802196815

Sell, D., & Sweeney, T. (2020). Percent consonant correct as an outcome measure for cleft speech in an intervention study. *Folia Phoniatrica et Logopaedica*, 72(2), 143–151. https://doi.org/10.1159/000501095

Shriberg, L. D. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research*, 36(1), 105–140. https://doi.org/10.1044/jshr.3601.105

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric. *Journal of Speech, Language, and Hearing Research*, 40(4), 708–722. https://doi.org/10.1044/jslhr.4004.708

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. https://doi.org/10.1044/jshd.4703.256

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, 5(3), 225–279. https://doi.org/10.3109/02699209108986113

Sitzman, T. J., Allori, A. C., & Thorburn, G. (2014). Measuring outcomes in cleft lip and palate treatment. *Clinics in Plastic Surgery*, 41(2), 311–319. https://doi.org/10.1016/j.cps.2013.12.001

Sweeney, T., Hegarty, F., Powell, K., Deasy, L., O'Regan, M., & Sell, D. (2020). Randomized controlled trial comparing Parent Led Therapist Supervised Articulation Therapy (PLAT) with routine intervention for children with speech disorders associated with cleft palate. *International Journal of Language & Communication Disorders*, 55(5), 639–660. https://doi.org/10.1111/1460-6984.12542

Tönz, M., Schmid, I., Graf, M., Mischler-Heeb, R., Weissen, J., & Kaiser, G. (2002). Blinded speech evaluation following pharyngeal flap surgery by speech pathologists and lay people in children with cleft palate. *Folia Phoniatrica et Logopaedica*, 54(6), 288–295. https://doi.org/10.1159/000066153

Umeda, N. (1978). Occurrence of glottal stops in fluent speech. *The Journal of the Acoustical Society of America,* 64(1), 88–94. https://doi.org/10.1121/1.381959

Vallino, L. D., Lass, N. J., Bunnell, H. T., & Pannbacker, M. (2008). Academic and clinical training in cleft palate for speech-language pathologists. *The Cleft Palate–Craniofacial Journal*, 45(4), 371–380. https://doi.org/10.1597/07-119.1

World Health Organization. (2002). *Global strategies to reduce the health care burden of craniofacial anomalies: Report of WHO meetings on international collaborative research on craniofacial anomalies.*

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Willadsen, E., Lohmander, A., Persson, C., Lundeborg, I., Alaluusua, S., Aukner, R., Bau, A., Boers, M., Bowden, M., Davies, J., Emborg, B., Havstam, C., Hayden, C., Henningsson, G., Holmefjord, A., Hölttä, E., Kisling-Møller, M., Kjøll, L., Lundberg, M., . . . Semb, G. (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. https://doi.org/10.1080/2000656x.2016.1254647

Zajac, D. J. (2015). The nature of nasal fricatives: Articulatory-perceptual characteristics and etiologic considerations. SIG 5 *Perspectives on Speech Science and Orofacial Disorders,* 25(1), 17–28. https://doi.org/10.1044/ssod25.1.17

Appendix A (p. 1 of 2)

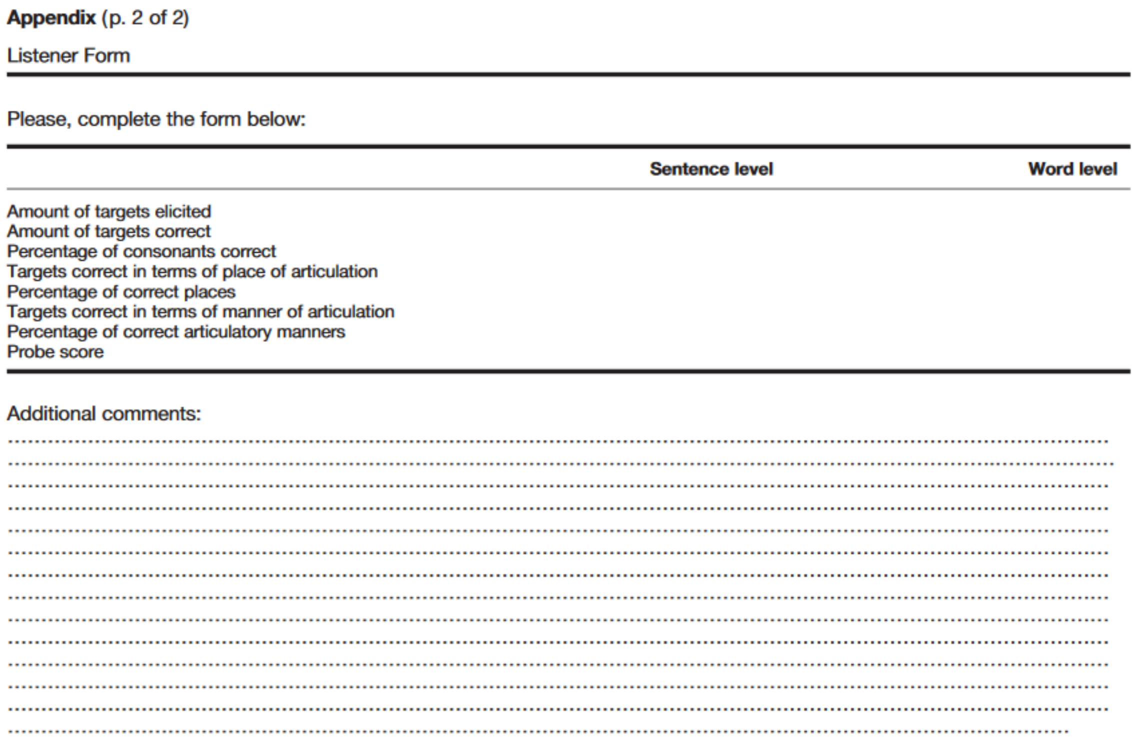
Listener Form

**Listening form**
**Sample number:**
**Date:**
**Rater:**

| Target consonant | Sentence level (Dutch) (Bruneel et al., 2020) | Phonetic transcription (target)[a] | Sentence level (English translation) | Phonetic transcription (production child) | Error classification |
|---|---|---|---|---|---|
| [p] | Papa riep opa. | [pɑpɑ ri'p o:pɑ] | Daddy called grandpa. | | |
| [t] | Tuur eet later. | [ty'r e:t la:tər] | Tuur eats later. | | |
| [b] | Bel boer Robbe. | [bɛl bu'r rɔbə] | Call farmer Robbe. | | |
| [d] | Lode doet de deur toe. | [lo:də du't də dø:r tu'] | Lode closes the door. | | |
| [f] | De toffe fee is lief. | [də tɔfə fe: ɪs li'f] | The nice fairy is sweet. | | |
| [v] | Eva viel voorover. | [e:vɑ vil vo:ro:vər] | Eva tripped. | | |
| [k] | De cake rook lekker. | [də ke:k ro:k lɛkər] | The cake smelled good. | | |
| [s] | Sara wil de losse jas. | [sa:rɑ wɪl də lɔsə jɑs] | Sara wants the loose coat. | | |
| [z] | Ze ziet de roze zoo. | [zə zit də ro:zə zo:] | She sees the pink zoo. | | |
| [x] | Ik goochel graag. | [ɪk ɣo:xəl ɣra:x] | I like magic. | | |
| [ɣ] | De egel gilt. | [də e:ɣəl ɣɪlt] | The hedgehoc screams. | | |
| [ʃ] | Liesje showt de sjaal. | [liʃə ʃo:wt də ʃa:l] | Liesje shows the scarf. | | |
| [ʒ] | De logé wou gelei. | [də lo:ʒe: wɑu ʒəlɛi ] | The guest wants jelly. | | |
| Sentences with s-clusters | Stella speelt liever de heks. | [stlɑ spe:lt li'vər də hɛks] | Stelle prefers to play the witch. | | |

| | Word level | Phonetic transcription (target)[a] | Sentence level (English translation) | Phonetic transcription (production child) | Error classification |
|---|---|---|---|---|---|
| [p] | Poes | [pus] | Cat | | |
| | Opa | [o:pa] | Grandpa | | |
| | Wip | [wɪp] | Seesaw | | |
| [t] | Toe | [tu] | Closed | | |
| | Water | [wa:tər] | Water | | |
| | Hoed | [hut] | Hat | | |
| [k] | Koe | [ku] | Cow | | |
| | Wekker | [wɛkər] | Alarm clock | | |
| | Koek | [kuk] | Cookie | | |
| [b] | Boer | [bur] | Farmer | | |
| | Baby | [be:bi:] | Baby | | |
| [d] | Dier | [di:r] | Animal | | |
| | Rode | [ro:də] | Red | | |
| | Ladder | [ladər] | Ladder | | |
| [f] | Fee | [fe:] | Fairy | | |
| | Wafel | [wa•fel] | Waffle | | |
| | Woef | [wuf] | Barking | | |
| [v] | Over | [o:vər] | Over | | |
| | Vuur | [vy:r] | Fire | | |
| [s] | Saus | [sɑus] | Sauce | | |
| | Lasso | [laso:] | Lariat | | |
| | Jas | [jas] | Coat | | |
| [z] | Zuur | [zy:r] | Sour | | |
| | Ezel | [e:zəl] | Donkey | | |
| [x] | Lach | [läx] | Smile | | |
| | Goochelaar | [ɣo:xəla:r] | Magician | | |
| [ɣ] | Geeuw | [ɣeuw] | Yawn | | |
| | Egel | [e:ɣəl] | Hedgehoc | | |
| [ʃ] | Sjaal | [ʃa:l] | Scarf | | |
| | Roosje | [ro:ʃə] | Rose | | |
| [ʒ] | Logé | lo:ʒe: | Guest | | |
| | Gelei | ʒəlɛi | Jelly | | |

[a] This column refers to the phonetic transcription of the correct target production.



**Appendix** (p. 2 of 2)

Listener Form

Please, complete the form below:

| | Sentence level | Word level |
|---|---|---|
| Amount of targets elicited | | |
| Amount of targets correct | | |
| Percentage of consonants correct | | |
| Targets correct in terms of place of articulation | | |
| Percentage of correct places | | |
| Targets correct in terms of manner of articulation | | |
| Percentage of correct articulatory manners | | |
| Probe score | | |

Additional comments:

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

..............................................................................................................................................

.........................................................................................................................................