Published in final edited form as:
Cleft Palate Craniofac J. 2016 January ; 53(1): 93–108. doi:10.1597/14-027.

The Americleft Speech Project: A Training and Reliability Study

Dr. Kathy L. Chapman, Ph.D.,
Professor, Department of Communication Sciences and Disorders, University of Utah, Salt Lake City, Utah

Dr. Adriane Baylis, Ph.D.,
Speech Scientist and Director of the Velopharyngeal Dysfunction Program, Nationwide Children's Hospital, and Assistant Professor-Clinical, Department of Plastic Surgery, The Ohio State University College of Medicine, Columbus, Ohio

Dr. Judith Trost-Cardamone, Ph.D.,
Professor Emeritus, Department of Communication Disorders and Sciences, California State University at Northridge, Northridge, California, and Speech Consultant, Ventura County Medical Center Cleft Lip and Palate Clinic, Ventura, California

Dr. Kelly Nett Cordero, Ph.D.,
Rehabilitation Therapies Supervisor of Clinical Practice and Outcomes and Craniofacial Speech-Language Pathologist for the Center for Craniofacial Services, Gillette Children's Specialty Healthcare, Saint Paul, Minnesota

Ms. Angela Dixon, M.A.,
Speech-Language Pathologist, Department of Audiology and Speech Pathology, Riley Hospital for Children at Indiana University Health, Indianapolis, Indiana

Ms. Cindy Dobbelsteyn, M.Sc.,
Speech-Language Pathologist, School of Human Communication Disorders, Dalhousie University, Halifax, Nova Scotia, Canada

Ms. Anna Thurmes, M.A.,
Speech-Language Pathologist/Program Coordinator, School of Dentistry, Department of Developmental and Surgical Sciences, University of Minnesota, Minneapolis, Minnesota

Dr. Kristina Wilson, Ph.D.,
Senior Speech-Language Pathologist and Clinical Researcher, Division of Speech, Language and Learning at Texas Children's Hospital, and Adjunct Assistant Professor, Department of Plastic Surgery at Baylor College of Medicine, Houston, Texas

Dr. Anne Harding-Bell, Ph.D.,
Module Coordinator, Post Graduate Cleft Palate Studies Programme, Department of Human Communication Sciences, University of Sheffield, Sheffield, United Kingdom

Address correspondence to: Dr. Kathy Chapman, Department of Communication Sciences and Disorders, University of Utah, 390 S 1530 E, Salt Lake City, UT 84112. kathy.chapman@health.utah.edu.

This paper was presented at the 69th Meeting of the American Cleft Palate–Craniofacial Association in San Jose, California, on April 18, 2012.

Dr. Triona Sweeney, Ph.D.,
Cleft Research Therapist, Speech & Language Therapy Department, Temple Street Children's University Hospital, Dublin, Ireland and Adjunct Professor, Speech & Language Therapy, Clinical Therapies Unit, University of Limerick, Limerick, Ireland

Dr. Gregory Stoddard, Ph.D., and
Co-Director, Study Design and Biostatistics Center, Department of Internal Medicine, University of Utah, Salt Lake City, Utah

Dr. Debbie Sell, Ph.D.
Principal Speech and Language Therapist, Speech and Language Therapy Department, and Senior Research Fellow, Centre for Outcomes and Experience Research in Children's Health, Illness and Disability (ORCHID), Great Ormond Street Hospital National Health Service Trust, London, United Kingdom

Abstract

Objective—To describe the results of two reliability studies and to assess the effect of training on interrater reliability scores.

Design—The first study (1) examined interrater and intrarater reliability scores (weighted and unweighted kappas) and (2) compared interrater reliability scores before and after training on the use of the Cleft Audit Protocol for Speech–Augmented (CAPS-A) with British English-speaking children. The second study examined interrater and intrarater reliability on a modified version of the CAPS-A (CAPS-A Americleft Modification) with American and Canadian English-speaking children. Finally, comparisons were made between the interrater and intrarater reliability scores obtained for Study 1 and Study 2.

Participants—The participants were speech-language pathologists from the Americleft Speech Project.

Results—In Study 1, interrater reliability scores improved for 6 of the 13 parameters following training on the CAPS-A protocol. Comparison of the reliability results for the two studies indicated lower scores for Study 2 compared with Study 1. However, this appeared to be an artifact of the kappa statistic that occurred due to insufficient variability in the reliability samples for Study 2. When percent agreement scores were also calculated, the ratings appeared similar across Study 1 and Study 2.

Conclusion—The findings of this study suggested that improvements in interrater reliability could be obtained following a program of systematic training. However, improvements were not uniform across all parameters. Acceptable levels of reliability were achieved for those parameters most important for evaluation of velopharyngeal function.

Keywords

Americleft; cleft palate; listener agreement; listener reliability; perceptual ratings; speech outcome; training

One of the primary outcome measures of cleft lip and palate management is speech. This measure generally includes perceptual judgments of articulation (including identification of compensatory articulations [CAs]) and resonance, along with other parameters associated with velopharyngeal function (e.g., identification of audible nasal emission, weak pressure consonants). Although clinicians and researchers have attempted to identify more objective measurement procedures (e.g., acoustic, aerodynamic), the "ear of the listener" is considered to be the "gold standard" for speech evaluation (e.g., McWilliams et al., 1990; Gerratt et al., 1993; Kreiman et al., 1993; Kent, 1996; Kuehn and Moller, 2000; Oates, 2009; Sweeney, 2011). At the same time, perceptual judgments "are susceptible to sources of error and bias" that may have serious consequences for the repeatability/reproducibility of these judgments (Kent, 1996, p. 7). Thus, researchers and clinicians using perceptual judgments for evaluation of speech are obligated to show that their findings can be replicated, which includes calculating and reporting agreement or reliability scores (Cordes, 1994). Although *reliability* and *agreement* are used interchangeably in the literature, they are actually different (de Vet et al., 2006; Kottner et al., 2011). Reliability measures "assess whether study objects, often persons, can be distinguished from each other, despite measurement errors" (de Vet, 2006, p. 1033). Agreement measures reflect the "degree to which scores or ratings are identical" (Kottner et al., 2011, p. 96). Reliability is more reflective of how the measure performs with a specific population; whereas, agreement relates more to the actual properties of the measure (de Vet, 2006). Agreement statistics such as percent agreement scores do not account for chance agreement, but reliability measures such as a kappa or intraclass correlation coefficient (ICC) remove chance agreement in their calculations (Cohen, 1960; Hallgren, 2012). Both intrarater (the same judge rates the same material on two separate occasions) and interrater (different judges independently rate the same material) reliability/agreement should be reported. High intrarater reliability demonstrates the ratings are reproducible under similar conditions; high interrater reliability demonstrates the judges are applying a common criterion in forming their ratings.

Numerous research studies and tutorials have addressed issues related to reliability for various types of perceptual data. For example, as noted by Cordes (1994), "Problems with the reliability of observational data are also among some of the most serious methodological concerns in some sub-areas of the discipline …" (p. 265). Almost 20 years later, the challenges associated with the use of perceptual data persist in voice (judgments of voice quality) (e.g., see Chan and Yiu, 2002; Shrivastav et al., 2005; Eadie and Baylor, 2006; Oates, 2009; Kreiman and Gerratt, 2010); articulation/phonology (phonetic transcription of speech data) (see Shriberg and Lof, 1991; Cucchiarini, 1996; Ball and Rahilly, 2002); stuttering (identification of stuttering events) (see Young, 1984; Cordes and Ingham, 1994; Lewis, 1994; Cordes, 2000); and cleft palate speech (judgments of resonance and other speech behaviors associated with clefting) (see Peterson-Falzone, 1996; Keuning et al., 1999; Gooch et al., 2001; Sell, 2005; Brunnegård and Lohmander, 2007; Sell et al., 2009).

Specific to cleft palate speech, Lohmander and Olsson (2004) performed a review of three international journals in an attempt to summarize the "state of the art" for studies examining perceptual assessment of cleft palate speech from 1980 to 2000. Of the 88 articles identified, reliability information was reported in only 51% of the articles: interrater and intrarater reliability was reported in 26%, interrater reliability in only 38%, and intrarater reliability in only 11%. These data suggest that reliability judgments were not performed for almost one half of studies. For those studies that have reported reliability for cleft-related speech characteristics such as hypernasality and audible nasal emission, there is considerable variance in the scores reported across studies. Moreover, it is sometimes difficult to compare across studies due to differences in how reliability was calculated (e.g., percent agreement, correlations, kappas, and single- or average-measure ICCs). However, examination of a representative sample of speech outcome studies reviewed by Lohmander (2011) suggested that interrater reliability for hypernasality ratings range from very poor (kappas of .41 and percent agreement scores of 39%) to excellent (kappas of 1.0 and percent agreement scores of 100%) across studies. There was less variability for ratings of audible nasal emission (percent agreement scores ranged from 69% to 100%); however, there were also fewer studies that reported reliability for this measure.

Perceptual judgments required for characterizing the speech sound productions of children with cleft palate are also challenging. In the only study that directly addressed transcription reliability of children with cleft palate, the mean interrater percent agreement scores for experienced speech-language pathologists (SLPs) were 39%, 69%, and 66% for CAs, non-CA errors, and correctly articulated sounds, respectively. Intrarater agreement, although better with a group mean of 76% for CAs, was still considered to be low (Gooch et al., 2001). Other reports of transcription reliability from speech outcome studies have also shown moderate levels of agreement (kappa = .44) for glottal articulations (Sell et al., 2001). However, reliability for transcription of CAs appeared to improve when these productions were grouped into categories such as nonoral cleft speech characteristics (CSCs) and scores were calculated for the entire category. Following this procedure, Sell et al. (2009) reported kappas in the .60-plus range, which is considered good agreement (Altman, 1991). Generally, interrater agreement for transcription of speech sounds improved when calculated for all sounds produced and not just for CAs, given that Lohmander et al. (2012) reported percent agreement scores of 86% to 100% in their study of speech outcomes. These scores were consistent with what has been reported in studies describing speech characteristics of children with cleft palate, that is, interrater agreement scores of 80% to 90% across studies (for a review, see Klinto et al., 2011).

A number of factors have been identified that make perceptual speech judgments difficult and subsequently impact reliability of these judgments (e.g., see Cordes, 1994; Kent, 1996; Eadie and Baylor, 2006, for comprehensive reviews). These include, but are not limited to (1) the speech analysis protocol used, (2) the procedures for data collection and analysis, and (3) listener characteristics (level of experience and training) (Cordes, 1994; Kent, 1996; Sell, 2005; Oates, 2009; Sell et al., 2009). Each of these three areas and their impact on reliability is addressed below.

Speech Analysis Protocol

As noted by Kent (1996), one of the basic issues related to perceptual analysis of speech disorders is lack of agreement about which parameters should be rated for a given disorder area. In cleft palate speech assessment, there has been much debate and controversy about what should be assessed and how it should be assessed (Lohmander and Olsson, 2004; Sell, 2005; Fujiwara et al., 2006). As a result, a variety of different protocols have been used to evaluate speech outcomes for speakers with repaired cleft palate (McWilliams and Philips, 1979; Dalston et al., 1988; Eurocleft Speech Group, 1994; Sell et al., 1994; Keuning et al., 1999; Sell et al., 1999; Eurocleft Speech Group, 2000; Dotevall et al., 2002; John et al., 2006; Henningsson et al., 2008). Whereas the content of these protocols is somewhat similar in terms of parameters evaluated and the types of judgments and scale values used, Henningsson et al. (2008) pointed out that nothing has been universally adopted or used widely enough to permit large-scale cross-center studies. Three recent protocols that have attracted international interest are the Cleft Audit Protocol for Speech–Augmented (CAPS-A) developed in the U.K. (John et al., 2006); the Swedish Articulation and Nasality Test (SVANTE) developed as part of the Scancleft project (Lohmander et al., 2005); and the Universal Parameters for Reporting Speech Outcomes (UPS) developed collaboratively by an international working group (Henningsson et al., 2008). The SVANTE is not accessible to non–Swedish-speaking audiences, and to our knowledge, the UPS has not been validated. The CAPS-A, however, has been shown to have validity, reliability, and applicability in several studies conducted in the U.K. (John et al., 2006; Sell et al., 2009; Britton et al., 2014). This protocol initially was developed as a quality assurance tool for evaluating 5-year-old children with a history of cleft palate, although it has been used more widely (Pereira et al., 2013). To date, no studies have been conducted comparing reliability of different speech assessment protocols. What is known, however, is that clinicians in the U.K. who received training on the use of the CAPS-A were able to reach acceptable levels of reliability on a variety of parameters related to speech outcome for children with cleft palate.

In addition to the speech assessment protocol to be used, the type of rating scale used might influence reliability. Traditionally, most protocols, especially those designed for clinical purposes, have used equal-appearing interval scaling (EAI) for rating of speech parameters such as hypernasality, hyponasality, and audible nasal emission, among others. However, recent studies have suggested that higher levels of reliability may be achieved by rating hypernasality using direct magnitude estimation (DME) (Zraick and Liss, 2000; Whitehill et al., 2002) or visual analog scaling (VAS) (Baylis et al., 2015). Baylis and colleagues also found higher levels of reliability for judgments of audible nasal emission using DME (Baylis et al., 2011) and VAS (Baylis et al., 2015) compared with EAI scaling.

Procedures for Data Collection and Analysis

Several authors have highlighted the need for standardized data collection, recording, and playback as part of the principles of perceptual speech assessment for any speech disorder (Grunwell et al., 1993; Kuehn and Moller, 2000; Gooch et al., 2001; Lohmander and Olsson, 2004; Sell, 2005; Sell et al., 2009). Yet, in the cleft palate literature, there is little standardization across studies for many of these variables. Specifically, variability is present for the sampling context, recording method, and the listening/ rating environment (see Lohmander and Olsson, 2004; Sell, 2005; Fujiwara et al., 2006; Sell et al., 2009). This is concerning because we know that for hypernasality, judgments are influenced by variables such as length of the sample (Spriestersbach and Powers, 1959; Daniel, 1971), intelligibility (McWilliams, 1954) and articulation competency of the speaker (Van Hattum, 1958), vowel context (Van Hattum, 1958; Spriestersbach and Powers, 1959), and voice quality (Imatomi, 2005). Furthermore, research has shown that when rating hypernasality, better reliability is noted (1) with low back versus high front vowels within a sentence (Watterson et al., 2007) and (2) for longer samples (sentence or connected speech) compared with shorter stimuli (vowels or syllables) (Spriestersbach and Powers, 1959; Counihan and Cullinan, 1970; Daniel, 1971).

Perceptual judgments required for phonetic transcription of children's speech are also influenced by external factors, some similar and some different from those described above for hypernasality (see McNutt et al., 1991; Shriberg and Lof, 1991; Kent, 1996; Lockhart and McLeod, 2013, for a review). Shriberg and Lof (1991) examined data from a group of trained transcribers to examine variables impacting reliability of phonetic transcription for children with speech sound disorders. They found that transcription reliability was influenced by (1) severity of the speech problem (as severity increased, transcription agreement decreased), (2) sampling context (agreement was better when transcribing conversational speech compared with single words), (3) word position (agreement was higher in word initial position and lower in word final position), and (4) level of transcription specificity (low levels of agreement with narrow transcription), among other things.

A recent study by Klinto et al. (2011) of children with cleft palate indicated that not only did children's error rates vary across speech contexts (single words, sentence imitation, retelling a narrative, and conversational speech), but interrater reliability was also context dependent (i.e., lowest for retelling a narrative, but good [80% to 90%] for the other three contexts).

Finally, several studies have focused on recording and playback conditions and how they influence reliability. There appears to be an advantage in terms of higher reliability scores for live judgments over audio recordings (Stevens and Daniloff, 1977) and video recordings over audio recordings (McNutt et al., 1991); although, this may differ across phonemes (see McNutt et al., 1991, for a summary). In contrast, reliability is not impacted by whether the listener uses headphones (high-quality or "regular" earphones) or speakers in the sound field (McNutt et al., 1991; Yeung, 2010) or whether the samples are recorded and played back using digital or analog recording equipment (Shriberg et al., 2005). Although it appears that differences in data collection and listening environments may not impact perceptual ratings equally, comparisons across studies are facilitated if procedures are standardized not only within but also between studies.

Listener Characteristics

One of the most important influences on reliability is the listener (Cordes, 1994). According to Kreiman et al. (1993), listeners may have different internal standards against which perceptual ratings are made, and these internal standards change across time, thus impacting both interrater and intrarater reliability. Internal standards may be influenced by a number of variables including the listener's experience rating the parameter(s) of interest and the training that is carried out.

Listener Experience

The effect of experience might vary depending on the nature of the perceptual task. For example, studies of listener experience suggested that differences exist between experienced and inexperienced listeners for voice quality ratings (Kreiman et al., 1990; Kreiman et al., 1992; Helou et al., 2010) and phonetic distinctions (Gooch et al., 2001; Wolfe et al., 2003; Munson et al., 2012) but not always for perceptual rating of dysarthric speech (Bunton et al., 2007). Type of experience might also be important. However, studies comparing college students with varying levels of experience have not always found an advantage for experienced listeners (Santelmann et al., 1999; Schellinger et al., 2008).

When judging the speech of speakers with cleft palate, there are variable findings across studies examining the impact of listener experience. Some studies found higher levels of reliability for experienced listeners compared with inexperienced listeners (Keuning et al., 1999; Gooch et al., 2001; Lewis et al., 2003), but others did not (e.g., Bradford et al., 1964; Tonz et al., 2002; Brünnegård et al., 2009). Some of the lack of agreement might be related to differing definitions for *experienced* and *inexperienced* as well as the nature of what was being rated. Clearly, this is an area that requires additional study.

Listener Training

Listener training has been used to improve validity and reliability of perceptual judgments in many areas of communication disorders including hearing impairment (e.g., Ellis and Beltyukova, 2008), stuttering (e.g., Cordes and Ingham, 1999), voice disorders (e.g., Bassich and Ludlow, 1986; Chan and Yiu, 2002, 2006; Eadie and Baylor, 2006), and cleft palate–related speech disorders (e.g., Lee et al., 2009). A majority of the studies indicated that training improved listeners' perceptual ratings (Chan and Yiu, 2002, 2006; Eadie and Baylor, 2006), and a combination of training with the use of anchors (external standards to serve as a reference point) has been shown to significantly improve reliability scores for perceptual judgments of several voice-quality parameters (Chan and Yiu, 2002; Eadie and Baylor, 2006). However, Ellis and Beltyukova (2008) studied the effect of familiarization training versus repeated exposure on naïve listeners' ability to judge speech intelligibility of children with profound hearing loss, using a word identification task in a storytelling context. Familiarization training consisted of listening to the samples along with printed transcripts of what the participants were saying. They found that repeated exposure, with or without familiarization training, improved word identification scores.

In the cleft palate literature, a number of studies have advocated the use of training to improve the reliability of perceptual judgments (e.g., McWilliams and Philips, 1990; Gooch et al., 2001; Lohmander et al., 2009; Sell et al., 2009), but only a few studies have attempted to show this experimentally. Lee et al. (2009) compared hypernasality judgments across three training conditions. The findings indicated that both interrater and intrarater reliability of hypernasality improved in the two practice conditions but not in the exposure-only condition, despite listeners having only limited experience with hypernasal speech. John et al. (2006) showed that multiple judges could obtain reliable results after limited but focused training. Sell et al. (2009) developed a training package to accompany the CAPS-A that specifies principles of speech sample selection, data acquisition, recording, playback, and listening guidelines for analysis. They subsequently reported on the effectiveness of this package in training a group of experienced and inexperienced SLPs (Sell et al., 2010).

This review of the literature suggests that although perceptual judgments of cleft palate speech have limitations and many sources of variability, steps can be taken to ensure that acceptable levels of reliability are achieved. Additionally, it appears that training on the assessment tools is desirable, irrespective of degree of experience. The purpose of this study is to determine whether training improves interrater reliability of listener ratings on the CAPS-A protocol in a cross-cultural group of English-speaking SLPs. The training and reliability results reported here will serve as the foundation for the analysis of the comparative speech outcome datasets for the Americleft Speech Group. The Americleft Speech Group is a subgroup of the Americleft Project, which grew out of the American Cleft Palate–Craniofacial Association Task Force on Inter-center Collaboration. This project was established with the goal of collaborative data collection and analysis for the reporting of intercenter outcomes related to the treatment of cleft lip and palate (Long et al., 2011).

Method

Overview

Two separate reliability studies were conducted, approximately 6 months apart, for this project. In Study 1, SLPs underwent training in the use of the CAPS-A using British- and Irish-English speech edits from different children from the caseloads of the three trainers. Edits were selected if they were particularly good examples of the different parameters and different scalar points, covering the range of severity levels with respect to resonance, audible nasal emission/nasal turbulence, and articulatory characteristics observed in children with repaired cleft palate. Following training, SLPs were asked to rate the speech samples of 10 novel speech edits of children from the U.K. and Ireland (referred to henceforth as the UK/I reliability speech samples) on three separate occasions: (1) prior to receiving training (pretraining), (2) immediately following training (posttraining 1), and (3) approximately 1 month following training (post-training 2) to determine the effect of training and to obtain intrarater and interrater reliability scores. In Study 2, modifications were made to the CAPS-A, and SLPs rated North American speech samples (referred to hereafter as NA reliability speech samples). The SLPs rated the NA reliability speech samples on two separate occasions approximately 4 and 5 months later to determine interrater and intrarater reliability scores, respectively. Results of Study 1 and Study 2 were then compared.

Study 1: UK/I Samples

SLP Raters—Nine certified SLPs from North America (NA) (the United States and Canada) who constituted the Americleft Speech Group participated in the training. All were affiliated with a cleft/craniofacial team, with the exception of one SLP who had previous cleft team experience and maintained her cleft speech assessment skills for research and teaching purposes as a professor (K.C.). All SLPs had a minimum of 6 years of experience with perceptual speech assessment of children with cleft palate, with the exception of one SLP who had just completed 1 year of mentored experience with one of the participating SLPs on a cleft palate team. None of the SLPs had a known hearing loss.

Material for the Study—The 10 novel UK/I reliability speech samples were from nine children with repaired cleft palate (with or without cleft lip) and one child who presented with noncleft velopharyngeal dysfunction. None of the children had diagnosed syndromes. The group ranged in age from 5 to 7 years of age. The samples were of good technical quality and for the purpose of testing reliability, represented a range of severity levels for speech characteristics often observed in children with repaired cleft palate. The sampling contexts included (1) a short spontaneous speech sample, (2) counting from 1 to 20 and 60 to 70, (3) a nursery rhyme, and (4) repetition of the Great Ormond Street Speech Assessment 98 (GOS.SP.ASS 98) sentences (Sell et al., 1999; John et al., 2006).

CAPS-A Training—The CAPS-A trainers led a 3-day training workshop on the CAPS-A materials and methodology for the SLP participants. The trainers included three of the CAPS-A authors (A.H., D.S., and T.S.), all of whom had extensive experience in the perceptual speech assessment of children with cleft palate. Each parameter of the CAPS-A was described in detail, and definitions for all terms were reviewed and discussed. The CAPS-A core parameter categories and subcategories, as well as scalar values for each rating (as described in Table 1) were presented and illustrated with video edits.

Training also reviewed the principles of audio and video recording, guidelines for playback and listening, and the structured listening and rating protocol used with the CAPS-A. Rating and transcription practice were integrated into the training and included significant time for consensus listening, discussion, and rating of samples. Time also was devoted to discussion of differences that would be anticipated when rating North American English samples due to differences in dialect/ accent as well as differences in SLP terminology for various rating parameters (e.g., “distinctiveness” versus “acceptability”) and articulation error types. The structured listening protocol shown in Table 2 (modified from Sell et al., 2009) illustrates the standard order in which the speech sample components were presented. It also specifies the listening medium (audio only or digital video with audio) for each of the parameters. For example, the intelligibility/distinctiveness rating was based on a single listening to the audio-only recording of the spontaneous speech sample, accompanied by a still image of the child’s face. Ratings for all other parameters (except grimace) were based on the visual (with audio) segments of counting, nursery rhyme recitation, and sentence repetition tasks. A rater could listen to these segments as many times as necessary.

The trainers devoted approximately 2 days to reviewing the administration and scoring procedures for the CAPS-A. Practice using the tool was facilitated by consensus ratings and phonetic transcription using a variety of video- and audio-recorded UK/I edits. These were presented to the group of SLP listeners with an external speaker placed relatively equidistant from all SLPs. During the training, all SLPs had copies of (1) the Appendix of the CAPS-A (John et al., 2006), which contained all parameter definitions, (2) a chart that listed and described the CSCs and included International Phonetic Alphabet/Extended International Phonetic Alphabet symbols and schematic illustrations depicting articulatory placement of specific error types, and (3) a copy of the GOS.SP.ASS sentences used in the samples to assist with speech ratings and transcription. The reader is referred to Sell et al. (2009) for a full description of the published CAPS-A training package for additional details.

During the consensus practice sessions, the SLPs could request that the sample be replayed as many times as needed to arrive at a rating for a given parameter. When there was disagreement, the ratings were discussed and replayed until a general consensus (typically ±1 scale point) was achieved. This pattern of consensus listening, which lasted on average about 30 to 45 minutes per sample, was repeated for all training samples.

Rating of Samples: Pretraining—Prior to the CAPS-A training session, the nine participating SLPs were asked to review the CAPS-A procedures (John et al., 2006; Sell et al., 2009) and to independently rate the 10 UK/I reliability speech samples to examine the effect of training on reliability ratings. These samples were provided on a DVD by the CAPS-A trainers prior to the training session. Formal listening instructions beyond those described in the John et al. (2006) article were not provided to the SLPs for this set of ratings. The CAPS-A rating forms were provided to each SLP to record ratings for each sample. All SLP ratings were performed independently while listening to the samples through external speakers or headphones.

Rating of Samples: Posttraining (Immediate and 1 Month Posttraining)—Immediately following the CAPS-A training, the nine SLPs independently rerated the 10 UK/I reliability speech samples. Seven samples were presented through an external speaker live to all SLPs. Due to time and travel constraints, the last three samples were rated independently by the SLPs through headphones or external speakers within 1 week of the final day of the training session. All 10 samples were rated again 1 month later to obtain posttraining reliability data. All ratings were conducted independently on a computer using earphones or high-quality speakers.

Study 2: NA Samples

Prior to undertaking rating of the NA reliability speech samples, minor modifications were made to the CAPS-A rating protocol and scoring form, both on conceptual grounds and to improve the ease of use with the NA speech samples. There were three primary modifications. First, the sentence stimuli for the NA training samples were replaced by the American English Sentence Sample (AESS) (Trost-Cardamone, 2012) (Appendix A). The AESS is comparable to the GOS.SP.ASS sentences used with the CAPS-A in that both target specific consonants, that is, sentences “loaded” with a given target sound. They are also similar in that they target pressure consonants and both include at least one sentence with low-pressure consonants only. The AESS sentences are also different from the GOS.SP.ASS sentences in that (1) the sentences targeting pressure consonants include no nasal consonants, (2) the sentence containing nasals includes no pressure consonants, and (3) any nontarget sound in a target sentence is an approximant/glide or liquid or is as close as possible to the same place of articulation as the target (Hutters and Henningsson, 2004; Henningsson et al., 2008; Lohmander et al., 2009). Second, minor changes were made to the core parameters of the CAPS-A and to their corresponding working definitions (Table 3). Third, the NA training sample ratings were based entirely on video recordings, not audio-only samples.

SLP Raters—Six of the nine original Americleft SLP listeners who participated in the CAPS-A training and Study 1 served as raters for Study 2. All were affiliated with a cleft/craniofacial team and had a minimum of 8 years of experience with cleft palate speech assessment. All six raters underwent pure-tone audiometry, and hearing was found to be within normal limits, bilaterally.

The NA Reliability Speech Samples—Ten video-recorded speech samples of English-speaking children from NA were selected from available pilot cases collected for the Americleft Speech Project, under respective institutional review board approval and with informed consent. Children in these samples were 5 to 10 years of age with repaired cleft palate (with or without cleft lip). One of the children had been diagnosed with velocardiofacial syndrome; the other nine had no diagnosed syndrome. Samples were recorded using a video camera with an external microphone, following the recommended recording guidelines provided in the CAPS-A training session (e.g., microphone placement, lighting). The samples were chosen from a pool of 26 samples, representing a range of severity levels and speech parameters of interest. The NA reliability speech samples were posted on a secure university-based website for the SLP raters to access remotely. The SLP raters were sent an electronic invitation and secure login information in order to access the samples from the website during the rating period.

Rating of Samples—The six SLP raters independently watched and analyzed the 10 NA reliability speech samples from their individual computers using high-quality headphones. Ratings were made using the Americleft modification of the CAPS-A rating form (CAPS-A-AM). Approximately 4 weeks later, each SLP rerated all 10 samples for the purpose of determining intrarater reliability, following the same aforementioned procedures.

Data Analysis

Interrater and intrarater reliability were computed using a weighted or unweighted kappa statistic (Stata, version 12; StataCorp., College Station, TX). The unweighted kappa was calculated for the nominal parameters: voice, noncleft speech immaturities/errors, perceived need for speech/language therapy, other speech errors, and recommended further investigation of velopharyngeal function (VPF). This is a type of “exact” agreement kappa in that the scores must be exactly the same to be considered in agreement. Weighted kappas were calculated for all other ordinal speech variables. The weights were assigned as $w_{ij} = 1 - (i - j)^2/(k - 1)^2$, so scores with smaller differences, or more closeness, were given more weight than scores with larger differences (Streiner and Norman, 1995; Fleiss et al., 2003). For example, for a scale with four possible scores, the weights are 1 if exactly the same, .89 if different by 1 point, .56 if different by 2 points, and 0 if different by 3 points. Strength of agreement descriptions were assigned using the labels suggested by Altman (1991), recognizing that these labels, like any other proposed labels, have some inherent arbitrariness.

To determine whether improvement in interrater reliability scores occurred following training, comparisons were made between the reliability scores obtained pretraining and posttraining 1 for the UK/I samples. To assess maintenance, comparisons were also made between posttraining 1 and posttraining 2 (UK/I reliability speech samples) and rating 1 and rating 2 (NA reliability speech samples). Intrarater reliability was also calculated between posttraining 1 and post-training 2 for the UK/I samples and between rating 1 and rating 2 for the NA samples. Finally, comparisons were made between the interrater and intrarater scores from the U.K. and the U.S. data sets.

Results

Study 1: UK/I Samples

Interrater Reliability: Pretraining—Interrater reliability was calculated using weighted and unweighted kappa statistics (Stata, version 12; StataCorp.) for the nine SLP raters on the CAPS-A. *Good agreement* (.61 to .80) was noted for 7 of the 13 parameters: intelligibility, voice, hypernasality, grimace, posterior oral CSCs, nonoral CSCs, and passive CSCs. *Moderate agreement* (.41 to .60) was noted for 3 of the 13 parameters: audible nasal emission, nasal turbulence, and need for speech/ language (S/L) intervention. *Fair agreement* (.21 to .40) was noted for the final three parameters: hyponasality, anterior oral CSCs, and noncleft speech immaturities/ errors (see Table 4). Table 5 shows the corresponding strength of agreement descriptors for the kappa statistic (Altman, 1991).

Interrater Reliability: Posttraining 1—Following training, improvement[1] was noted for four of the CAPS-A parameters. Higher kappas/corresponding strength of agreement categories were noted for hyponasality (.25 to .48), audible nasal emission (.43 to .72), nasal turbulence (.51 to .68), and anterior oral CSCs (.34 to .47). Lower kappas/strength of agreement categories were noted for two parameters: grimace (.71 to .44) and noncleft speech immaturities/errors (.28 to .20) (Table 4).

Interrater Reliability: Posttraining 2—Kappas/corresponding strength of agreement categories were maintained from posttraining 1 to posttraining 2 for a majority of parameters: voice, hyponasality, nasal turbulence, anterior oral CSCs, posterior oral CSCs, nonoral CSCs, passive CSCs, and noncleft speech immaturities/errors. Improvement in kappas/strength of agreement categories were noted for intelligibility (.76 to .82), hypernasality (.76 to .82), grimace (.44 to .67), and need for S/L intervention (.57 to .64). Higher kappas were noted for anterior oral CSCs (.47 to .51), posterior oral CSCs (.62 to .69), and nonoral CSCs (.65 to .73), but the changes did not result in different strength of agreement category assignments. Level of agreement fell close to the pretraining level (.43) for one variable: audible nasal emission (.72 to .57) (Table 4).

Intrarater Reliability: Posttraining 1 and Posttraining 2—Intrarater reliability was calculated for the two posttraining ratings. Very good agreement was noted for five of the parameters (intelligibility, hypernasality, posterior oral CSCs, nonoral CSCs, and passive CSCs). Good agreement was obtained for seven of the parameters (voice, hyponasality, audible nasal emission, nasal turbulence, grimace, noncleft speech/immaturities errors, and need for intervention). Anterior oral CSCs was the only category that showed moderate agreement (Table 6).

[1] A score was considered to be improved if a higher kappa was seen that also resulted in a change in the strength of agreement descriptor (Altman, 1991).

Study 2: NA Sample

Interrater Reliability: Rating 1—Interrater reliability was calculated (weighted and unweighted kappa statistic) for the six SLP raters and 11 parameters on the CAPS-A-AM using samples collected in NA. Interrater reliability scores indicated good agreement for 4 of 11 parameters: hypernasality, audible nasal emission, nonoral CSCs, and recommend VPF evaluation. Moderate agreement was noted for 2 of 11 parameters: anterior oral CSCs and passive CSCs. Fair agreement was noted for three parameters: intelligibility, hyponasality, and recommend speech therapy. Poor agreement was noted for voice and posterior oral CSCs (see Table 7).

Interrater Reliability: Rating 2—Interrater reliability scores calculated for the same six raters approximately 1 month later showed stable kappas/strength of agreement categories for six parameters: intelligibility, voice, hypernasality, posterior oral CSCs, recommend speech therapy, and recommend VPF evaluation. Interrater reliability improved for hyponasality (.39 to .67) and passive CSCs (.55 to .72). Lower kappas/levels of agreement categories were noted for audible nasal emission (.71 to .53), anterior oral CSCs (.45 to .38), and nonoral CSCs (.78 to .60) (see Table 7).

The kappas obtained for the NA samples appear to represent low reliability for some of the parameters; however, it is likely that these scores are associated with an anomaly that has been reported to occur with the kappa and ICC formulas for certain data sets. When scores cluster in a corner of the cross-tabulation table for categorical ratings, or the range of ratings for the study participants is very narrow for continuous ratings, the resulting kappa or ICC will be smaller than it should be based on the data (and often is a kappa or ICC of zero) (Stoddard, 2012). This occurs because there is insufficient variability among the participants (in our case, the severity of the speech samples) relative to the agreement that could occur by chance, so the formulas for kappa and ICC break down.

In the case of the NA data, it appears that the samples did not have sufficient diversity within several of the speech parameters to yield an accurate kappa. Given this situation, and in order to describe the interrater reliability more accurately, mean percent agreement scores were also calculated for each of the parameters. As can be seen in Table 7, those parameters with the lowest kappas (i.e., intelligibility, voice, hyponasality, posterior oral CSCs, and recommended speech therapy) had percent agreement scores equal to or greater than many of the parameters showing higher kappas (e.g., hypernasality).

The kappa formula problem might also explain the poorer interrater reliability scores on the UK/I reliability ratings for hyponasality. When mean percent agreement scores were calculated for this measure for the UK/I data, the scores were 81%, 80%, and 87% for pretraining, posttraining 1, and posttraining 2, respectively.

Intrarater Reliability: NA Samples—Intrarater reliability was calculated (weighted and unweighted kappa statistic) for the six SLP raters for the same 11 parameters described above (see Table 8). The findings indicated very good agreement for 4 of the 11 parameters: hypernasality, anterior oral CSCs, nonoral CSCs, and recommend VPF evaluation. Good agreement was noted for five parameters: intelligibility, voice, hyponasality, audible nasal emission, and passive CSCs. Finally, moderate agreement was noted for posterior oral CSCs and recommend speech therapy.

Comparison Between UK/I and NA Samples

Interrater Reliability—Due to the nature of the samples and the noted problem with the kappa statistic, it is difficult to compare interrater reliability for the UK/I and NA data sets. However, a comparison of the weighted/unweighted kappas at equivalent points for the two data sets (i.e., posttraining 1 for the UK/I and rating 1 for NA data sets) showed similar scores for hypernasality, audible nasal emission, and anterior oral CSCs. Although the kappa for nonoral CSCs was higher for the NA samples than the UK/I samples, the strength of agreement category was good agreement for both. The kappas for the UK/I samples were higher than those obtained for the NA samples for six parameters: intelligibility, voice, hyponasality, posterior oral CSCs, passive CSCs, and need for intervention/recommend speech therapy. We found it interesting that in the NA data, all of these parameters, with the exception of passive CSCs, yielded problematic kappas. If we compare the mean percent agreement scores for these parameters, higher percent agreement scores were noted in the NA data for half of the variables and lower scores were noted for the other half (Table 9).

Intrarater Reliability—A comparison of the intrarater reliability scores for the two data sets indicated similar ratings for 5 of the 10 parameters: voice, hypernasality, hyponasality, audible nasal emission, and nonoral speech errors. Lower kappas were noted for the NA data compared with the UK/I data for four parameters: intelligibility, posterior oral CSCs, passive CSCs, and need for intervention/recommend speech therapy. The kappas for passive CSCs were very similar for the UK/I and NA data (.81 and .78, respectively); although, the strength of agreement category assignment differed. Finally, a higher kappa was noted for anterior oral CSCs for the NA data (Table 10).

Discussion

The purpose of this study was to describe the results of two reliability studies that were carried out using a tool designed to evaluate speech outcomes for intercenter collaborative studies. In the first study, we also attempted to test the effect of training on interrater reliability scores. The study design did not allow us to examine the influence of training on intrarater reliability scores because two different pretraining sessions were not conducted. However, as is often seen, intrarater reliability scores were typically higher than interrater reliability scores for both Study 1 and Study 2.

Consistent with other investigations, this study showed that interrater reliability can be improved following a program of systematic training (Chan and Yiu, 2002, 2006; Eadie and Baylor, 2006; Ellis and Beltyukova, 2008; Lee et al., 2009) and that training effects are maintained for at least 1 month posttraining. However, the effects of training were not consistent across all speech outcome parameters, given that reliability scores improved for some parameters but not others. For example, interrater reliability improved for 6 of the 13 parameters following training (at either posttraining 1 or posttraining 2), six parameters did not change (or the change was not maintained), and kappas for one parameter decreased following training. It is not clear why ratings for some parameters improved more than others. However, the following factors may have played a role: (1) variability in the amount of time given to the various speech parameters during training, (2) the relative difficulty of the parameters rated, and/or (3) the raters' experience with and training on the different parameters prior to this study.

The reliability results obtained here are not directly comparable to those obtained by Sell et al. (2009), because they reported average-measure ICCs, which tend to be higher than single-measure ICCs (Hallgren, 2012). In this study, weighted or unweighted kappas were used, which are interchangeable with single-measure ICCs (two-way mixed model) but are lower than average-measure ICCs when calculated on the same data set (e.g., Hallgren, 2012; Stoddard, 2012). For example, the average-measure ICC (two-way mixed model) for hypernasality on the NA data set was .93, compared with the weighted kappa (or single-measure ICC) of .70 that was reported here. In both studies, however, higher levels of agreement were obtained for intelligibility, hypernasality, grimace, posterior oral CSCs, nonoral CSCs, and passive CSCs; moderate levels for audible nasal emission and need for intervention; and lower levels for anterior oral CSCs and noncleft speech immaturities/ errors.

It is interesting, and also similar to Sell et al. (2009), that many of the parameters showing higher levels of interrater reliability were those typically associated with velopharyngeal function (i.e., hypernasality, posterior oral CSCs, nonoral CSCs, and passive CSCs). This is encouraging, especially in the case of hypernasality, given that Lohmander and colleagues (Brunnegård and Lohmander, 2007; Lohmander et al., 2012) described the challenges of obtaining good interrater reliability for hypernasality. Another measure related to velopharyngeal function, audible nasal emission, reached only moderate levels of agreement in our study as well as in the Sell et al. (2009) study. The lower level of listener agreement for this parameter might be related to the finding by Baylis et al. (2011) and Baylis and colleagues (2015) showing that audible nasal emission is a prothetic phenomenon and that validity and reliability of audible nasal emission judgments can be improved using a ratio-based rating procedure such as VAS rather than the EAI method that was used here. It also may be that audible nasal emission lends itself less well to scaling and that a binary judgment of present or absent, as recommended by Henningsson et al. (2008), is more accurate and would yield better reliability.

The low interrater reliability seen for the parameter anterior oral CSCs is not atypical (Sell et al., 2009). This parameter requires raters to make judgments about whether a dental/ dentalization, lateral/lateralization, or palatal/palatalization of the target consonant occurred. In many cases, this is indicated by diacritics and results in allophonic rather than phonemic differences. Shriberg and Lof (1991) found that even transcribers from the same lab who spent between 12 and 20 hours a week transcribing together (consensus transcription) were not able to achieve acceptable levels of agreement using narrow transcription (i.e., use of diacritics). As explained by these authors, "The most direct explanation for the low levels of narrow phonetic transcription agreement is that children's speech productions frequently contain confusing acoustic cues relative to the phoneme and allophone boundaries expected by the ambient community" (Shriberg and Lof, 1991, p. 255). They concluded that the poor reliability scores suggest that the judgments may be beyond our perceptual capabilities and went on to say that "narrow phonetic transcription may be unreliable for certain of the purposes for which it currently is used in communicative disorders" (Shriberg and Lof, 1991, p. 273).

It should be noted also that one of the anterior oral CSCs, palatal/palatalization, might result in a phoneme substitution, commonly referred to as the *middorsum palatal stop* (MDPS). These have been observed in the speech of children with cleft palate but are not sounds that occur in English. Work by Santelmann and colleagues (1999) suggested that although listeners were able to discriminate MDPSs from other stops, they were identified as alveolar or velar stops. Acoustic analysis of MDPSs, alveolar stops, and velar stops produced by the same children suggested that although MDPSs were similar to other stops acoustically, acoustic cues varied for individual speakers, which may also contribute to the low levels of reliability noted for this category (Sussman and Chapman, 2000). Additionally, Gibbon and Crampin (2001) found that an adult speaker with a repaired cleft palate produced MDPSs with a lateral release. If a lateral release is an articulatory feature of MDPSs for all speakers with cleft palate, it may partially explain why MDPSs are difficult to identify (Gibbon and Crampin, 2001). Whereas improvement was noted with training for anterior oral CSCs, this category continued to show one of the lowest levels of reliability posttraining. Whether reliability can be improved with additional training is yet to be determined.

Another variable with lower interrater reliability was noncleft speech immaturities/errors. The CAPS-A instructs the rater to indicate "if any non-cleft speech immaturities/ errors are present that require an SLT [speech-language therapist] to monitor or intervene with treatment" (John et al., 2006). Sell et al. (2009) provided a convincing list of reasons for the low level of agreement associated with this outcome, including limited training, subjective nature of the judgments, and overlap between some cleft and noncleft speech immaturities/ errors (see Sell et al., 2009, for a more detailed description). Furthermore, whether the child's production should be considered an error or developmentally appropriate depends on the normative data that is being used. Looking across 10 studies that examined consonant development in young children, there was considerable variability in acquisition of the /s/ sound for English-speaking children. For example, the age of acquisition was listed as 3 years (Prather et al., 1975; Smit et al., 1990 [for girls]; Dodd et al., 2003), 4 years (Arlt and Goodban, 1976; Chirlian and Sharpley, 1982), 4½ years (Templin, 1957; Kilminster and Laird, 1978), 5 years (Smit et al., 1990 [for boys]), and 5½ years (Anthony et al., 1971) across studies. This judgment becomes even more difficult if consideration is given to the actual type of error that is being produced (Smit, 1993). Due to these issues, this parameter was deleted for the version of the CAPS-A-AM used in Study 2.

Interrater Reliability NA Samples

The second reliability study was conducted following changes/clarifications in the CAPS-A protocol (i.e., CAPS-A-AM) based on the results of Study 1. The expectation was that the reliability scores would be higher due to changes in the scoring protocol and use of samples with children speaking Canadian or American English, yet that did not appear to be the case. Much has been written about the problems with using the kappa statistic when the parameters to be rated are infrequently occurring (Viera and Garrett, 2005) or when a majority of the data fall into one category of the scale (Byrt et al., 1993; Stoddard, 2012). In the current data set, many parameters with low kappas had percent agreement scores that were higher than the percent agreement scores seen for other parameters with good kappas. This suggests that the "decrease" in kappa scores from Study 1 to Study 2 was not a true decline but a known statistical anomaly with kappa. An examination of the individual ratings for those parameters that showed lower kappas in Study 2 compared with Study 1 (intelligibility, voice, hyponasality, posterior CSCs, passive CSCs, and need for intervention) supports this proposal. For example in the case of intelligibility, there are five possible categories ranging from *normal* to *impossible to understand*. Yet, none of the samples in Study 2 received a rating of 1 or 4, and 67% of the samples were rated as 3. In contrast, the samples in Study 1 were assigned scores from 0 to 4, with the scores being fairly evenly distributed across all 10 samples. This lack of variability was also noted for the parameters need for intervention (all but one sample was rated as needing intervention), posterior oral CSCs (rated as present in one sample), voice (rated as present in two samples), and hyponasality (rated as present in three samples). In contrast, only one other parameter, passive CSCs, showed a drop in agreement; but for this parameter, the percent agreement score was also low, suggesting that the kappa more accurately reflected the true reliability. This was confirmed by an examination of the frequency of occurrence of passive CSCs, which showed that about one half of the samples contained examples of these speech characteristics, which was similar to Study 1.

These analyses suggest that differences in the samples rated in Study 1 versus Study 2 accounted for the drop in agreement scores. The lack of variability in the NA samples was likely related to the fact that whereas these samples were chosen to represent a range of severity levels and speech behaviors, the pool of samples available for rating was somewhat restricted. In addition to causing issues with regard to the kappa statistic, the lack of diversity of the NA samples may also have impacted the difficulty of the rating task, especially for the parameter intelligibility, because a majority of the samples clustered in the middle of the severity range. Although not proven for intelligibility specifically, it has been reported that it is typically easier to rate speech that is normal or severely disordered compared with samples representing the midrange of the scale (Kent, 1996; Shrivastav et al., 2005), which appeared to be the case for the NA samples for intelligibility.

In summary, the findings of this study suggest that ratings of speech outcome can be reliably performed on the speech of children with cleft palate for a majority of the parameters on the CAPS-A and the CAPS-A-AM. Furthermore, it shows that training can and should be used to improve interrater reliabilty and validity of speech ratings for researchers and clinicians involved in the study and treatment of children with cleft palate. At the same time, although improvements were noted for some of the parameters following training, it was not uniform across parameters. This is not surprising given that we know that not all perceptual categories are equal in terms of stability of ratings (Kent, 1996). Additional training, including consensus listening and strict adherence to category definitions (among other things), may result in higher agreement for some parameters. However, it might be that in some cases we have reached our "perceptual limits" for certain parameters. If this proves to be true, and the judgments are not repeatable, consideration should be given to excluding them from research reports, as suggested by Shriberg and Lof (1991). Fortunately, for examination of speech outcome in children with cleft palate, acceptable levels for reliability were achieved for a majority of those parameters important for overall evaluation of speech articulation and velopharyngeal function.

Future research should examine the influence of training on intrarater reliability scores. Studies should focus on identifying those protocols and scaling methods that result in the highest levels of both interrater and intrarater reliability. Additionally, attention should be directed toward studying observer drift, which can be particularly problematic in the case of longitudinal studies or studies where data analysis continues over a long period of time without provisions for recalibration of listeners.

Finally, it is not enough to report interrater and intrarater reliability if we do not use the appropriate statistical procedures to calculate reliability. Continued attention to issues of reliability and validity of the measures that we use in clinical and research assessments of individuals with communication impairments, including those related to cleft lip and palate, should be considered an important priority because those measures are the cornerstone of our research and clinical endeavors.

Acknowledgments

This research was funded in part by the American Cleft Palate–Craniofacial Association, the Cleft Palate Foundation, and the National Institutes of Health–National Institute of Dental and Craniofacial Research R21DE023519.

References

Altman, DG. Practical Statistics for Medical Research. London: Chapman and Hall/CRC Press; 1991. p. 404

Anthony, A., Bogle, D., Ingram, TTS., McIsaac, MW. The Edinburgh Articulation Test. Edinburgh: E & S Livingstone; 1971.

Arlt PB, Goodban MT. A comparative study of articulation acquisition as based on a study of 240 normals, aged three to six. Lang Speech Hear Serv Sch. 1976; 7:173–180.

Ball MJ, Rahilly J. Transcribing disordered speech: the segmental and prosodic layers. Clin Linguist Phon. 2002; 16:329–344. [PubMed: 12185981]

Bassich CJ, Ludlow CL. The use of perceptual methods by new clinicians for assessing voice quality. J Speech Hear Disord. 1986; 51:125–133. [PubMed: 3702360]

Baylis AL, Munson B, Moller KT. Perceptions of audible nasal emission in speakers with cleft palate: a comparative study of listener judgments. Cleft Palate Craniofac J. 2011; 48:399–411. [PubMed: 20572776]

Baylis AL, Whitehill TL, Chapman KL. Americleft Speech Group. Validity and reliability of visual analog scaling for assessment of speech outcomes in children with repaired cleft palate. Cleft Palate Craniofac J. 2015; 52:660–670. [PubMed: 25322442]

Bradford LJ, Brooks AR, Shelton RL. Clinical judgment of hypernasality in cleft palate children. Cleft Palate J. 1964; 6:329–335.

Britton L, Albery E, Bowden M, Harding-Bell A, Phippen G, Sell D. A cross-sectional cohort study of speech in 5-year-olds with cleft palate ± lip to support development of national audit standards. Cleft Palate Craniofac J. 2014; 51:431–451. [PubMed: 24635034]

Brunnegård K, Lohmander A. A cross-sectional study of speech in 10- year-old children with cleft palate: results and issues of rater reliability. Cleft Palate Craniofac J. 2007; 44:33–44. [PubMed: 17214536]

Brunnegård K, Lohmander A, van Doorn J. 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. 2009; 44:656–674. [PubMed: 18821109]

Bunton K, Kent RD, Duffy JR, Rosenbek JC, Kent JF. Listener agreement for auditory-perceptual ratings of dysarthria. J Speech Lang Hear Res. 2007; 50:1481–1495. [PubMed: 18055769]

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993; 46:423–429. [PubMed: 8501467]

Chan KMK, Yiu EML. A comparison of two perceptual voice evaluation training programs for naive listeners. J Voice. 2006; 20:229–241. [PubMed: 16139475]

Chan KMK, Yiu EML. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res. 2002; 45:111–126. [PubMed: 14748643]

Chirlian NS, Sharpley CF. Children's articulation development: some regional differences. Aust J Hum Commun Disord. 1982; 10:23–30.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20:37–46.

Cordes AK. Individual and consensus judgments of dysfluency types in the speech of persons who stutter. J Speech Lang Hear Res. 2000; 43:951–964. [PubMed: 11386481]

Cordes AK. The reliability of observational data: I. Theories and methods for speech-language pathology. J Speech Lang Hear Res. 1994; 37:264–278.

Cordes AK, Ingham RJ. Effects of time-interval judgment training on real-time measurement of stuttering. J Speech Lang Hear Res. 1999; 42:862–879. [PubMed: 10450907]

Cordes AK, Ingham RJ. The reliability of observational data: II. Issues in the identification and measurement of stuttering events. J Speech Lang Hear Res. 1994; 37:279–294.

Counihan DT, Cullinan DL. Reliability and dispersion of nasality ratings. Cleft Palate J. 1970; 7:261–270. [PubMed: 5266336]

Cucchiarini C. Assessing transcription agreement: methodological aspects. Clin Linguist Phon. 1996; 10:131–155.

Dalston RM, Marsh JL, Vig KW, Witzel MA, Bumsted RM. Minimal standards for reporting the results of surgery on patients with cleft lip, cleft palate, or both: a proposal. Cleft Palate J. 1988; 25:3–7. [PubMed: 3422597]

Daniel HJ. Nasality ratings of single words, phrases, and running speech samples obtained from cleft palate children. Folia Phoniatr Logop. 1971; 23:41–49.

de Vet HCW, Terwee CB, Know DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006; 59:1033–1039. [PubMed: 16980142]

Dodd B, Holm A, Hua Z, Crosbie S. Phonological development: a normative study of British English-speaking children. Clin Linguist Phon. 2003; 17:617–643. [PubMed: 14977026]

Dotevall H, Lohmander-Agerskov A, Ejnell H, Bake B. Perceptual evaluation of speech and velopharyngeal function in children with and without cleft palate and the relationship to nasal airflow patterns. Cleft Palate Craniofac J. 2002; 39:409–424. [PubMed: 12071789]

Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. J Voice. 2006; 20:527–544. [PubMed: 16324823]

Ellis LW, Beltyukova S. Effects of training on naïve listeners' judgments of the speech intelligibility of children with severe-to-profound hearing loss. J Speech Lang Hear Res. 2008; 51:1114–1123. [PubMed: 18664708]

Bronsted K, Grunwell P, Henningsson G, Jansonius K, Karling J, Meijer M, Ording D, Sell D, Vermei-Zieverink E, Wyatt R. Eurocleft Speech Group. A phonetic framework for the cross-linguistic analysis of cleft palate speech. Clin Linguist Phon. 1994; 8:109–125.

Grunwell P, Bronsted K, Henningsson G, Jansonius K, Karling J, Meijer M, Ording D, Wyatt R, Sell D, Vermei-Zieverink E. Eurocleft Speech Group. A six-centre international study of the outcome of treatment in patients with clefts of the lip and palate: the results of a cross-linguistic investigation of cleft palate speech. Scand J Plast Reconstr Surg Hand Surg. 2000; 34:219–229. [PubMed: 11020918]

Fleiss, JL., Levin, B., Pai, MC. Statistical Methods for Rates and Proportions. 3. Hoboken, NJ: John Wiley & Sons; 2003. p. 609

Fujiwara Y, Henningsson G, Ainoda N. A review of Japanese articles on perceptual assessment of speech sounds in individuals with cleft palate. Jpn J Logoped Phoniatr. 2006; 47:252–257.

Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. J Speech Lang Hear Res. 1993; 36:14–20.

Gibbon FE, Crampin L. An electropalatographic investigation of middorsum palatal stops in an adult with repaired cleft palate. Cleft Palate Craniofac J. 2001; 38:96–105. [PubMed: 11294548]

Gooch JL, Hardin-Jones M, Chapman KL, Trost-Cardamone JE, Sussman J. Reliability of listener transcriptions of compensatory articulations. Cleft Palate Craniofac J. 2001; 38:59–67. [PubMed: 11204684]

Grunwell, P., Sell, D., Harding, A. Describing cleft palate speech. In: Grunwell, P., editor. Analysing Cleft Palate Speech. London: Whurr; 1993.

Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012; 8:23–34. [PubMed: 22833776]

Helou LB, Solomon NP, Henry LR, Coppit GL, Howard RS, Stojadinovic A. The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V): ratings of postthyroidectomy voice. Am J Speech Lang Pathol. 2010; 19:248–258. [PubMed: 20484704]

Henningsson G, Kuehn D, Sell D, Sweeney T, Trost-Cardamone J, Whitehill T. Universal parameters for reporting speech outcomes in individuals with cleft palate. Cleft Palate Craniofac J. 2008; 45:1–17. [PubMed: 18215095]

Hutters B, Henningsson G. Speech outcome following treatment in cross-linguistic cleft palate studies: methodological implications. Cleft Palate Craniofac J. 2004; 41:544–549. [PubMed: 15352862]

Imatomi S. Effects of breathy voice source on ratings of hypernasality. Cleft Palate Craniofac J. 2005; 42:641–648. [PubMed: 16241176]

John A, Sell D, Sweeney T, Harding-Bell A, Williams A. The cleft audit protocol for speech-augmented: a validated and reliable measure for auditing cleft speech. Cleft Palate Craniofac J. 2006; 43:272–288. [PubMed: 16681400]

Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. Am J Speech Lang Pathol. 1996; 5:7–23.

Keuning KHD, Wieneke GH, Dejonckere PH. The intrajudge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: the effect of judges and speech samples. Cleft Palate Craniofac J. 1999; 36:328–333. [PubMed: 10426599]

Kilminster M, Laird E. Articulation development in children aged three to nine years. Aust J Hum Commun Disord. 1978; 6:23–30.

Klinto K, Slameh E, Svensson H, Lohmander A. The impact of speech material on speech judgment in children with and without cleft palate. Int J Lang Commun Disord. 2011; 46:348–360. [PubMed: 21575075]

Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Steiner DL. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int J Nurs Stud. 2011; 48:661–671. [PubMed: 21514934]

Kreiman J, Gerratt BR. Perceptual assessment of voice quality: past, present and future. SIG3 Perspect Voice Voice Disord. 2010; 22:62–67.

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. J Speech Lang Hear Res. 1993; 36:21–40.

Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. J Speech Lang Hear Res. 1990; 33:103–115.

Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. J Speech Hear Res. 1992; 35:512–520. [PubMed: 1608242]

Kuehn DP, Moller KT. Speech and language issues in the cleft palate population: the state of the art. Cleft Palate Craniofac J. 2000; 37:348–348. [Accessed March 7, 2012] Available at http://dx.doi.org/10.1597/1545-1569(2000)037%3C0348:SALIIT%3E2.3.CO;2.

Lee A, Whitehill T, Ciocca V. Effect of listener training on perceptual judgment of hypernasality. Clin Linguist Phon. 2009; 23:319–334. [PubMed: 19399664]

Lewis KE. Reporting observer agreement on stuttering event judgments: a survey and evaluation of current practice. J Fluency Disord. 1994; 19:269–284.

Lewis KE, Watterson TL, Houghton SM. The influence of listener experience and academic training on ratings of nasality. J Commun Disord. 2003; 36:49–58. [PubMed: 12493637]

Lockhart R, McLeod S. Factors that enhance English-speaking speech-language pathologists' transcription of Cantonese-speaking children's consonants. Am J Speech Lang Pathol. 2013; 22:523–539. [PubMed: 23813201]

Lohmander, A. Surgical interventions and speech outcomes in cleft lip and palate. In: Howard, S., Lohmander, A., editors. Cleft Palate Speech: Assessment and Intervention. West Sussex: Wiley-Blackwell; 2011. p. 55-85.

Lohmander, A., Borell, E., Henningsson, G., Havstram, C., Lundeborg, I., Persson, C. SVANTE—Svenskt Artikulations-och Nasalitets Test [manual]. Malmo: Pedogogisk Design; 2005.

Lohmander A, Friede H, Lilja J. Long-term, longitudinal follow-up of individuals with unilateral cleft lip and palate after the Gothenburg primary early veloplasty and delayed hard palate closure protocol: speech outcomes. Cleft Palate Craniofac J. 2012; 49:657–671. [PubMed: 22364610]

Lohmander A, Olsson M. Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature. Cleft Palate Craniofac J. 2004; 41:64–70. [PubMed: 14697067]

Lohmander A, Willadsen E, Persson C, Henningsson M, Bowden B, Hutters B. Methodology for speech assessment in the Scandcleft project—an international randomized clinical trial on palatal surgery: experiences from a pilot study. Cleft Palate Craniofac J. 2009; 46:347–362. [PubMed: 19642772]

Long RE, Hathaway R, Daskalogiannakis J, Mercado A, Russell K, Cohen M, Semb G, Shaw W. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate, part 1. Principles and study design. Cleft Palate Craniofac J. 2011; 48:239–243. [PubMed: 21219224]

McNutt JC, Wicki L, Paulsen J. Judgments of phoneme errors under four modes of audio-visual presentation. J Speech Lang Pathol Audiol. 1991; 15:37–42.

McWilliams BJ. Some factors in the intelligibility of cleft-palate speech. J Speech Hear Disord. 1954; 19:524–528. [PubMed: 13222468]

McWilliams, BJ., Morris, HL., Shelton, RL. Cleft Palate Speech. Philadelphia: BC Decker; 1990.

McWilliams, BJ., Philips, BJ. Velopharyngeal Incompetence. Philadelphia: BC Decker; 1990.

McWilliams, BJ., Philips, BJ. Velopharyngeal Incompetence: Audio Seminars in Speech Pathology. Philadelphia: WB Saunders; 1979.

Munson B, Johnson JM, Edwards J. The role of experience in the perception of phonetic detail in children's speech: a comparison between speech-language pathologists and clinically untrained listeners. Am J Speech Lang Pathol. 2012; 21:124–139. [PubMed: 22230182]

Oates J. Auditory-perceptual evaluation of disordered voice quality. Folia Phoniatr Logop. 2009; 61:49–56. [PubMed: 19204393]

Pereira V, Tuomainen J, Sell D. Perceptual outcomes of maxillary osteotomy in patients with cleft lip and palate. Int J Lang Commun Disord. 2013; 48:640–650. [PubMed: 24165361]

Peterson-Falzone SJ. The relationship between timing of cleft palate surgery and speech outcome: what have we learned, and where do we stand in the 1990s? Semin Orthod. 1996; 2:185–191. [PubMed: 9161287]

Prather EM, Hedrick DL, Kern CA. Articulation development in children aged two to four years. J Speech Hear Disord. 1975; 40:179–191. [PubMed: 1234929]

Santelmann H, Sussman J, Chapman K. Perception of middorsum palatal stops from the speech of three children with repaired cleft palate. Cleft Palate Craniofac J. 1999; 36:233–242. [PubMed: 10342611]

Schellinger, SK., Edwards, J., Munson, B., Beckman, MK. The role of listener expectations on judgments of children's /s/ productions. Presented at the Symposium on Research in Child Language Disorders; June 2008; Madison, Wisconsin.

Sell D. Issues in perceptual speech analysis in cleft palate and related disorders: a review. Int J Lang Commun Disord. 2005; 40:103–121. [PubMed: 16101269]

Sell D, Grunwell P, Mildinhall S, Murphy T, Cornish TC, Williams A, Bearn D, Shaw WC, Murray J, Sandy J. Cleft lip and palate care in the United Kingdom—the Clinical Standards Advisory Group (CSAG) study. Part 3: speech outcomes. Cleft Palate Craniofac J. 2001; 38:30–37. [PubMed: 11204679]

Sell D, Harding A, Grunwell P. Revised GOS.SP.ASS (98): speech assessment for children with cleft palate and/or velopharyngeal dysfunction. Int J Lang Commun Disord. 1999; 34:7–33.

Sell D, Harding A, Grunwell P. A screening assessment of cleft palate speech (Great Ormond Street Speech Assessment). Eur J Disord Commun. 1994; 29:1–15. [PubMed: 8032102]

Sell, D., Harding-Bell, A., Sweeney, T., Freeman, J., John, A. CAPS-A Speech Audit Tool Training Programmes—are they necessary?. Presented at the Meeting of the Craniofacial Society of Great Britain and Ireland; April 2010; Liverpool, United Kingdom.

Sell D, John A, Harding-Bell A, Sweeney T, Hegarty F, Freeman J. Cleft Audit Protocol for Speech (CAPS-A): a comprehensive training package for speech analysis. Int J Lang Commun Disord. 2009; 44:529–548. [PubMed: 18821108]

Shriberg LD, Lof GL. Reliability studies in broad and narrow phonetic transcription. Clin Linguist Phon. 1991; 5:225–279.

Shriberg LD, McSweeny JL, Anderson BE, Campbell TF, Chial MR, Green JR, Hauner KK, Moore CA, Rusiewicz HL, Wilson DL. Transitioning from analog to digital audio recording in childhood speech sound disorders. Clin Linguist Phon. 2005; 19:335–359. [PubMed: 16019779]

Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. J Speech Lang Hear Res. 2005; 48:323–335. [PubMed: 15989395]

Smit AB. Phonologic error distributions in the Iowa-Nebraska Articulation Norms. J Speech Lang Hear Res. 1993; 36:533–537.

Smit AB, Hand L, Freilinger JJ, Bernthal JE, Bird A. The Iowa articulation norms project and its Nebraska replication. J Speech Hear Disord. 1990; 55:779–798. [PubMed: 2232757]

Spriestersbach DC, Powers GR. Nasality in isolated vowels and connected speech of cleft palate speakers. J Speech Hear Res. 1959; 2:40–45. [PubMed: 13655290]

Stevens I, Daniloff RG. Trouble with /s/: a methodological study of factors affecting the judgment of misarticulated /s/. J Commun Disord. 1977; 10:207–220. [PubMed: 903406]

Stoddard, GJ. Biostatistics and Epidemiology Using Stata: A Course Manual. Salt Lake City, UT: University of Utah School of Medicine; Available at http://www.ccts.utah.edu/biostats/?pageId=5385 [Accessed February 20, 2012]

Streiner, DL., Norman, GR. Health Measurement Scales: A Practical Guide to Their Development and Use. New York: Oxford University Press; 1995. p. 118

Sussman, J., Chapman, KL. Acoustic analysis of mid-dorsum, alveolar, and velar stop consonants. Presented at the American Speech- Language-Hearing Association Convention; November 2000; Washington, DC.

Sweeney, T. Nasality—assessment and intervention. In: Howard, S., Lohmander, A., editors. Cleft Palate Speech Assessment and Intervention. Oxford: Wiley and Blackwell; 2011. p. 199-216.

Templin, MC. Certain Language Skills in Children [monograph]. Minneapolis: University of Minnesota Press; 1957.

Tönz M, Schmid I, Graf M, Mischler-Heeb R, Weissen J, Kaiser G. Blinded speech evaluation following pharyngeal flap surgery by speech pathologists and lay people in children with cleft palate. Folia Phoniatr Logop. 2002; 54:228–295.

Trost-Cardamone, JE. American English Sentence Sample: a controlled sample for assessing cleft palate speech outcome. Presented at the Meeting of the American Cleft Palate–Craniofacial Association; April 2012; San Jose, California.

Van Hattum RJ. Articulation and nasality in cleft palate speakers. J Speech Hear Disord. 1958; 1:383–387. [PubMed: 13599159]

Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–363. [PubMed: 15883903]

Watterson T, Lewis K, Allord M, Sulprizio S, O'Neill P. Effect of vowel type on reliability of nasality ratings. J Commun Disord. 2007; 40:503–512. [PubMed: 17391692]

Whitehill TL, Lee ASY, Chun JC. Direct magnitude estimation and interval scaling of hypernasality. J Speech Lang Hear Res. 2002; 45:80–88. [PubMed: 14748640]

Wolfe V, Martin D, Borton D, Youngblood HC. The effect of clinical experience on cue trading for the /r-w/ contrast. Am J Speech Lang Pathol. 2003; 12:221–228. [PubMed: 12828535]

Yeung, AC. Dissertation. Pokfulam, Hong Kong: University of Hong Kong; 2010. Effects of Listening Conditions on Perceptual Ratings of Hypernasal Speech.

Young, MA. Identification of stuttering and stutterers. In: Curlee, RF., Perkins, WH., editors. Nature and Treatment of Stuttering: New Directions. San Diego: College-Hill; 1984. p. 13-30.

Zraick RI, Liss JM. A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. J Speech Lang Hear Res. 2000; 43:979–988. [PubMed: 11386483]

APPENDIX A American English Sentence Sample (Trost-Cardamone, 2012)

These sentences are structured to assess consonant targets in *all positions of occurrence* in English (word initial, medial, and final). Modifications to the sentence sample were made to conform to the Cleft Audit Protocol for Speech–Augmented (CAPS-A) (John et al., 2006). The CAPS-A assesses sounds *in word initial and word final positions only.* Initial and final targets assessed are bolded and underlined. Sentence No. 11 has an alternate sentence for assessing all three positions (“Laura will wear a lily”). Sentences No. 21 through 23 were not scored, given that they are not scorable in the CAPS-A protocol. Sentence No. 25 was not included in the sample.

1.	/m/	**M**om ’n’ Amy are ho**m**e
2.	/p/	Puppy will **p**ull a ro**p**e
3.	/b/	Buy **b**aby a bi**b**
4.	/f/	A fly **f**ell off a lea**f**
5.	/v/	I lo**v**e every **v**iew
6.	/θ/	**Th**irty-two tee**th**
7.	/ð/	**Th**e other feather
8.	/n/	Anna **kn**ew no o**n**e
9.	/t/	Your **t**urtle ate a ha**t**
10.	/d/	**D**o it today for Da**d**
11.	/l/	**L**aura will ye**ll**
12.	/s/	Sissy **s**aw Sally ra**c**e
13.	/z/	**Z**oey ha**s** roses
14.	/ʃ/	**Sh**e washed a di**sh**
15.	/tʃ/	Wa**tch** a **ch**oo-choo
16.	/dʒ/	Geor**g**e saw **G**igi
17.	/ŋ/	We are hangi**ng** on
18.	/k/	A **c**ookie or a ca**k**e
19.	/g/	**G**ive Aggie a hu**g**
20.	/h/	Hurry ahead **H**arry
21.	/r/	**R**ay will arrive early
22.	/w/	We **w**ere away
23.	/m,n,ŋ/	We ra**n** a lo**ng** **m**ile

24.	/s/ clusters	I **sp**y a **st**arry **sk**y
25.	nasals, high pressure, low pressure targets—Summer is sunny, winter is windy and cold	

TABLE 1

CAPS-A Core Speech Parameters Assessed and Corresponding Scalar Values[*]

Parameter	**Scalar Values for Rating**
Intelligibility/distinctiveness	5-point scale: 0–4
Voice characteristics	Binary rating: 0 or 1
Resonance	
○ Hypernasality	5-point scale: 0–4
○ Hyponasality	3-point scale: 0–2
Nasal airflow	
○ Audible nasal emission	3-point scale: 0–2
○ Nasal turbulence	3-point scale: 0–2
Grimace	Binary rating: 0 or 1
Consonant production—cleft speech characteristics (CSCs)	
○ Anterior oral CSCs (dentalization, lateralization, palatalization)	3-point scale: 0–2
○ Posterior oral CSCs (double articulation, backed to velar/uvular)	3-point scale: 0–2
○ Nonoral CSCs (pharyngeal articulation, glottal articulation, active nasal fricatives, double articulation)	3-point scale: 0–2
○ Passive CSCs (weak and/or nasalized consonants, nasal realization of plosives and/or suspected passive nasal fricative, gliding of fricatives/affricates)	3-point scale: 0–2
Noncleft speech immaturities/errors	Binary rating: 0 or 1
Evidence of influencing factors (general comments on child’s speech, language, or hearing)	Qualitative; not rated
Perceived need (speech and language therapy required for cleft speech problems at some point)	Yes__ No__

[*] CAPS-A =Cleft Audit Protocol for Speech–Augmented (John et al., 2006).

A rating of 0 means there is no deviation; a value of 8 is assigned when the parameter cannot be rated.

TABLE 2

Listing Protocol for Rating Speech Samples Using the CAPS-A[*]

Sample Component	Medium	Action
Spontaneous speech	Audio only—still facial image	Rate *Intelligibility/Distinctiveness* based on one listening; no relistening
Counting 1–20, 60–70, Elicited nursery rhyme, Sentence repetition	Audio only—still facial image	Rate all parameters except *Grimace*
Counting 1–20, 60–70, Elicited nursery rhyme, Sentence repetition	Digital video with audio	Rate all parameters; review, revise ratings except for *Intelligibility*; review, revise consonant transcriptions

[*] Modified from Sell et al. (2009).

TABLE 3

Americleft Modifications to CAPS-A Parameters and Definitions

Parameters	Modifications
Grimace	Deleted
Audible nasal emission & nasal turbulence	Collapsed into one rating category labeled "Audible Nasal Emission/Nasal Turbulence"
Active nasal fricative & passive nasal fricative	Collapsed into the single error category of "Nasal Fricative"
Definitions	*Modifications*
"Perceived Need" category and subcategory of "Speech and language therapy required for cleft speech problems at some point" (yes/no)	Modified to "Perceived Need for Speech Management" with two subcategories of **1** Recommend speech therapy (yes/no) **2** Recommend further investigation of velopharyngeal function (yes/no)
Intelligibility/distinctiveness	Working definition of this parameter modified from: "Rate the ability of an unfamiliar listener to understand speech that is heard" to "Rate your ability to understand speech that is heard"

TABLE 4

Interrater Reliability for UK/I Samples: Kappas and Strength of Agreement Pre, Post 1, and Post 2

	Pre		Post 1		Post 2	
Parameter	**Kappa**	**Strength of Agreement**[*]	**Kappa**	**Strength of Agreement**	**Kappa**	**Strength of Agreement**
Intelligibility	.75	Good	.76	Good	.82	Very good
Voice	.66	Good	.63	Good	.63	Good
Hypernasality	.70	Good	.76	Good	.82	Very good
Hyponasality	.25	Fair	.48	Moderate	.50	Moderate
Audible nasal emission	.43	Moderate	.72	Good	.57	Moderate
Nasal turbulence	.51	Moderate	.68	Good	.66	Good
Grimace	.71	Good	.44	Moderate	.67	Good
Anterior oral CSCs	.34	Fair	.47	Moderate	.51	Moderate
Posterior oral CSCs	.67	Good	.62	Good	.69	Good
Nonoral CSCs	.67	Good	.65	Good	.73	Good
Passive CSCs	.66	Good	.75	Good	.74	Good
Noncleft speech errors	.28	Fair	.20	Poor	.16	Poor
Need for intervention	.56	Moderate	.57	Moderate	.64	Good

[*] Altman (1991).

TABLE 5

Strength of Agreement for Kappa Statistic[*]

Value of κ	Strength of Agreement
.00–.20	Poor
.21–.40	Fair
.41–.60	Moderate
.61–.80	Good
.81–1.00	Very good

[*] Altman (1991).

TABLE 6

Intrarater Reliability for UK/I Samples: Kappas and Strength of Agreement Post 1 to Post 2*

	Post 1 – Post 2	
Parameter	**Kappa**	**Strength of Agreement***
Intelligibility	.84	Very good
Voice	.79	Good
Hypernasality	.84	Very good
Hyponasality	.70	Good
Audible nasal emission	.70	Good
Nasal turbulence	.77	Good
Grimace	.62	Good
Anterior oral CSCs	.60	Moderate
Posterior oral CSCs	.81	Very good
Nonoral CSCs	.84	Very good
Passive CSCs	.81	Very good
Noncleft speech errors	.62	Good
Need for intervention	.69	Good

*Altman (1991).

TABLE 7

Interrater Reliability for NA Samples: Kappas, Strength of Agreement, and Mean % Agreement Rating 1 and Rating 2

	Rating 1			Rating 2		
Parameter	**Kappa**	**Strength of Agreement**[*]	**M % Agreement**	**Kappa**	**Strength of Agreement**	**M % Agreement**
Intelligibility	.29	Fair	64	.28	Fair	48
Voice	.10	Poor	71	.18	Poor	69
Hypernasality	.71	Good	40	.70	Good	49
Hyponasality	.39	Fair	73	.67	Good	70
Audible nasal emission	.71	Good	71	.53	Moderate	69
Anterior oral CSCs	.45	Moderate	47	.38	Fair	49
Posterior oral CSCs	.06	Poor	79	.00	Poor	78
Nonoral CSCs	.78	Good	83	.60	Moderate	71
Passive CSCs	.55	Moderate	59	.72	Good	76
Recommend STx[†]	.26	Fair	81	.21	Fair	77
Recommend VPF Eval[†]	.69	Good	78	.63	Good	80

[*] Altman (1991).

[†] STx = speech intervention; VPF Eval = velopharyngeal function evaluation.

TABLE 8

Intrarater Reliability for the NA Samples: Kappas and Strength of Agreement

Parameter	Kappa	Strength of Agreement[*]
Intelligibility	.70	Good
Voice	.75	Good
Hypernasality	.85	Very good
Hyponasality	.62	Good
Audible nasal emission	.75	Good
Anterior oral CSCs	.81	Very good
Posterior oral CSCs	.46	Moderate
Nonoral CSCs	.85	Very good
Passive CSCs	.78	Good
Recommend STx[†]	.56	Moderate
Recommend VPF Eval[†]	.83	Very good

[*] Altman (1991).

[†] STx = speech intervention; VPF Eval = velopharyngeal function evaluation.

TABLE 9

Comparison of Interrater Reliability for UK/I (Post 1) and NA (Rating 1) Samples: Kappas, Strength of Agreement, and Mean % Agreement

	UK Post 1			NA Rating 1		
Parameter	Kappa	Strength of Agreement*	M % Agreement	Kappa	Strength of Agreement	M % Agreement
Intelligibility	.76	Good	48	.26	Fair	64
Voice	.66	Good	81	.10	Poor	71
Hypernasality	.70	Good	46	.71	Good	40
Hyponasality	.48	Moderate	80	.39	Fair	73
Audible nasal emission	.72	Good	63	.71	Good	71
Anterior oral CSCs	.47	Moderate	54	.45	Moderate	47
Posterior oral CSCs	.62	Good	64	.06	Poor	79
Nonoral CSCs	.65	Good	63	.78	Good	83
Passive CSCs	.75	Good	75	.55	Moderate	59
Need for intervention†	.57	Moderate	71	.26	Fair	81

*Altman (1991).

†This was titled “Recommend STx” in the NA data.

TABLE 10

Comparison of Intrarater Reliability for UK/I and NA Samples: Kappas and Strength of Agreement

	UK/I Samples		NA Samples	
Parameter	**Kappas**	**Strength of Agreement**[*]	**Kappas**	**Strength of Agreement**
Intelligibility	.84	Very good	.70	Good
Voice	.79	Good	.75	Good
Hypernasality	.84	Very good	.85	Very good
Hyponasality	.70	Good	.62	Good
Audible nasal emission	.70	Good	.75	Good
Anterior oral CSCs	.60	Moderate	.81	Very good
Posterior oral CSCs	.81	Very good	.46	Moderate
Nonoral CSCs	.84	Very good	.85	Very good
Passive CSCs	.81	Very good	.78	Good
Need for intervention[†]	.69	Good	.56	Moderate

[*] Altman (1991).

[†] This was titled “Recommend STx” in the NA data.