University of Alberta

Development and Evaluation of the Speech Intelligibility Probe for Children with Cleft Palate Version 5 (SIP-CCLP Ver. 5)

by

Carrie Lynne Gotzke

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Rehabilitation Science

Faculty of Rehabilitation Medicine

Fall 2012
Edmonton, Alberta

Abstract

The *Speech Intelligibility Probe for Children with Cleft Palate* (*SIP-CCLP)* is a computer-mediated word imitation measure of intelligibility that targets the speech error patterns of English-speaking children with cleft palate. Previous evaluations provided support for *SIP-CCLP* as a reliable and valid measure of speech intelligibility for children with cleft palate but revisions were recommended to improve its sensitivity, efficiency, utility and rigour for research and clinical applications. This thesis describes the construction and validation of *SIP-CCLP Ver. 5* as a discriminative health status measure of intelligibility for children with cleft palate.

Audio recordings of *SIP-CCLP* form 1 and 2 words, conversational speech and imitated sentences were obtained from 21 children with cleft palate, ranging in age from 37 to 84 months. Fourteen children completed a second set of *SIP-CCLP* recordings. Recordings were played back to listeners (i.e., 108 university students) who completed four word identification tasks independently. The percentage of words identified correctly served as the intelligibility score for each task. The 100-word conversational speech sample was transcribed phonetically to determine percentage of consonants correct (PCC). Two expert speech-language pathologists rated each child's hypernasality and voice severity from a standard speech sample.

Reliability coefficients were greater than 0.9 for all evaluations (i.e., test-retest, alternate forms, inter-rater and intra-rater reliability; internal consistency), indicating that *SIP-CCLP* scores are stable when differentiating between

individuals over time, forms, listeners, and items. Validity was assessed by examining the relationships of *SIP-CCLP* scores to 1) scores from other measures of intelligibility and 2) measures of related constructs. *SIP-CCLP* scores were correlated positively with intelligibility scores obtained from a 100-word conversational sample and an imitated sentence task. *SIP-CCLP* scores were correlated positively with PCC and negatively with hypernasality ratings. Between 51% and 69% of the variance in *SIP-CCLP* scores was explained by these two predictors. Based on these results, *SIP-CCLP Ver. 5* meets or surpasses established criteria for reliability and validity as a discriminative measure of speech intelligibility. It fills an identified need for an efficient, child-specific measure of intelligibility, with established rigour for children with cleft palate as young as three years of age.

Acknowledgement

This work was funded by the Cleft Palate Foundation, Killam Trusts and University of Alberta.

I would first like to express my gratitude to my doctoral supervisor, Dr. Megan Hodge. Her model as a clinician, researcher and professional is one that I continue to try and emulate. I would also like to thank my supervisory and examining committee members for their guidance: Dr. Terrance Nearey, Dr. Karen Pollock, Dr. Eric Parent, Dr. Tim Bressmann and Dr. Kathy Chapman. I would also like to acknowledge Dr. Todd Rogers for his assistance with generalizability study design.

I am indebted to Cindy Humphrey at the Alberta Children's Hospital and Lynn Smith at Alberta Health Services who went above and beyond to help recruit child participants and Sheila Ennis at the Glenrose Rehabilitation Hospital who shared her "ear" for voice and resonance and her clinical database of children's recordings. Thank you.

I would also like to thank the children and their families for participating. It was a please and an honor to make ice cream sundaes with all of you. I would also like to express my gratitude to the many university students who served as listeners and the raters and speech-language pathologists who shared their expertise.

Lastly, I would like to thank my husband Joe for his unwavering love, support and encouragement during my studies and my son Sean for his smiles that have brightened each and every day.

Table of Contents

Chapter 1

Chapter 2

Chapter 4

List of Tables

List of Figures

Chapter 1

Introduction

Overview

This work describes the development and evaluation of the *Speech Intelligibility Probe for Children with Cleft Palate Version 5* (*SIP-CCLP Ver. 5*) as a *discriminative* health status measure of intelligibility for young English-speaking children with cleft palate. Health status measures provide information about individuals or groups at a single point in time (Greenhalgh, Long, Brettle & Grant, 1998; Kirshner & Guyatt, 1985). The aim of a discriminative health status measure is "to distinguish between individuals or groups on an underlying dimension when no external criterion or gold standard is available" (Kirshner & Guyatt, 1985; p. 27). This chapter provides context for the work by reviewing speech characteristics of children with cleft palate and outlining the process of test development and validation for discriminative health status measures. Chapter two describes the development of *SIP-CCLP Ver. 5* and includes an evaluation of its content-related validity by a group of experts. Chapter three reports the results of an evaluation of the reliability of *SIP-CCLP Ver. 5* based on classical test theory. Supplementary evaluations of the reliability of *Ver. 5* using generalizability and item response theory are included in Appendix E and F, respectively. Chapter four reports the results of an evaluation of the criterion and construct-related validity of *Ver. 5* and includes a comparison of the error types identified for the child participants with those expected based on the experts' evaluation. The results of an item analysis and subsequent identification of items

to be revised and deleted are reported in Appendix D. Chapter five reports the effects of repeated exposure to *SIP-CCLP* stimuli and presents guidelines for administering the *Ver. 5* closed-set response task to listeners. The final chapter discusses the findings in relation to the research questions and presents recommendations for scoring and interpreting *SIP-CCLP Ver. 5* results. It also addresses the limitations and strengths of the current work and provides directions for future research. This is followed by concluding statements about the outcomes of the dissertation and the suitability of *SIP-CCLP Ver. 5* as a discriminative health status measure of intelligibility for young children with cleft palate.

Statement of the Problem

Intelligibility can be defined as the degree to which an individual's spoken message is recovered by a listener (Kent, Weismer, Kent & Rosenbek, 1989). Measures of intelligibility reflect the cumulative impact of a talker's articulation, resonance, voice and prosody on the speech signal. Intelligibility corresponds to the level of communicative activity (i.e., execution of a task) in the International Classification of Functioning, Disability and Health (ICF) framework (WHO, 2003). Clinically, intelligibility is often estimated through the use of equal-appearing interval scales (Whitehill, 2002). However, estimation of intelligibility using rating scales has a number of limitations. Their validity for measuring intelligibility is questionable as listeners are unable to divide intelligibility into equal intervals (Schiavetti, 1992). Other speech characteristics, such as hypernasality and nasal air emission, may make it difficult for listeners to focus

only on intelligibility when rating (Konst, Weersink-Braks, Rietveld & Peters, 2000; Witzel, 1995). Rating scales provide a single score describing intelligibility but they do not provide insight into the speech characteristics underlying the intelligibility deficit (Kent et al., 1989). Finally, rating scales provide scores at an ordinal scale of measurement, which limits their analysis to nonparametric statistics.

Methods of measuring intelligibility in which unfamiliar listeners orthographically transcribe a speech sample word-by-word (i.e., open-set response task) and/or identify a spoken word from a word list (i.e., closed-set response task) are considered to be valid, allow listeners to focus on intelligibility and provide ratio level data (Schiavetti, 1992). However, they provide little information about what is underlying the intelligibility deficit unless they are developed using a phonetic contrast approach to intelligibility assessment (Boothroyd, 1985; Kent et al., 1989). In this approach, unfamiliar listeners identify a spoken word from a set of choices that are minimally contrastive word pairs selected to target the speech error patterns for the population of interest. This approach allows the test user to assess the impact of a speaker's articulation errors and abnormal resonance and voice quality on intelligibility scores.

Characteristics of the listener, such as familiarity with the speaker, the speech characteristics of the disordered population, and the test stimuli, are recognized as variables that may affect a speaker's intelligibility scores (Walshe, Miller, Leahy & Miller, 2008). These variables may also affect the reliability of measures used to obtain intelligibility scores. For example, use of a single

listener who is familiar with the speaker and the test stimuli may yield higher scores than using a panel of listeners who are unfamiliar with the speaker. Furthermore, depending on the purpose of measurement, characteristics of the listener or listener group may affect the validity of the scores obtained. For example, if the purpose of measurement is to obtain an index of severity that describes the intelligibility of the speaker in his/her home environment, using a listener familiar with the speaker to obtain an intelligibility score appears justified. However, if the purpose of measurement is to describe intelligibility in other environments with unfamiliar listeners (e.g., store clerks), controlling listener familiarity with the speaker and the test stimuli are important to obtaining a valid measure.

The *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP)* is a computer-mediated measure of single-word intelligibility that uses a phonetic contrast approach to target the speech error patterns of children with cleft palate (Connolly, 2001; Feltz, McClure & O'Hare, 2002; Gotzke, 2003; Hodge & Gotzke, 2007). Gotzke (2005) evaluated the reliability and validity (construct and criterion) of *SIP-CCLP* with 30 children with and without cleft palate (15 children per group). Results showed acceptable inter-rater reliability and validity for the preliminary procedures applied and revisions were recommended to improve its sensitivity to the speech error patterns of children with cleft palate, efficiency by reducing the number of test items and utility for research and clinical situations that require alternate forms of an intelligibility

measure. This dissertation undertook to address these recommendations and to provide a rigorous evaluation of the revised version using test theory.

Speech of Children with Cleft Palate

Based on a review of 34 articles that described speech outcomes for children with cleft palate after surgery, Lohmander (2011) concluded that "good speech" could be expected for 50-60% of three-year-olds and 60-70% of four to five-year-olds and "normal speech" could be expected for 70-80% of six to eight-year-olds. Good speech was defined by absence of articulation errors, hypernasality and nasal emission and *good* intelligibility. Chapman and Willadsen (2011) identified variables that influence speech outcomes for children with and without cleft palate. For children with cleft palate, cleft type and severity, treatment (e.g., type and timing of surgery, availability and quality of interdisciplinary treatment) and remaining structural issues post-surgical closure (e.g., presence of fistulas, competence of the velopharyngeal mechanism) may all affect speech outcomes. In addition, speech outcomes for children with cleft palate are influenced by variables that they share with all children: co-morbidities (e.g., family history, syndrome), hearing status (e.g., frequency of otitis media, hearing loss) and treatment, access to and quality of speech therapy, and child and family characteristics.

Intelligibility. The intelligibility of speakers with cleft palate has been found to be lower than the intelligibility of age-similar speakers without craniofacial abnormalities. Konst et al. (2000) compared the intelligibility scores of 20 children with cleft palate and eight children without cleft palate at age two-

and-a-half years. Scores were obtained by having unfamiliar listeners orthographically transcribe 10 spontaneous utterances and then rate intelligibility for the same sample using a 10-point scale in which "1" was described as "unintelligible" and "10" was described as "intelligible." The percentage of words identified correctly and intelligibility ratings were significantly lower for the children with cleft palate. Merrick, Kunjur, Watts and Markus (2007) also reported a significant difference in intelligibility ratings obtained using a four-point descriptive scale for 50 children with cleft palate and 50 age and gender-matched controls, ranging in age from three to sixteen years. Gotzke (2005) compared the intelligibility of 15 children with cleft palate and 15 children without cleft palate or a speech sound disorder[1] on the *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP)* and on a 100-word spontaneous speech sample. The percentage of words identified correctly served as the intelligibility scores for both the *SIP-CCLP* and the spontaneous sample. Intelligibility scores were significantly lower for the 15 children with cleft palate than for the 15 age-similar children without cleft palate for both samples. Other studies, including Prins and Bloomer (1968), Phillips and Harrison (1969), and Zajac, Plante, Lloyd and Haley (2011), have reported lower intelligibility scores for children with cleft palate compared to children without cleft palate.

[1] Articulation of the children without cleft palate in Gotzke (2005) was screened using standard assessment subtests (e.g., *Fluharty-2*; Fluharty, 2001). All children scored at or above the 16th percentile on the screen. The articulation skills of the speakers without craniofacial abnormalities in Konst et al. (2000) and Merrick et al. (2007) were not described.

Relationship to articulation, resonance and voice. McWilliams (1954) found a significant positive correlation ($r = .72$, $p < .01$) between intelligibility, determined by orthographic transcription, and articulation scores. McWilliams concluded that as the number of articulation errors increases, intelligibility decreases. Zajac et al. (2011) also found a significant positive correlation ($r = .78$, $p < .001$) between mean intelligibility scores, determined by orthographic transcription of single words by five listeners, and a measure of articulation (i.e., percentage of consonants correct, as determined by phonetic transcription of the twelve words in the Preschool Screening form of the *Assessment of Phonological Processes-Revised* (Hodson, 1986)) for 21 children with cleft lip and palate ranging in age from five years to nine years, five months.

Keuning, Wieneke, van Wijngaarden and Dejonckere (2002) examined the relationship between nasality and intelligibility for 43 speakers with cleft palate. Nasality was assessed by visual analog scale and an acoustic measure of nasal resonance ("nasalance"). Intelligibility was evaluated using a visual analog scale. Anchors on both scales were defined as "normal" (left side) and "extremely deviant" (right side). The perceptual ratings of nasality and intelligibility had positive correlations for material containing nasals ($r = .63$) and for samples without nasals ($r = .60$), such that samples rated as more deviant in terms of hypernasality were also rated as more deviant in terms of intelligibility. Although significant positive correlations were also found between intelligibility ratings and nasalance scores for both types of stimuli, these were lower (with nasals: $r = .37$; without nasals: $r = .34$) than those reported for the perceptual rating of nasality.

Whitehill and Chun (2002) examined the relationships among intelligibility, nasality and articulatory accuracy for 20 Cantonese-speaking children with repaired cleft palate ranging in age from 5 years, 1 month to 15 years, 4 months. Intelligibility, defined as the percentage of words identified correctly, was based on children's single word productions and evaluated using a closed-set minimal pair response task. Nasality of oral and nasal-loaded sentences was rated using seven-point equal interval scales, where seven represented unacceptable speech or severe nasality. Articulatory accuracy was described by the percentage of phonemes transcribed phonetically as correct in the single-word speech sample. A significant correlation was found between intelligibility and articulatory accuracy scores ($r = .77$, $p < .01$) but the relationship between nasality and intelligibility was not significant (oral sentences: $r = -.38$; nasal sentences: $r = -.41$).

Two studies were identified that included voice when examining the relationships of speech characteristics to intelligibility for children with cleft palate. Moller and Starr (1984) investigated the relationships of resonance, articulation and voice to intelligibility for 100 speakers with cleft palate who ranged in age from 2 to 42 years. Listeners judged a sample consisting of conversational speech, imitated sentences, reading, counting and sustained vowels. Intelligibility was determined using percent estimation, while resonance, voice and articulation were rated on an eight-point scale, where "0" indicated no distortion/deviation and "7" indicated severe distortion/deviation. Significant negative correlations were found between intelligibility and resonance ($r = -.49$, p

< .01) and between intelligibility and articulation (r = -.89, p < .01). The relationship between intelligibility and voice was not significant (r = -.21). Konst, Rietveld, Peters and Weersink-Braks (2003) used multiple regression to examine the predictive relationship of nine speech variables (i.e., palatalization, lateralization of /s/, backing, glottal articulation, hyperkinetic voice, hypernasality, nasal emission, nasal fricative and correctness of articulation) to intelligibility. Ratings were obtained for 15 utterance samples of spontaneous speech from 20 children with cleft palate and 8 age-similar peers without cleft palate between 2 years, 5 months and 2 years, 7 months. Each variable was rated using a seven-point equal-appearing interval scale. For intelligibility and correctness of articulation, a higher score corresponded to "better speech" (p. 599), while for all other variables, a higher score corresponded to "more disordered speech." Correctness of articulation, lateralization and backing accounted for 93% of the variation in intelligibility.

Summary. Whitehill (2002) reviewed 57 articles that included a measure of intelligibility for speakers with cleft palate. Intelligibility was used to describe outcomes following surgery, speech therapy or orthodontic treatment in 22 studies and to provide a measure of severity in 28 studies. Reliability of intelligibility measures was reported in 30 of the 57 papers. Reports of inter-rater or intra-rater reliability, use of a consensus model, or previous reports of reliability by members of the same team were used as evidence of reliability. The majority of studies used a rating scale to measure intelligibility, a method whose validity has been questioned (e.g., Schiavetti, 1992). Whitehill (2002) raised concerns about the

state of intelligibility measurement in speakers with cleft palate including the reliability and validity of measures used to evaluate intelligibility and the limited amount of research focused on understanding intelligibility in this population.

There is a lack of word identification measures of speech intelligibility that are appropriate for preschool and early school-age children during the period when between 20 and 50% of children with cleft palate are expected to have some degree of speech intelligibility deficit (Lohmander, 2011). The commercially available *Children's Speech Intelligibility Measure (CSIM)* (Wilcox & Morris, 1999) was developed to provide an "objective measure of single-word intelligibility of children ages 3 years to 10 years, 11 months whose speech is considered unintelligible" (Wilcox & Morris, 1999; p. 1) to monitor progress during treatment. However, many of the *CSIM* test words are not expected to be in the vocabulary of young children, as "nearly half of the words" (Wilcox & Morris, 1999; p. 17) were not found in Hall, Nagy and Linn's (1984) published list of the most frequently used words in the spontaneous speech of children ranging in age from four years, six months to five years. No picture context is provided for the word stimuli to be imitated. In the speech intelligibility test for children with repaired cleft lip and palate developed by Zajac et al. (2011), appropriateness of the stimulus words for children was considered with most of the words (85%) taken from Hall et al.'s list (1984). However, this measure has not been used with children with cleft palate younger than five years of age. Furthermore, a phonetic contrast approach was not used to select the word stimuli. The *Test of Children's Speech* or *TOCS+* (Hodge & Gotzke, 2010) is a set of

word identification measures designed to measure the intelligibility of young children's imitative word and sentence productions. The items in the word format were selected using a phonetic contrast approach and target speech error patterns of individuals with dysarthria (Hodge & Gotzke, 2011; Kent et al., 1989).

Test Development

Test development refers to the process of developing items and combining them to form an instrument that measures some aspect of an individual's skills, abilities, interest, attitudes or knowledge (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999). The process of test development has been defined as a series of steps or phases, beginning with a statement of purpose and identification of the scope of the construct and ending with the development of guidelines for administering, scoring and interpreting test scores (AERA, APA, & NCME, 1999; Crocker & Algina, 1986).

Prior to developing items, the test developer establishes the test specifications. These specifications describe the test content, how items will be structured, how responses will be formatted (e.g., number of response alternatives) and how the test will be administered and scored (AERA, APA & NCME, 1999). Test specifications should also include definitions for the desired psychometric properties of the items. These specifications are used to develop an item pool. Items in the pool are reviewed for fit with the test specifications (AERA, APA & NCME, 1999; Crocker & Algina, 1986). Items that pass the review are pilot tested to obtain information on item quality (e.g., difficulty and

discrimination) with participants who are representative of the target population for whom the test is intended. Finally, items are assembled into a test, and procedures for administering, scoring and interpreting results are developed. Throughout test development, documentation of all methods is very important (AERA, APA & NCME, 1999).

Purpose of testing. When constructing health status and outcome measures, development and evaluation criteria are defined by the desired purpose for which the measure is being designed (Kirshner & Guyatt, 1985). Health status measures provide information about individuals or groups at a single point in time and may serve either a discriminative or predictive purpose (Greenhalgh et al., 1998; Kirshner & Guyatt, 1985). As noted in the Overview, the aim of a discriminative health status measure is "to distinguish between individuals or groups on an underlying dimension when no external criterion or gold standard is available" (Kirshner & Guyatt, 1985; p. 27). The purpose of intelligibility testing may be discriminative if the goal of measurement is to describe differences in severity of a speech disorder among a group of speakers with the same underlying condition. The purpose of a predictive health status measure is "to classify individuals into a set of predefined measurement categories when a gold standard is available, either concurrently or prospectively, to determine whether individuals have been classified correctly" (Kirshner & Guyatt, 1985; p. 27). Hearing screening is a possible example of a predictive measure, as results are later confirmed through full audiological assessment.

Outcome measures provide information about differences within a patient over time and serve an evaluative purpose (Greenhalgh et al., 1998; Kirshner & Guyatt, 1985). Such evaluative measures aim to assess "the magnitude of longitudinal change in an individual or group on the dimension of interest" (Kirshner & Guyatt, 1985; p. 28). The Pre-Kindergarten version of the American Speech-Language-Hearing Association (ASHA) National Outcome Measurement System (as cited in Thomas-Stonell, McConney-Ellis, Oddson, Robertson & Rosenbaum, 2007, p. 75) is an example of an evaluative outcome measure used to quantify functional change in six communication areas following speech and/or language intervention. For speakers with cleft palate, intelligibility has been used to evaluate change following surgery (e.g., Maegawa, Sells & David, 1998), prosthetic intervention (e.g., Konst et al., 2003) and speech therapy (e.g., Prins & Bloomer, 1965).

Kirshner and Guyatt (1985) identified six steps in test construction and validation that may be approached differently, depending on the purpose for which the measure is being developed: item selection, item scaling, reliability, validity, item reduction and responsiveness. It is important to note that the requirements for developing and validating a test for one purpose may be counter to the requirements for another purpose. For example, ease of interpretation is the guiding criteria for scaling items in a discriminative instrument, while the ability to register change guides item scaling for evaluative instruments. Furthermore, relevancy of different steps in test construction and validation differs depending on the purpose (Scientific Advisory Committee of the Medical Outcomes Trust,

2002). For example, while assessment of responsivity is important in validating evaluative instruments, it is not relevant when validating discriminative and predictive instruments. As a result, the purpose for which a measure is developed and validated should be clearly described by test developers. If a measure is to be used for a different purpose than described, the test user must justify the new use, which may require re-evaluating items, as well as reliability, validity and responsivity of the measure for the new purpose (AERA, APA & NCME, 1999). The following discussion focuses on criteria to be used in construction and validation of a discriminative health status measure as this is the purpose of developing *SIP-CCLP Ver. 5*.

Item selection. Kirshner and Guyatt (1985) identified three criteria to be used when selecting items for a discriminative measure. The first criterion is that that selected items should focus on features that are influenced by the condition. For example, children with cleft palate may actively compensate for velopharyngeal dysfunction by changing the place of articulation of alveolar sounds to a more backed location (Chapman & Willadsen, 2011). Therefore, in constructing a discriminative single-word measure of intelligibility for young children with cleft palate, items that target alveolar sounds would be selected preferentially over items that target sounds produced at a more backed location (e.g., glottal). The other two criteria are that items should be applicable to all examinees and stable over time. Therefore, if a single-word intelligibility measure is designed for young children, words that are in the vocabulary of

children would be selected preferentially over words that are not (e.g., blue vs. violet).

Item scaling. Item scaling refers to the options available for examinees in responding to each item (Kirshner & Guyatt, 1985). For discriminative measures, test developers want to minimize variability between examinees in terms of how an item is interpreted by including items with short response sets. Therefore, items with two clearly defined responses (e.g., nasal air emission is present, nasal air emission is absent) are preferred over items with responses that may be interpreted differently by examinees (e.g., mild, mild-to-moderate, moderate, moderate-to-severe, severe).

Reliability. Reliability has been defined as "the degree to which an instrument is free from random error" (Scientific Advisory Committee of the Medical Outcomes Trust, 2002) and the "desired consistency of test scores" (Crocker & Algina, 1986). Reliability of a measure is the result of the interaction among the instrument, the population of examinees and the testing situation (Streiner & Norman, 2008). Kirshner and Guyatt's (1985) criterion for the reliability of a discriminative measure is that inter-examinee variation be large and stable, such that there are high correlations between scores obtained on the test over multiple testing occasions.

Validity. Test validation is a process in which evidence supporting the proposed score interpretation is accumulated. AERA, APA and NCME (1999) described five sources that can provide evidence of validity: test content (i.e., content-related validity), response processes, internal structure, relations to other

variables (i.e., construct and criterion-related validity), and consequences of testing. According to Kirshner and Guyatt (1985), test developers of a discriminative measure are most concerned with evidence based on correlations between test scores and variables that are related to the construct that the test is purported to measure. For example, test developers of a discriminative measure of intelligibility for children would examine the relationships of articulation, resonance and voice to scores obtained from the measure. Discriminative measures are developed to distinguish between individuals when no external criterion or gold standard is available; therefore, it is not possible to evaluate criterion-related validity. However, relationships between the discriminative measure and measures that are hypothesized to be the gold standard may be investigated.

Item reduction. Item reduction refers to the process used by test developers to identify the final set of items that will be included in a test. The test developer calculates and evaluates item parameters that describe how a sample of examinees responds to each item on the test (Crocker & Algina, 1986). If the purpose of a test is discriminative, items that discriminate between examinees with different degrees of ability on the construct would be retained, while others that do not discriminate would be deleted.

Responsiveness. Responsiveness is defined as "the power of the index to detect a difference when one is present" (Kirshner & Guyatt, 1985, p. 34). Responsiveness is of concern for evaluative outcome measures only.

Summary. The purpose of a test guides the development of test and item specifications such as how items will be formatted and scored (AERA, APA & NCME, 1999). It also guides how items will be selected for the final form and how reliability and validity of this form will be evaluated. The test developer is responsible for documenting how these evaluations were conducted and for describing details of the sample used. This information allows potential test users to assess the adequacy of the evidence supporting a test's reliability, validity and responsivity and the appropriateness of the test for their measurement situation. With each change to the measure, the test developer must reevaluate reliability and validity with samples of subjects from the population for whom the test is intended. As a result, the process of test development is often conducted over a series of studies.

Purpose

The overarching purpose of this dissertation was to develop and evaluate a discriminative health status measure of intelligibility for young English-speaking children with cleft palate (*SIP-CCLP Ver. 5*), following the steps summarized in the preceding section. The specific objectives were to 1) develop *SIP-CCLP Ver. 5* based on recommendations by Gotzke (2005) and guidelines for single word lists to be used in assessment of cleft palate speech ((European Collaboration in Craniofacial Anomalies (EUROCRAN), 2009; Sell, Harding & Grunwell, 1999); 2) evaluate the content-related validity of the error patterns in *Ver. 5*; 3) evaluate the test-retest, alternate forms, inter-rater and intra-rater reliability of *Ver. 5* scores; 4) evaluate the criterion and construct-related validity of *Ver. 5* scores; 5)

develop guidelines for administering *Ver. 5* to children and obtaining judgments from listeners by examining the effect of listener familiarity with speaker, test stimuli and listening task on *Ver. 5* scores; and 6) recommend guidelines for interpreting scores obtained using *SIP-CCLP Ver. 5*.

Figure 1-1 shows a flow chart of the project objectives and their components. The development of *SIP-CCLP Ver. 5,* including item selection (i.e., error patterns, stimulus words and forms), item scaling and software development, and the evaluation of content-related validity, is described in Chapter 2. The evaluation of *Ver. 5* as a discriminative measure of speech intelligibility is outlined in Chapters 3 (reliability) and 4 (criterion and construct-related validity), as well as in Appendix D (item reduction). The development of guidelines for administering *SIP-CCLP Ver. 5* to children and listeners and interpreting scores are described in Chapters 5 and 6. Chapter 6 also provides a general discussion of the results and a set of conclusions about the outcome of the dissertation. Readers may find it useful to refer to the figure as they navigate the dissertation document. Chapters 2, 3, 4 and 5 each address a separate objective and are presented as a set of "stand-alone" but related article manuscripts.

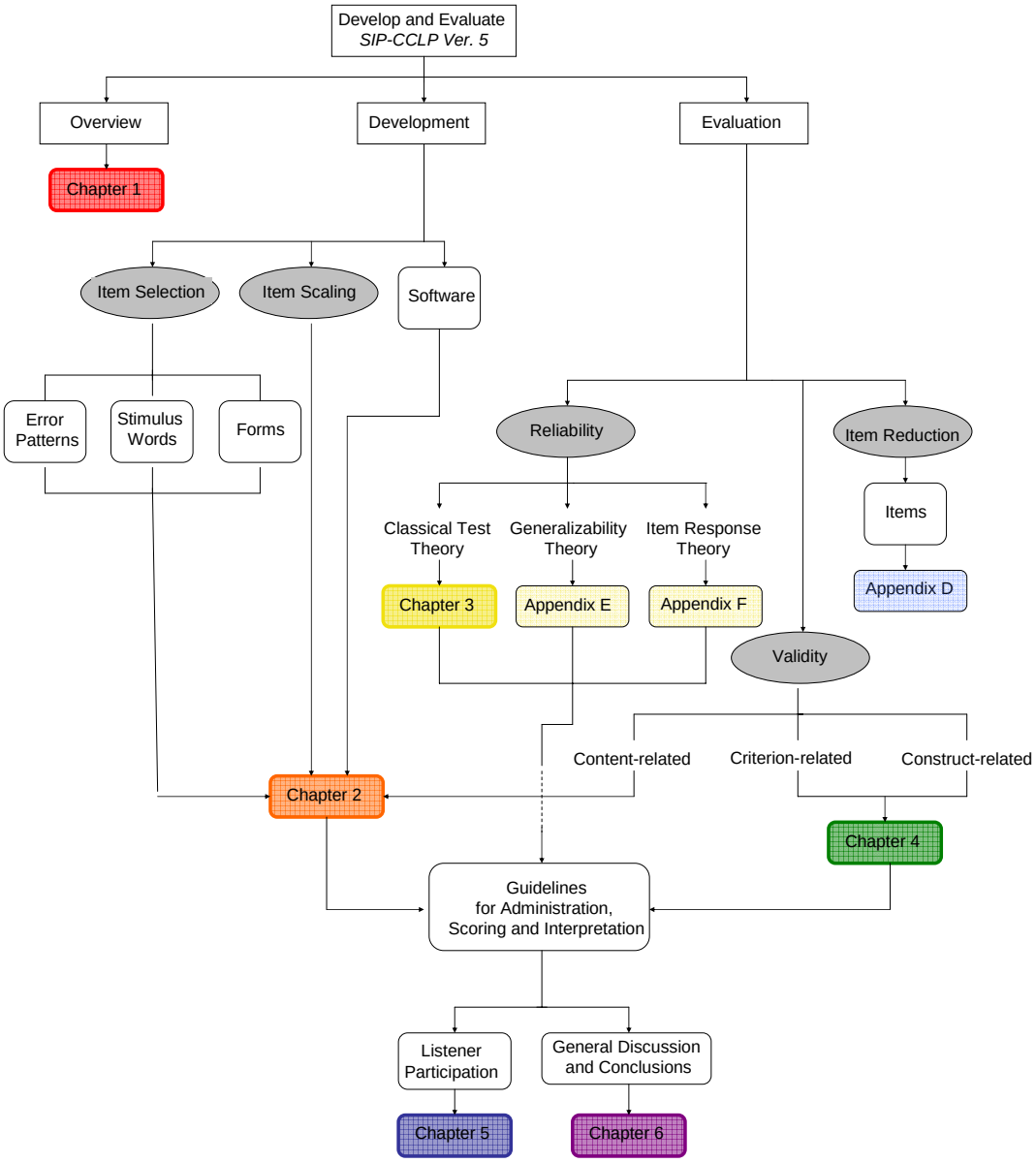

Figure 1-1. Flow chart outlining the relationships of the project objectives to the contents of each chapter.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Boothroyd, A. (1985). Evaluation of speech production of the hearing impaired: Some benefits of forced-choice testing. *Journal of Speech and Hearing Research, 28*, 185-196.

Chapman, K. L., & Willadsen, E. (2011). The development of speech in children with cleft palate. In S. Howard & A. Lohmander (Eds.), *Cleft Palate Speech: Assessment and Intervention* (pp. 23-40). West Sussex, UK: Wiley- Blackwell.

Connolly, S. (2001). *A phonetic contrast approach to assessing intelligibility in children with cleft palate.* (Unpublished master's project). University of Alberta, Edmonton, AB.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

European Collaboration in Craniofacial Anomalies. (2009). *Single word lists*. Retrieved from http://www.eurocran.org/content.asp?contentID=1387

Feltz, C., McClure, K., & O'Hare, J. (2002). *Speech intelligibility probe for children with cleft palate (SIP-CCLP): A preliminary assessment of validity and reliability*. (Unpublished master's project). University of Alberta, Edmonton, AB.

Fluharty, N. (2001). *Fluharty Preschool Speech and Language Screening Test (Second edition)*. Austin, TX, Pro-Ed.

Gotzke, C. L. (2005). *Speech intelligibility probe for children with cleft palate version 3: Assessment of reliability and validity*. Unpublished master's thesis). University of Alberta, Edmonton, AB.

Gotzke, C. L. (2003). *Continuing assessment of the validity and reliability of the speech intelligibility probe for children with cleft palate version 2*. (Unpublished master's project). University of Alberta, Edmonton, AB.

Greenhalgh, J., Long, A. F., Brettle, A. J., & Grant, M. J. (1998). Reviewing and selecting outcome measures for use in routine practice. *Journal of Evaluation in Clinical Practice, 4*(4), 339-350.

Hall, W. S., Nagy, W. E., & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hodge, M., & Gotzke, C. L. (2007). Preliminary results of an intelligibility measure for English-speaking children with cleft palate. *Cleft Palate-Craniofacial Journal, 44*(2), 163-174.

Hodge, M., & Gotzke, C. (2010). Stability of intelligibility measures for children with dysarthria and cerebral palsy. *Journal of Medical Speech Language Pathology, 18(4),* 61-65.

Hodge, M. M., & Gotzke, C. L. (2011). Minimal pair distinctions and intelligibility in preschool children with and without speech sound disorders. *Clinical Linguistics & Phonetics*: 25(10): 853-863.

Hodson, B. (1986). *The Assessment of Phonological Processes* (Revised Edition). Danville, IL: Interstate.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Keuning, K. H., Wieneke, G. H., Van Wijngaarden, H. A., & Dejonckere, P. H. (2002). The correlation between nasalance and a differentiated perceptual rating of speech in Dutch patients with velopharyngeal insufficiency. *Cleft Palate-Craniofacial Journal, 39(3)*, 277-284.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disorders, 38*(1), 27-36.

Konst, E. M., Rietveld, T., Peters, H. F. M., & Weersink-Braks, H. (2003). Use of a perceptual evaluation instrument to assess the effects of infant orthopedics on the speech of toddlers with cleft lip and palate. *Cleft Palate-Craniofacial Journal, 40*(6), 597-605.

Konst, E. M., Weersink-Braks, H., Rietveld, T., & Peters, H. (2000). An intelligibility assessment of toddlers with cleft lip and palate who received and did not receive presurgical infant orthopedic treatment. *Journal of Communication Disorders, 33,* 483- 501.

Lohmander, A. (2011). Surgical intervention and speech outcomes in cleft lip and palate. In S. Howard & A. Lohmander (Eds.), *Cleft Palate Speech: Assessment and Intervention* (pp. 55-86). West Sussex, UK: Wiley-Blackwell.

Maegawa, J., Sells, R .K., & David, D. J. (1998). Speech changes after maxillary advancement in 40 cleft lip and palate patients. *Journal of Craniofacial Surgery, 9*(2), 177-182.

McWilliams, B. J. (1954). Some factors in the intelligibility of cleft-palate speech. *Journal of Speech and Hearing Research, 19*, 524-527.

Merrick, G. D., Kunjur, J., Watts, R., & Markus, A. F. (2007). The effect of early insertion of grommets on the development of speech in children with cleft palates. *British Journal of Oral and Maxillofacial Surgery, 45*, 527-533.

Moller, K. T., & Starr, C. D. (1984). The effects of listening conditions on speech ratings obtained in a clinical setting. *Cleft Palate Journal*, *21(2)*, 65-69.

Phillips, B. J., & Harrison, R. J. (1969). Articulation patterns of preschool cleft palate children. *Cleft Palate Journal, 6*, 245-253.

Prins, D., & Bloomer, H. H. (1965). A word intelligibility approach to the study of speech change in oral cleft patients. *Cleft Palate Journal, 2*, 357-368.

Prins, D., & Bloomer, H. H. (1968). Consonant intelligibility: A procedure for evaluating speech in oral cleft subjects. *Journal of Speech and Hearing Research, 11*, 128-137.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11,* 193-205.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement and management* (pp. 119-155). Amsterdam, NL: John Benjamins.

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

Streiner, D.L., & Norman, G.R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford, UK: Oxford University Press.

Thomas-Stonell, N., McConney-Ellis, S., Oddson, B., Robertson, B., & Rosenbaum, P. (2007). An evaluation of the responsiveness of the pre-kindergarten ASHA NOMS. *Canadian Journal of Speech-Language Pathology and Audiology, 31(2)*, 74-82.

Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language & Communication Disorders, 43(6)*, 633-648.

Whitehill, T. (2002). Assessing intelligibility in speakers with cleft palate: A critical review of the literature. *Cleft Palate-Craniofacial Journal, 39(1)*, 50-58.

Whitehill, T., & Chun, J. C. (2002). Intelligibility and acceptability of speakers with cleft palate. In F. Windsor, M. L. Kelly, & N. Hewlett (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 405-415). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilcox, K., & Morris. S. (1999). *Children's Speech Intelligibility Measure*. San Antonio, TX: The Psychological Corporation.

Witzel, M. A. (1995). Communicative impairment associated with clefting. In R. J. Shprintzen & J. Bardach (Eds.), *Cleft palate speech management: A multidisciplinary approach* (pp. 137-166). St. Louis, MO: Mosby.

World Health Organization. (2003). *ICF checklist version 2.1a, Clinician form for international classification of functioning, disability and health*. Retrieved from http://www.who.int/classifications/icf/training/icfchecklist.pdf .

Zajac, D. J., Plante, C., Lloyd, A., & Haley, K. L. (2011). Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate. *Cleft Palate-Craniofacial Journal, 48(5)*, 538-549.

Chapter 2

Development of the *Speech Intelligibility Probe for Children with Cleft Palate Version 5 (SIP-CCLP Ver. 5)*

Rationale for *SIP-CCLP*

The *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP) Version 5* is a computer-mediated word imitation measure of speech intelligibility, developed as part of the *Test of Children's Speech Plus (TOCS+)* project (http://www.tocs.plus.ualberta.ca/). The *TOCS+* project developed a set of word identification tasks to measure the intelligibility of children's speech using word and sentence stimuli (e.g., *TOCS+* intelligibility measures; Hodge, Daniels & Gotzke, 2009) and supporting software for computer-mediated recording, playback and analysis. Like the *TOCS+* word intelligibility measure (Hodge et al., 2009), *SIP-CCLP* stimulus items were developed using a phonetic contrast approach to intelligibility assessment (see Kent, Weismer, Kent & Rosenbek, 1989). In this approach, words are selected that differ from other real words by a phoneme that differs in one or two articulatory features (i.e., manner, place, voicing). While the words in the *TOCS+* word intelligibility measure (Hodge et al., 2009) were selected to target the speech error patterns of individuals with dysarthria, the words in *SIP-CCLP* were selected to target the speech error patterns of children with cleft palate. Like its predecessors, *Children's Intelligibility Probe for Cleft Palate (CIP-CLP*) (Connolly, 2001), *SIP-CCLP Ver. 1* (Feltz et al., 2002), *SIP-CCLP Ver. 2* (Gotzke, 2003), *SIP-CCLP Ver. 3* (Gotzke, 2005) and *SIP-CCLP Ver. 4* (Gotzke & Hodge, 2008),

Version 5 was developed to assess the impact that speech error patterns associated with cleft palate have on intelligibility. *SIP-CCLP* software, based, in part, on the *TOCS+* software (Hodge et al., 2009), is used to elicit children's word productions in response to a pre-recorded verbal model and picture cue and record these directly to the computer as digital audio (.wav) files. In previous versions, listener judges also used the software to complete both open- (i.e., orthographic transcription) and closed-set (i.e., multiple choice) response tasks. Intelligibility scores are generated from the listeners' responses. In the open-set response task, a child's recordings are played to listeners who type in the word that they perceive the child to say. The percentage of words produced by the child that are identified correctly by listeners provides the intelligibility score for the open-set response task. In the closed-set response task, words that differ by one phoneme from the target word serve as the foil response options for each item. The percentage of items identified correctly serves as the intelligibility score. The closed-set task also provides information about the child's ability to make phonetic contrasts identifiable to listeners. *SIP-CCLP Ver. 5* uses a revised version of the closed-set response task only.

The following section reviews five considerations for a phonetic intelligibility measure that were outlined by Kent et al. (1989) and provide the rationale behind development of *SIP-CCLP*.

1. The measure provides a means to identify reasons for the intelligibility deficit. Whitehill and Chau (2004) developed a phonetic intelligibility measure for Cantonese speakers with cleft palate. In their group of 15 speakers with repaired

cleft lip and palate, they observed that "some speakers who had similar intelligibility scores had very different phonetic contrast error profiles" (p. 348), suggesting that the measure was able to provide unique information with respect to what error patterns were underlying the intelligibility deficit in each speaker. Like Kent et al. (1989), Whitehill and Chau (2004) suggested that this information could be used to guide intervention.

2. A phonetic intelligibility measure allows for analysis at different levels. Whitehill and Chau (2004) and Hodge and Gotzke (2007) analyzed the type of error contrast patterns identified for their speakers, but they could have also analyzed frequency of error patterns by sound (e.g., /t, s/), place of articulation (e.g., alveolar, velar) or manner of articulation (e.g., stops, fricatives and affricates). Gotzke (2005) used *SIP-CCLP Ver. 3* closed-set results for 15 children with and 25 children without cleft palate to analyze errors by manner, place and voicing. She found that the percentage of errors was significantly higher for the children with cleft palate on the phonetic contrast items targeting stops, fricatives, affricates, and liquids, alveolar place of articulation and voicing.

3. An intelligibility measure developed using a phonetic contrast approach should be sensitive to the "potential speech deficiencies" (p. 489) of the population of interest. For children with cleft palate, structural abnormalities of the speech mechanism remaining after surgical repair of the cleft, such as malocclusion, ectopic, missing or rotated teeth, malformations of the alveolar ridge, fistulas, and velopharyngeal dysfunction, may affect articulation and resonance, resulting in speech sound substitution, omission, addition, and distortion errors (Chapman,

1993). Whitehill and Chau (2004) cited two criteria for including phonetic contrast items that were sensitive to an error pattern in their intelligibility measure for Cantonese speakers with cleft palate: the error pattern was reported in at least two previous studies, and it was applicable to Cantonese.

4. The results of the intelligibility measure direct further acoustic and physiologic assessment of the talker. For example, if listeners identified initial consonant deletion, acoustic assessment could be used to determine if the consonant was actually deleted or if a glottal stop was used in its place. This could reveal more information about the articulatory nature of the perceived speech error and better inform the focus of treatment for the perceived error pattern.

5. The results of the phonetic intelligibility measure reflect the results of articulation testing because they use different approaches to assess speech sound production. Phonetic transcription of a speech sample to determine percentage of consonants correct (PCC) is one way to measure articulatory competence (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997). Gotzke (2005) compared mean percentage of consonants correct in the *SIP-CCLP* based on listeners' responses in the closed-set response task, phonetic transcription of the *SIP-CCLP* stimulus words and phonetic transcription of a 100-word spontaneous sample by a trained transcriber, and found no significant differences in PCC across the three conditions. This suggests that the phoneme identification results obtained using the closed-set response task confirmed those that would be obtained using phonetic transcription. However, Gotzke (2005) did not investigate if the specific

error patterns identified by listeners matched those identified through phonetic transcription.

History of *SIP-CCLP*

The phonetic content, format of the closed-set response task, and the software platform of *SIP-CCLP* were developed in several stages. In addition, initial evaluations of the reliability and validity of the later versions were conducted. These developments and evaluations are represented in sequential versions and are summarized in the next section to provide context for the development of *Version 5*.

Target error patterns. Connolly (2001) originally identified candidate error patterns for children with cleft palate through literature review and consultation with a speech-language pathologist with clinical experience with this population. She used this information to develop 160 phonetic contrast items. Feltz et al. (2002) added 11 phonetic contrast items targeting consonant clusters in *SIP-CCLP Ver. 1* based on Chapman (1993) who reported that percentage of occurrence for cluster simplification (e.g., /stov/ → /tov/) was significantly higher for three-year-old children with cleft palate compared to children without cleft palate. In *Ver. 2*, Gotzke (2003) added phonetic contrast items targeting substitution of stops for fricatives and affricates (e.g., target: "sail," foil: "tail"), as well as additional exemplars of patterns already tested in earlier versions, increasing the number of phonetic contrast items to 194. No additional phonetic contrast items were added in development of *Ver. 3*. However, Gotzke (2005) recommended phonetic contrast items targeting place and manner preference

errors be added to *Ver. 4* after examination of error patterns identified by listeners using the "blank" option in *Ver. 3* and comparison with the Cantonese measure of speech intelligibility for children with cleft palate developed by Whitehill and Chau (2004).

Closed-set response task. In *SIP-CCLP Ver. 1*, each phonetic contrast item was presented with four choices for judging: the target word, two foils and a "?" in the closed-set response task. One foil targeted the contrast of interest, while the other foil targeted a different feature than the one of interest. For example, for the error pattern in which alveolar stops are heard as alveolar nasals, the target word "bat" was contrasted with "mat" (i.e., targeted manner) and "pat" (i.e., targeted voicing not manner). The "?" option allowed listener judges to indicate that they were unable to identify the target and to reduce their likelihood of guessing.

In *Ver. 2*, Gotzke (2003) revised the options in the closed-set response task. In some *SIP-CCLP Ver. 1* phonetic contrast items, both foils targeted *SIP-CCLP* error patterns. For example, in the item described previously, the first foil (i.e., "mat") targets a manner preference error, while the second foil (i.e., "pat") targets a voicing error. In *Versions 2 and 3*, contrasts items were balanced in that for the two words in a contrast pair, each word was a target in one item and a foil in a second item. Weismer (2008) stated that this was an important feature of phonetic intelligibility measures because it did not bias the measure to certain error patterns. Gotzke (2003) revised the response options in the closed-set response task to simplify interpretation of responses and ensure that each phonetic

contrast item tested a single error pattern. Listener response options in *Ver. 2* are the target (e.g., "bat"), its foil (e.g., "mat"), a "blank" and "can't identify." If the sound or word identified differed from the options provided, listeners were instructed to use the "blank" button to type in what is heard. This option allowed listeners to capture instances of sound substitutions, omissions and additions for which minimal pairs had not been developed. The "can't identify" button replaced the "?" option to be used by listeners when they are unable to identify the target as an English sound (Hodge & Gotzke, 2007).

Confidence and distortion ratings. In the earliest versions, confidence ratings were included to increase the sensitivity of the *SIP-CCLP* to distortions and non-English substitutions that characterize the speech of children with cleft palate. Listener judges were asked to rate their certainty that the word chosen was the word actually said. Connolly (2001) used a three-point confidence rating system where "1" was described as "very certain," "2" was described as "somewhat confident" and "3" was described as "you do not know what was said" or "not confident." Feltz et al. (2002) revised this to a two-point confidence rating where "1" was described as "very confident" and "2" was described as "somewhat confident."

Gotzke (2003) revised the confidence ratings to distortion ratings in *Ver. 2.* In *Ver. 2*, listeners rated the child's production of the sound as "clear" or "distorted" after they identified the sound by choosing one of the minimally contrastive foils or typing in a word in the "blank" option. No distortion rating is made if the "can't identify" option is chosen. Gotzke (2003) anticipated that

"changing the focus of the rating system would better capture the nonstandard substitutions and distortions that characterize cleft palate speech" (p. 25).

No changes were made to the distortion rating system by Gotzke (2005) in the development of *Ver. 3*. In *Ver. 4*, Gotzke and Hodge (2008) used the distortion ratings to capture all three types of errors characterizing the speech of children with cleft palate (i.e., substitutions, omissions and distortions) in a single "phonetic accuracy score." Listener responses were recoded such that "correct/clear" responses were given two points and "correct/distorted" responses were given one point. All other responses, including substitutions and omissions were given zero points. The sum of the points divided by the number of possible points (i.e., number of contrast items X 2) provided a phonetic accuracy score for each listener's set of responses.

Software platform. In earlier versions, pictures representing the target words were presented on cards (Connolly, 2001) or as a power point show (Feltz et al., 2002) with the examiner providing the model of the target word for the child to imitate. Children's word productions were audio-recorded and then converted to digital audio files. These digital audio files were used to create listening tapes for presentation of the stimulus words to listeners. For the open-set response task, listeners wrote down what they heard. For the closed-set response, listeners marked which choice best matched what they heard and rated their confidence in their choice on the response form provided. Listener response forms were scored by hand and the percentage of words identified correctly (i.e., intelligibility score) was calculated.

Software was developed using *Macromedia Authorware 6* (Macromedia, Inc., 2001) to administer *Ver. 2*. At the beginning of administration to a child, the software generated a unique presentation order of the target words, a unique test identifier and folders in which all files associated with the child were saved. Child productions were elicited in response to a picture cue displayed on the computer screen and an auditory model produced by the examiner, and recorded directly to computer as digital audio (.wav) files using a sampling rate of 48 kHz and 16 bit quantization. The *TOCS+* Universal Sound Server (Young, Hodge & Daniels, 2004) was embedded in the software to set the recording and playback levels of the computer. A short beep and the appearance of a frog in the top left corner of the screen cued the child when it was time to repeat the word. Musical animations that appeared after every twenty items provided short breaks during administration. Other software features included the option to stop testing and return to it at a later time and to redo the practice items to help the child understand the task. The judging tasks were also administered using the software. For the open-set response task, listeners typed the word that was heard, as the audio file for each item was presented. For evaluation of intra-rater agreement, the software randomly selected 12 words for repeated presentation in the open-set response task. For the closed-set response task, the software generated a unique presentation order for each listener. Listeners were shown four choices ("buttons") on the computer screen as the audio file for the item was presented (i.e., target, its foil, "blank" and "can't identify"). If the target, its foil or "blank" was chosen, the listener also rated the sound production as "clear" or "distorted."

Listener response forms were saved to computer as text files. Like earlier versions, responses were scored by hand. However, the computerized format allowed error patterns to be determined with greater efficiency, as closed-set response forms from multiple judges could be easily combined into a single Excel spreadsheet. In *Ver. 3,* the software was revised to randomly choose 20 phonetic contrast items to be repeated in the closed-set response task for evaluation of intra-rater agreement. No other changes were made to the recording or open-set judging component of the *Ver. 3* software.

To create *Ver. 4*, the recording and judging components of the software were revised and an analysis component was added. In the recording component, models of the target words were recorded from a male speaker of Western Canadian English and embedded in the software. Test options were added to allow the user to choose whether the pictures, pre-recorded models, animations and "beep" to cue responses are turned "on" or "off" during administration. The option of playing instructions to the child prior to administration was added. The instructions contain pictures and spoken models illustrating the task, as well as opportunities to practice the task. The judging component of the software was changed to allow: 1) randomization of the playback order of the child's recordings for each listener (open-set only); 2) creation of separate Microsoft Excel (.xls) files for listener's responses for the test items and the items repeated for intra-rater agreement; and 3) automatic calculation of the total number correct (open-set only). An analysis component was added in creation of the *Ver. 4* software. The analysis component compiled and analyzed data for a child from three

listener's responses on the closed-set response task. Prior to analyzing the listeners' responses, the software first checked the listener's response files for entries in the "blank" response option and then, presented these for the examiner to verify or recode as the target or foil word. After the check is completed, the software generates an analysis file (.xls) and saves it in the child's folder. The analysis file compares the listeners' responses on the phonetic contrast items and summarizes the results. Three scores are presented: mean percentage of phonetic contrast items correct that are assigned a clear rating, mean percentage of phonetic contrast items that are assigned a distorted rating, and phonetic accuracy score. The analysis file also summarizes the number of phonetic contrast items correct and incorrect. A phonetic contrast item is described as correct if a minimum of two of the three listeners identified the contrast correctly and incorrect if a minimum of two of the three listeners chose the foil or did not agree on the response.

Evaluation of reliability and validity. Gotzke (2005) evaluated the reliability and validity of *SIP-CCLP Ver.* 3 using results from 15 children with cleft palate and 25 children with typical speech development and no history of craniofacial abnormalities. This subject pool included children recruited by Feltz et al. (2002) and Gotzke (2003). Gotzke (2005) concluded that, based on the results of this evaluation, the *SIP-CCLP* showed promise as a reliable and valid measure of speech intelligibility for children with cleft palate but recognized that revisions were needed to improve the sensitivity, efficiency and utility of *SIP-CCLP* for research and clinical applications. Gotzke (2005) made the following

recommendations: a) reduce the number of words elicited and items judged to decrease the length of the task for both children and listeners (i.e., increase efficiency); b) develop two forms to reduce child and listener familiarity with the stimulus words; and c) increase the number of response options to six to reduce the chance probability of listeners choosing the target and increase the sensitivity of the closed-set response task to the error patterns of children with cleft palate. Gotzke (2005) reported the results of a two-part item analysis that was conducted to address the first recommendation using empirical results obtained from 45 children (i.e., children from Feltz et al., 2002 and Gotzke, 2005). Gotzke (2005) first determined the phonetic contrast items in which a minimum of 90% of the listeners chose the target (115 items). Any items identified in step one but identified as errors for the children with cleft palate (i.e., two of the three listeners chose the foil or typed in a response in the "blank" which was judged to be representative of an error pattern unique to children with cleft palate) were then returned to the stimulus set. By applying this procedure, Gotzke (2005) excluded 94 phonetic contrast items from the closed-set response task. This reduced the number of word stimuli to be elicited from 124 to 80 and the number of phonetic contrast items to be judged from 194 to 108. However, at the time, Gotzke (2005) did not undertake a prospective evaluation of the revised content of *Ver. 4* or address the other two recommendations.

Development of *SIP-CCLP Version 5*

The remainder of this chapter describes the development of a substantially revised version of *SIP-CCLP* (*Ver. 5*) based on recommendations by Gotzke (2005) and guidelines for single-word lists to be used in assessment of cleft palate speech (European Collaboration in Craniofacial Anomalies (EUROCRAN), 2009; Sell, Harding & Grunwell, 1999). This includes the results of a three-part evaluation of content relevance and representativeness (i.e., content validity) of the error patterns in V*er. 5*. Kirshner and Guyatt (1985) stated that development of health status and outcome measures is guided by whether the purpose is discriminative, predictive or evaluative. *SIP-CCLP Ver. 5* was developed to be a discriminative measure of speech intelligibility that provides an index of severity of speech disorder in English-speaking children with cleft palate. The following description of *SIP-CCLP Ver. 5* is organized by the error patterns and phonetic content represented, the addition of a second form, results of several analyses of the vocabulary content in the two forms with respect to suitability for young children, analyses of word frequency and neighbourhood density of the multiple choice options in the closed-set response task, and revisions to the software.

Content validity of error patterns. Six different types of error patterns found in the speech of children with cleft palate are targeted in *SIP-CCLP Ver. 3* and *4*: manner preference errors (MPE), place preference errors (PPE), sibilant errors (SE), glottal errors (GE), voicing errors (VE) and cluster errors (CE). Hodge and Gotzke (2007) revised the description of place preference errors to include glottal errors, reducing the number of error categories to five. Within

these categories, error patterns were classified as cleft-related, developmental or unknown based on the extant literature. Cleft-related errors were associated with velopharyngeal dysfunction, structural differences at the alveolar ridge, fistulae, malocclusion or dental anomalies (e.g., ectopic teeth). Developmental errors were "the result of naturally occurring 'phonologic processes' that the child gradually eliminates as he matures" (p. 752; Peterson-Falzone, 1990). Unknown errors were those that could not be identified as cleft-related or developmental or were not reported during normal speech development. Testing both members of each phonetic contrast word pair provided opportunities for errors to occur in the unknown category.

The validity of the classification of error patterns as cleft-related, developmental and unknown was questioned during development of *Ver. 5*, because there is not a "gold standard" for attributing the cause of errors to cleft palate or phonologic processes that occur during development. For example, the cleft-related place preference error in which an oral stop is identified as a glottal stop (e.g., "pat" identified as "pa") may be an example of the developmental pattern final consonant deletion. Therefore, in *Ver. 5,* the classification of error patterns as cleft-related, developmental and unknown was eliminated. However, the objective of including error patterns that characterized the speech of children with cleft palate and speech disorders was retained. To meet this objective, the validity of the error patterns in *SIP-CCLP* was reexamined. Validity can be defined as "the degree to which the instrument measures what it purports to measure" (Scientific Advisory Committee of the Medical Outcomes Trust, 2002,

p. 200). Validity of the *SIP-CCLP* error patterns was assessed using a three part process. Direct empirical support was determined based on the number of children with cleft palate for whom each error pattern was identified by listeners (Gotzke, 2005). Next, a literature review was conducted to determine if published studies that reported phonological and/or phonetic analyses of the speech of children with cleft palate supported each of the error patterns tested in *SIP-CCLP*. After these two steps were completed, a panel of experts was recruited to assess content validity of the candidate error patterns.

Results from SIP-CCLP Ver. 3 and literature review by error category. This section describes details of the errors identified for the 15 children with cleft palate by type of error pattern reported by Gotzke (2005) and the supporting references identified in the literature review that describe the occurrence of error patterns for children with cleft palate, organized by the five main error categories. Appendix A provides a summary of this information.

1. *Manner preference errors*. A manner preference error occurs when the substituted sound is produced using a different manner of articulation than the target sound (Gotzke, 2005). Subtypes of manner preference errors are designated by a lowercase letter.

a. Substitution of sonorants (i.e., liquid, glide, nasal) for obstruents (i.e., stop, fricative, affricate). Listeners identified glide or liquid substitution for an obstruent for five children with cleft palate (six instances) (Gotzke, 2005). Stokes and Whitehill (1996) identified "gliding" errors, in which /s, l/ and the Cantonese palatal affricate, /ts/, were identified as /j/, based on phonetic transcription, in the

speech of four of seven Cantonese-speaking children with repaired cleft palate. In summarizing error patterns for children between four years, six months and seven years, six months, Harding and Grunwell (1996) listed /j/ as a cleft-type substitution for /s/ and as a non-cleft substitution for /t/ and /d/ and /w/ as a non-cleft substitution for /p/ and /f/. However, they did not describe any other glide/liquid substitution patterns for obstruents. Morris and Ozanne (2003) and Chapman (1993) listed the error pattern "liquid/glide replacement" in which a "liquid/glide is substituted for another consonant (e.g., [w∧n] for "sun")" (p. 470 of the former) in their descriptions of the phonological processes identified in the speech of children with cleft palate. Chapman (1993) reported a mean occurrence of this process of less than two instances in each of the three age groups (three-year-olds, four-year-olds and five-year-olds) for a sample of 30 children with cleft palate ranging from 3; 1 to 6; 0 years of age. Based on these findings, liquid/glide replacement of obstruents was included in *SIP-CCLP Ver. 5*.

Fifteen instances of substitution of a nasal for an oral stop were identified for six children with cleft palate by Gotzke (2005). This error pattern has been described in several studies of cleft palate speech (Harding & Grunwell, 1996; Prins & Bloomer, 1965; Chapman, 1993; Bzoch, 1965). Substitution of nasals for oral fricatives has also been reported (Harding & Grunwell, 1996). Two instances of this error pattern were identified for one child with cleft palate by Gotzke (2005). Based on these results, the error pattern of nasals for oral stops was expanded to nasals for obstruents.

Gotzke (2005) also recommended inclusion of the error pattern nasals for liquids. She identified this error pattern for two children with cleft palate (nine instances). This pattern was also described by Lynch, Fox and Brookshire (1983) and Prins and Bloomer (1965) for children with cleft palate. Based on these results, this error pattern was also included in *Ver. 5*.

b. Stopping of fricatives or affricates (e.g., "fail" identified as "bale"). Twenty instances of this error pattern were identified by Gotzke (2005) for seven of the children with cleft palate, justifying its inclusion in *Ver. 5*. Using the "blank" option, listeners also identified eight instances where an affricate was identified as a fricative for three children with cleft palate. As this error pattern has also been identified by Chapman (1993) and Morris and Ozanne (2003), it was included in *Ver. 5*.

c. Gliding of liquids. Gliding has been described as a developmental error pattern (Chapman, 1993; Morris & Ozanne, 2003). Using the type-in response afforded by the "blank" option, listeners identified 16 instances of liquid simplification in which /r, l/ targets are substituted with /w, j/, respectively, for five of the children with cleft palate (Gotzke, 2005). Based on these findings, gliding of liquids was included in *Ver. 5*.

d. Substitution of obstruents for liquids and glides. Gotzke (2005) reported one instance of an error in which a liquid was identified as a fricative (one child). This error pattern was described by Chapman and Hardin (1992) and Chapman (1993). Therefore, it was included in *Ver. 5*.

e. Substitution of oral stops for nasals. Gotzke (2005) reported two instances of errors in which nasals were identified as oral stops (one child). Substitution of oral stops for nasals would be expected if children with cleft palate were also hyponasal (Peterson-Falzone, Trost-Cardamone, Karnell & Hardin-Jones, 2006). As noted by Kummer (2008), hyponasality may be a complication of surgery for velopharyngeal dysfunction or of other craniofacial conditions (e.g., choanal stenosis) in speakers with cleft palate. Consequently, this error pattern was included in *Ver. 5*.

f. Substitution of affricates for oral stops. Gotzke (2005) reported two instances of errors in which oral stops were identified as affricates (one child). Chapman (1993) and Lynch, Fox and Brookshire (1983) described affrication in which affricates are substituted for stops. Chapman (1993) also described substitution of affricates for fricatives for three and four-year-old children with cleft palate. Gotzke (2005) identified three instances of this error for one child with cleft palate. Based on these findings, the error patterns substitution of affricates for oral stops and substitution of affricates for fricatives were included in *Ver. 5*.

2. *Place preference errors.* Gotzke (2005) described place preference errors as occurring when a sound is produced at a different place of articulation than the target sound. Support for place preference error subtypes is described in the following section.

a. Substitution of alveolar stops for bilabial stops. Listeners identified six instances of this error pattern for four of the children with cleft palate (Gotzke, 2005). This pattern has been described for children with cleft palate by Chapman (1993) and the Eurocleft Speech Group (1993). Based on these findings, this error pattern was included in *Ver. 5*.

b. Substitution of velar stops for alveolar stops. Listeners identified 18 instances of this error pattern for six of the children with cleft palate (Gotzke, 2005). Using the "blank" option, listeners' identified seven errors in which /k, g/ were substituted for /tʃ/ and /dʒ/ for one child with cleft palate. In *Ver. 3*, only backing of alveolar stops (i.e., /t, d/ produced as /k, g/) was targeted in the phonetic contrast items. The literature also provides examples of backing of sounds other than alveolar stops to velar place of articulation (i.e., substitution of /k/ and /g/ for /p, f, v/ and /s/) (Harding & Grunwell, 1996; Stokes & Whitehill, 1996). Ruiter, Korsten-Meijer and Goorhuis-Brouwer (2009) noted that both /t/ and /d/ were often substituted by /k/ in their group of children with cleft palate, suggesting that voicing may not be maintained in the substitution error. Based on these findings, phonetic contrast items targeting substitution of velar stops for fricatives and affricates were added to *Ver. 5*. Some examples of two feature errors in which both place of articulation and voicing differ in the target versus the foil were included (e.g., "dee" identified as "key").

c. Substitution of glottal stops for oral stops or affricates or glottal fricatives for oral stops. Four instances of glottal stop substitution for oral stops for three children with cleft palate were identified (Gotzke, 2005). These errors

may represent instances of consonant deletion not glottal stop substitution, as substitution is not verified through phonetic transcription. Listeners did not identify any instances of the other two patterns. However, these patterns have been described in other studies of the speech of children with cleft palate (e.g., Chapman, 1993; Harding & Grunwell, 1996) and so they were maintained in *Ver. 5*. Substitution of glottal stops and fricatives for glides, liquids and nasals has also been described (e.g., Lynch, Fox & Brookshire, 1983) but was not tested in *Ver. 3*. These error patterns were added to *Ver. 5*. Gotzke (2005) recommended adding new phonetic contrast items targeting the substitution of the glottal fricative for oral fricatives or affricates and the substitution of glottal stop for fricatives, as six instances of the former and four instances of the latter were identified for some of the children with cleft palate (one and three, respectively). These patterns have been described for other speakers with cleft palate (Morris & Ozanne, 2003; Stokes & Whitehill, 1996; Bzoch, 1965). Based on these findings, these error patterns were also included in *Ver. 5*.

d. Substitution of alveolar stops for velar stops. Listeners identified this type of error for six children with cleft palate (18 instances) (Gotzke, 2005). This error pattern has been described by Chapman (1993) and Harding and Grunwell (1996). Based on these findings, it was included in *Ver. 5*.

e. Substitution of alveolar stops for bilabial stops. Seven instances of this error pattern were identified for five children with cleft palate (Gotzke, 2005). Harding and Grunwell (1996) described this pattern as a "non-cleft realization" ("errors that are either developmental or idiosyncratic deviations unlikely to be

related to the cleft palate," p. 55). Lynch et al. (1983) also identified this error pattern in a speaker with cleft palate. Based on these findings, this error pattern was included in *Ver. 5*.

3. *Voicing errors.* Voicing errors occur when a voiced obstruent (e.g., /b/) is substituted for a voiceless obstruent (e.g., /p/) or vice versa.

a. Substitution of voiced sounds for voiceless sounds. Gotzke (2005) identified this error as a cleft-related pattern if the targeted sound was in word final position (e.g., "rope" identified as "robe") and as a developmental pattern if the targeted sound was in word initial or word medial position (e.g., "fail" identified as "veil"). One cleft-related voicing error was identified by listeners for the 15 children with cleft palate and 10 developmental voicing errors were identified for six children. Chapman (1993) described initial voicing of voiceless stops as a developmental pattern. These error patterns were included in *Ver. 5.*

b. Substitution of voiceless sounds for voiced sounds (e.g., "jeep" identified as "cheep"). Listeners identified 38 instances in which a voiceless sound was substituted for a voiced sound in word initial or word medial position (13 children) and seven instances of this error pattern in word final position (4 children) (Gotzke, 2005). Scherer, Williams and Proctor-Williams (2008) reported differences in the use of voiced and voiceless sounds in children with and without cleft palate at 30 months of age. Children were identified as using a speech sound if it was phonetically transcribed at least twice in a 50-utterance connected speech sample. Fewer children with unilateral cleft lip and palate used voiced stops and fricatives compared to the children without cleft palate.

Furthermore, for each place of articulation, the percentage of children with cleft palate using voiceless stops and fricatives was greater than the percentage using voiced stops and fricatives. Bzoch (1965) reported a tendency for three to five-year-old children with cleft palate to produce voiceless fricatives and affricates for voiced targets. Based on these findings, this pattern was included in *Ver. 5*.

4. *Sibilant errors.* Sibilant errors were described as a substitution error in which a sibilant is produced at a different place of articulation. Support for sibilant error subtypes is described in the following section.

a. Substitution of palatal sibilants for alveolar sibilants (i.e., palatalization). Listeners identified 15 instances in which a palatal fricative was substituted for alveolar sibilants for seven children with cleft palate (Gotzke, 2005). Palatalization has been described as a substitution error for children with cleft palate by Morris and Ozanne (2003) and as a common distortion for this population by the Eurocleft Speech Group (1993) and Albery and Grunwell (1993). Ruiter et al. (2009) noted that fronting or palatalization of /s/ was identified commonly in the speech of their group of children with cleft palate (mean age = 72 months). Ruiter et al. (2009) reported that /f/ and /s/ were most often nasalized or substituted by nasal friction for the toddlers with cleft palate in their study. Listeners identified four errors in which a fricative was identified as a fricative-nasal cluster (e.g., "sip" identified as "snip) for one child with cleft palate, which may be the result of nasal air emission or nasalization of the fricative (Hodge & Gotzke, 2007). Harding and Grunwell (1996) described this

pattern as a possible cleft-type error for /s/. Based on these findings, this error pattern was included in *Ver. 5*.

b. Substitution of labiodental fricatives for alveolar sibilants (i.e., weakening). Listeners identified four instances in which a labiodental was substituted for alveolar sibilants for three children with cleft palate (Gotzke, 2005). Weakening may be an example of the "other fronting" pattern described by Chapman (1993). Based on these findings, this error pattern was included in *Ver. 5*.

c. Substitution of alveolar sibilants for palatal sibilants or interdental fricatives for alveolar or palatal sibilants (fronting). Fourteen instances of these error patterns were identified for six of the children with cleft palate (Gotzke, 2005). Both palatal fronting and substitution for interdental fricatives have been described as developmental patterns (Chapman, 1993; Morris & Ozanne, 2003; Harding & Grunwell, 1996). Based on these findings, these error patterns were included in *Ver. 5*.

d. Substitution of alveolar sibilants for labiodental or interdental fricatives. Listeners identified nine instances of this error pattern for six of the children with cleft palate (Gotzke, 2005). Harding and Grunwell (1996) described substitution of /s/ for /f/ as a "non-cleft realization" (p. 348). Lynch et al. (1983) identified this error pattern in a speaker with cleft palate. Based on these findings, this error pattern was included in *Ver. 5*.

5. *Cluster errors.* Cluster errors were defined initially by Gotzke (2005) as "errors in which a consonant is deleted from a cluster" (p. 23). While Gotzke originally classified all occurrences of cluster reduction as developmental, Hodge and Gotzke (2007) reclassified errors in which the first obstruent was deleted from an obstruent-obstruent cluster as cleft-related (e.g., "stew" identified as "two") and errors in which the obstruent was deleted from an obstruent-sonorant cluster as developmental (e.g., "block" identified as "lock"). Deletion or weakening of an obstruent in an obstruent-obstruent cluster was suggested to be a possible consequence of velopharyngeal dysfunction. Two instances of this error pattern were identified for two children with cleft palate (Gotzke, 2005). Listeners identified six instances in which an obstruent was deleted from an obstruent-sonorant cluster for four children with cleft palate. This type of cluster error has been described by Ruiter et al. (2009) and Morris and Ozanne (2003). Deletion of a sonorant from an obstruent-sonorant cluster was also tested in *Ver. 3*, but no instances of this error pattern were identified for the children with cleft palate (Gotzke). Chapman (1993) provided a further example of a cleft-related cluster error in which a cluster was reduced to a velar consonant as a result of backing (e.g., "stove" identified as "cove"). While this error pattern was not identified for the 15 children with cleft palate evaluated by Gotzke (2005), it was included in *SIP-CCLP Ver. 5*, as were the other patterns described.

Expert assessment of SIP-CCLP Ver. 5 candidate error patterns. The third step in determining the phonetic error patterns to be included in *SIP-CCLP Ver. 5* was a review of the candidate error patterns by content experts to assess content relevance and representativeness. A common means of assessing content-related validity is to have items in a measure evaluated by a panel of experts familiar with the subject matter of the measure and/or the population for whom the measure is intended to be used (Crocker & Algina, 1986; Scientific Advisory Committee of the Medical Outcomes Trust, 2002). The results of the panel's assessment were used to verify that the candidate error patterns identified using *SIP-CCLP Ver. 3* results and in the review of the literature were representative of the speech error patterns of young children with cleft palate.

Content experts. Content experts were identified from the references that provided support for the error patterns targeted in *SIP-CCLP* and contacts in the speech-language pathology community. Eleven academics with expertise analyzing the speech of children with cleft palate were contacted by email and invited to participate. Nine experts agreed to participate and were sent a cover letter outlining the purpose of the study, information letter and consent form electronically. Once consent was received, instructions for completing the review process, a review form listing the error patterns assessed in *SIP-CCLP* and a questionnaire about the expert's background and experience were also sent electronically. These documents are provided in Appendix B. Eight experts completed the consent form and ratings task. One completed the ratings several

months after the other seven experts. The ninth rater was unable to complete the ratings due to personal circumstances.

All eight experts described their current academic position as a professor or lecturer. Seven experts indicated that they have held that position for over 15 years. Six experts indicated that they concurrently held a clinical position as a speech-language pathologist and had done so for at least five years. The number of years of experience that experts have had analyzing the speech of children with cleft palate ranged from 15 to 20 years (one expert) to over 30 years (three experts). All eight experts responded "yes" to the question "Do you consider yourself to have expert knowledge about the speech characteristics of children with cleft palate?" Six of the eight judges' experience was with English-speaking children with cleft palate. One judge's experience was with Swedish-speaking children and one judge's experience was with Cantonese-speaking children. All judges were fluent English speakers and writers.

Procedure. Experts were instructed to focus on children between three and seven years of age. They were instructed to identify which *SIP-CCLP Ver. 5* candidate error patterns occur rarely (i.e., <10%) in the speech of young children with cleft palate who also have a speech disorder and to provide comments about their ratings. Experts whose experience is with children who speak a language other than English were instructed to write "not applicable" beside error patterns that contained sounds not within the children's phonological system. Then, they were instructed to write down any error patterns that should be added to those listed to provide adequate representation of the speech error patterns of children

with cleft palate who also have a speech disorder. A limitation of this method of rating is that it required experts to rely on their recall to make a judgment about whether each error pattern occurs in more or less than 10% of the young children with cleft palate who also have a speech disorder.

Results. Experts were considered to have agreed on the rating of the error patterns if a minimum of five experts chose the same response. By this definition, agreement was obtained for 34 of the 39[2] error patterns. Agreement was not obtained for two manner preference errors (i.e., substitution of nasals for liquids, substitution of glottal stops for fricatives); two place preference errors (i.e., substitution of alveolar stops for velar stops, weakening); and one cluster error (i.e., deletion of an obstruent in an obstruent-sonorant cluster). To evaluate whether the rating system and/or description was contributing to the lack of agreement, experts were contacted and asked to rate these error patterns a second time. The description of the error patterns and the ratings were modified to clarify the task. After the second rating task, experts agreed that three of the five error patterns occur in fewer than 10% of children with cleft palate. Agreement was still not obtained for two error patterns (substitution of nasals for liquids and alveolar stops for velar stops). These two error patterns were retained in *Ver. 5*, as each error pattern was identified in the speech of at least one child with cleft palate in evaluation of *Ver. 3*.

[2] Experts rated 40 error patterns but the description of one pattern was incorrect (substitution of voiceless obstruents for voiced in final position should have been substitution of voiced obstruents for voiceless in final position). Therefore, ratings collected for this error pattern were not included in this analysis. All voicing error patterns were rated as occurring in less than 10% of children with cleft palate by at least 6 of the 8 judges.

Experts agreed that eleven error patterns occur in more than 10% of children with cleft palate who also have a speech disorder. Three of these were manner preference errors (stopping, gliding and substitution of nasals for obstruents); four were place preference errors (substitution of glottal stops for affricates, substitution of glottal stops for oral stops, substitution of glottal fricatives for oral fricatives, and substitution of velars for obstruents); two were sibilant errors (palatalization and fronting) and two were cluster errors (deletion of an obstruent in an obstruent-obstruent cluster and backing with cluster reduction). These 11 error patterns were retained in the *SIP-CCLP Ver. 5* pool.

Experts agreed that 26 patterns occur in less than 10% of children with cleft palate who also have a speech disorder. Comments on these 26 error patterns were reviewed. Experts noted some age-related differences for the occurrence of some patterns. For example, one expert noted that the pattern "substitution of fricatives for affricates" may be identified in the speech of three-year-old children but it is usually resolved by four or five years of age. Experts also noted that some patterns may be more common in the speech of children with non-cleft velopharyngeal dysfunction, children diagnosed with 22q11 deletion syndrome, or children with oromotor disorders than children with cleft palate (e.g., substitution of voiceless stops for voiced). Of the 26 patterns identified by listeners as rarely occurring, four were not identified for the 15 children with cleft palate who participated in the evaluation of *SIP-CCLP Ver. 3* (substitution of oral stops for liquids and glides, substitution of glottal fricatives for oral stops, deletion of a sonorant in an obstruent-sonorant cluster, and substitution of glottal

fricatives for sonorants). These four error patterns were deleted from the pool of error patterns to be tested in *SIP-CCLP Ver. 5*. The remaining 22 error patterns were retained in *SIP-CCLP Ver. 5* as each error pattern was identified in the speech of at least one child with cleft palate in evaluation of *Ver. 3*.

Experts recommended several examples and descriptions of error patterns to add to *SIP-CCLP Ver. 5* to provide adequate representation of the speech error patterns of children with cleft palate with a speech disorder. These included distortions (e.g., lateralization of sibilants), place errors (e.g., palatal dorsal production of lingual alveolars) and compensatory articulations (e.g., nasal fricatives). Some distortion errors described by the experts are included in the *SIP-CCLP Ver. 5* error patterns as substitution errors (e.g., palatalization of fricatives – "ship" identified as "sip"). Additional distortion errors can be captured in *SIP-CCLP Ver. 5* by listener ratings of the child's production of the target sound as "clear" or "distorted." The place errors and compensatory articulations described by experts cannot be captured using sounds in the American English phonological system. The response option "can't identify" in the closed-set response task provides a means to capture these errors by listeners. Two experts noted that the error patterns in which glottal stops are substituted for oral stops, fricatives or affricates may actually capture instances of initial or final consonant deletion. As noted previously, acoustic analysis may reveal which error pattern is present. No error patterns were added to *SIP-CCLP Ver. 5* on the basis of experts' descriptions.

Summary of error patterns included in SIP-CCLP Ver. 5. The outcome of the three part process for determining the error patterns to include in *SIP-CCLP Ver. 5* (review of the errors identified by listeners for the children with cleft palate in Gotzke (2005), review of the literature describing the speech error patterns of this population, expert evaluation of content-related validity of the candidate error patterns) provides support for the validity of the 35 error patterns included in *SIP-CCLP Ver. 5*. These include 11 error patterns in the manner preference category, 11 error patterns in the place preference category, six error patterns in the voicing category, four error patterns in the sibilant category and three error patterns in the cluster category. Appendix A lists the error patterns by category included in *Ver. 5*.

Addition of a second form. Previous versions of *SIP-CCLP* have one form. In research settings, listeners in intelligibility studies are often recruited from students enrolled at postsecondary institutions (e.g., Gotzke, 2005), resulting in a large pool of potential listeners. In clinical settings, however, the pool of available listeners may be limited. As a result, an individual may serve as a listener multiple times and may become familiar with or learn the items in an intelligibility measure with a limited number of stimulus items (Ziegler, Hartmann & von Cramon, 1988; Ziegler & Zierdt, 2008). Child participants may also learn the stimulus words if the test is administered multiple times. To avoid familiarization effects, Ziegler et al. (1988) suggested that measures of intelligibility should include more than one form.

Three approaches have been used to develop multiple forms. Yorkston and Beukelman (1980) used random selection of items from 50 pools of 12 phonologically similar items to create multiple forms. While this method permits the development of a large number of forms, evaluation of parallel forms reliability is challenging (Ziegler et al., 1988) and phonetic content may vary considerably among forms. A second approach is to create item sets specific to each of several forms (e.g., Monsen, Moog & Geers, 1988). This method allows the test developer control over the phonetic content of each form, allowing the same error patterns to be analyzed across forms (Ziegler et al., 1988). A third approach is to develop sets of equivalent items in an item bank from which items are selected for each administration of a test. Ziegler and Zierdt (2008) used an item banking approach in the development of the Munich Intelligibility Profile (MVP). This approach prevents listeners from becoming familiar with the stimulus items in a particular form, as the individual items in one administration of a test will be different from the items in another administration. It also increases the number of unique test forms that can be created from the same set of items. For example, if an item bank contained three sets each with two items, eight unique three-item test forms could be created. If each item was tied to a form, only two unique three-item test forms would be possible. While this approach has the advantage of creating multiple, equivalent and unique test forms, it also has disadvantages. A large number of items need to be developed, which is challenging when one is restricted to vocabulary appropriate for young children. In addition, items need to be evaluated with large samples to obtain stable item

parameters, which was not feasible for the scope of the current project. A computer database is also needed to store, retrieve and maintain items. Therefore, the second approach was used to create items sets specific to two alternate forms for *SIP-CCLP Ver. 5* that were controlled for content and could be analysed for the same error patterns.

Each form of *SIP-CCLP Ver. 5* has 63 stimulus items. Each form has the same four practice words: "fight", "peel", "rat," and "slow." Practice words are included to provide child participants with the opportunity to become familiar with the imitation and recording task. Each target sound appears in the same number of words in each form (e.g., form 1 and 2 each have two words targeting /f/). In the closed-set response task, the corresponding phonetic contrast items in form 1 and 2 target the same error patterns for all items but those targeting /f/ in final position, /θ/ in initial position and /st/ in CCV and CCVC words. It was not possible to identify real words which targeted the same error patterns for these items. The two forms have 12 words in common (robe, deer, fail, V, veil, pass, zee, zoo, chew, cheese, jeep, and badge), constituting 19% of the stimulus words in each form. Common items were necessary when it was not possible to identify different words with parallel syllable structure for the target sounds (e.g., only one word with consonant-vowel structure, targeting /v/, and appropriate for children was identified – the letter name "V") or words that targeted the same sound and error patterns (e.g., only one word targeting /ʤ/ and contrastive with /ʃ, ʧ, l/, and appropriate for children was identified - "jeep").

Considerations for selecting *SIP-CCLP Ver. 5* stimulus words. Two factors were considered in selecting *SIP-CCLP Ver. 5* stimulus words: 1) their appropriateness for sampling the speech of individuals with cleft palate and 2) their appropriateness for young children. The European Collaboration in Craniofacial Anomalies (EUROCRAN; 2009), among others, has developed guidelines to use when constructing single-word assessment measures for individuals with cleft palate. These guidelines concern the position and phonetic context of the target sound in the stimulus words, the number of times a target sound should be tested, and the test sound inventory. The second consideration in selecting *Ver. 5* stimulus words was their appropriateness for young children. Accuracy of word production may be affected by whether the words are present in children's expressive vocabulary and by how many other words are similar to the target word in their vocabulary (Vitevitch & Stamer, 2006). Therefore, these factors (i.e., age of acquisition and neighbourhood density) were also considered in the selection of word stimuli for *Ver. 5*. The following sections describe the EUROCRAN (2009) guidelines for sampling cleft palate speech, age of acquisition and neighbourhood density and their application to *Ver. 5* stimulus words. A list of the stimulus words in both forms is included in Appendix C.

EUROCRAN (2009) guidelines for phonetic content. Guidelines for the phonetic content of single-word lists that are to be used in the assessment of cleft palate speech characteristics have been developed (European Collaboration in Craniofacial Anomalies (EUROCRAN), 2009; Sell, Harding & Grunwell, 1999). These guidelines were endorsed by Henningsson, Kuehn, Sell, Sweeney, Trost-

Cardamone, and Whitehill (2008) in their description of speech sampling considerations for the "universal parameters for reporting speech outcomes in individuals with cleft palate." Each of five guidelines is described in the next section, accompanied by a comparison of *SIP-CCLP Ver. 5* stimulus words.

1. EUROCRAN (2009) recommended that when selecting stimulus items, words containing a single target pressure consonant (e.g., see) be preferred, while words containing nasal consonants and vowels of different height (in multisyllabic words) be avoided. If words with multiple consonants are selected (e.g., consonant-vowel-consonant (CVC) structure), it is recommended that words in which the other consonant is a glide or liquid (e.g., sell; Sell, Harding, & Grunwell, 1999) or a consonant with the same place of articulation as the target (e.g., cake; Lohmander, Willadsen, Persson, Henningsson, Bowden & Hutters, 2009) be selected preferentially over words in which the other consonant is an obstruent produced at a different place of articulation than the target. It is recommended that the target sound be in a linguistically stressed position in the word as sounds are "most distinctly articulated, most easily recognizable and minimally influenced by the phonetic content" (p. 349, Lohmander et al., 2009) in this position. Finally, because word initial position is often a stressed position across languages, Henningsson and Hutters (2004) (as cited in Henningsson et al., 2008, p. 11) recommended words with a consonant-vowel (CV) syllable structure be selected to allow easy comparison of cross-linguistic error patterns.

In the word stimuli for *SIP-CCLP Ver. 3,* all target consonants are represented in word-initial position in either a word with CV syllable structure or

with CVC structure in which the other consonant is a liquid or glide. In creating *Ver. 5* word stimuli, these types of words were maintained to facilitate use of the recorded *SIP-CCLP* stimulus words. It is important to note that the EUROCRAN (2009) speech sampling guidelines were developed to standardize assessment and promote cross-language comparison of cleft palate speech characteristics (e.g., articulation and resonance), not for developing a minimal pair word list to assess intelligibility. As the purpose of the *SIP-CCLP* is to provide a measure of severity of a child's speech intelligibility deficit, sampling words with a variety of syllable structures and degrees of phonetic complexity were included to provide additional information about the nature of the intelligibility deficit than could be acquired by limiting stimuli to simple word structures.

2. EUROCRAN (2009) recommended that target sounds be tested three times in "strong" position (i.e., "the position where the test sound is most distinctly articulated, most easily recognizable and minimally influenced by the context. This position usually implies that the consonant occurs in word- or syllable-initial stressed position") and two times in other positions, while Henningsson et al. (2008) recommended that a target consonant be tested a minimum of two times. In *SIP-CCLP Ver. 5*, the majority of target consonants are tested at least once in word-initial and word-final position. Eight target consonants do not follow this specification (i.e., /m, w, Θ, v, j, l, r, ŋ/) due to phonotactic constraints in English (e.g., /ŋ/ does not occur in word initial position) and difficulties identifying stimulus words appropriate for young children that targeted the consonant (e.g., /Θ/).

3. EUROCRAN (2009) recommended that all obstruents, liquids, glides and one or more nasals be assessed in a single word test. In *Ver. 5*, all consonants are included, except /ð/ and /ʒ/ ,as it was difficult to identify age-appropriate and/or picturable words containing these sounds, and /h/, as it was never identified as being in error in previous evaluations (Gotzke, 2005).

4. EUROCRAN (2009) suggested that all clusters relevant to the language should be tested when developing a specific speech assessment tool for speakers with cleft palate. In particular, clusters with oral non-pressure consonants and clusters with nasal consonants should be included because of different degrees of "loading" on the velopharyngeal mechanism. Clusters with oral non-pressure consonants (e.g., /sl/) were described as having minimum loading (i.e., velopharyngeal port remains closed throughout), whereas clusters with nasals (e.g., /sn/) were described as having maximum loading (i.e., velum moves from closed to open (for nasal) to closed (for vowel)). In a cross-linguistic examination of the occurrence of cleft-type speech characteristics, Grunwell et al. (2000) found that error scores were highest on sentences containing /s/, /sp/ and /sm/ in their sample of 131 children with unilateral cleft lip and palate. Furthermore, error scores were higher for sounds produced at the alveolar place of articulation (i.e., /t, d, nt, s/) than for sounds produced at other places of articulation. In the *Cleft Audit Protocol for Speech – Augmented (CAPS-A)*, John, Sell, Sweeney, Harding-Bell and Williams (2006) recommend that consonant production be assessed for /s/ clusters, specifically /st, skr/, and /sl/. Sounds produced at the alveolar place of articulation have been identified as most affected by cleft palate

in other studies (Gotzke, 2003; Harding & Grunwell, 1996). In *SIP-CCLP Ver. 5*, consonant clusters with sounds produced at the alveolar place of articulation (i.e., /st, sl, sn, str, sp, sk, tr, dr/) are targeted in word initial position. Final clusters are not targeted.

5. EUROCRAN (2009) and Henningsson et al. (2008) recommended that in development of assessment tools for speakers with cleft palate, a minimum of ten items should contain a high vowel (i.e., /i, I, u, ʊ/) to allow assessment of hypernasality. In vowel spectra, nasalization introduces an antiformant below F1 that reduces F1 amplitude and increases its bandwidth and centre frequency by 50 to 100 Hz (Pickett, 1999). Nasalization also introduces antiformants in the region of F2 and F3 that reduce their peak amplitudes and in some cases, flatten the spectral peaks (Pickett, 1999). Vowels are rarely identified as being in error for speakers with cleft palate (Peterson-Falzone, Hardin-Jones & Karnell, 2010). However, vowels may be less intelligible in the speech of children with cleft palate due to hypernasality and/or compensatory strategies involving tongue height and mouth opening adopted to reduce the perception of hypernasality. In S*IP-CCLP Ver. 3*, there were eleven items that targeted vowels, as a consequence of testing two error patterns: addition of oral stops or affricates before or after a vowel. These items were never identified as errors for the children with or without cleft palate by Gotzke (2005); therefore, they were not included in *Ver. 5*. However, the recommendation to include a minimum of ten items with a high vowel (i.e., /i, I, u, ʊ/) to allow assessment of the influence of hypernasality on

word intelligibility was followed in *Ver. 5*. Form 1 and 2 have a high vowel in 27 and 28 stimulus words, respectively.

Age of word acquisition. Measures that describe age of word acquisition may be used to determine the likelihood that a word is in the vocabulary of young children. Age of acquisition has been defined as the "month in which 50% of the children … were reported to comprehend or produce" a word (Goodman, Dale & Li, 2008; p. 521) or the age at which words are learned (Clark & Paivio, 2004). Age of acquisition has been determined from estimates based on parent report questionnaires that inventory young children's vocabulary development (e.g., Fenson, Marchman, Thal, Dale, Reznick & Bates, 2007), from frequency counts in children's dictionaries (Clark & Paivio, 2004), or from ratings completed by adults who rate the age at which they learned a word (Bird, Franklin & Howard, 2001). Frequency of usage counts based on young children's spontaneous speech is another way to determine the likelihood that a word may be present in a young child's lexicon (e.g., Stemach & Williams, 1988; Hall, Nagy & Linn, 1984; Kolson, 1960).

Appropriateness of the stimulus words in the two forms of *SIP-CCLP Ver. 5* was evaluated for young children by examining frequency usage counts of pre-kindergarten and/or first-grade children's spontaneous speech (Stemach & Williams, 1988; Kolson, 1960). Frequency of usage counts for younger children were not identified. Fourteen of the 114 stimulus words were not listed in the frequency usage counts reported by Stemach and Williams (1988), and Kolson (1960): the letter names "V", "zee", "G", "J", "K", "As", "Ks", and "fail", "veil",

"zap", "Lee", "bash", "Sue," or "spear." No letter names were listed in either database. Evans, Bell, Shaw, Moretti and Page (2006), showed a card listing all 26 uppercase letters in random order to Canadian kindergarten children (n = 149; average age: 5 years, 9 months) and asked the children to name each letter. The percentage of correct responses was 98.0% for "A", 83.2% for "G", 83.9% for "J", 89.3% for "K", 73.2% for "V", and 94.6% for "Z". Letter naming was also assessed for children in California by Treiman, Tincoff, Rodriguez, Mouzaki and Francis (1998) using the same method as Evans et al. (2006). The percentage of correct responses ranged from 79% for G to 97% for X for 38 five-year-old children and from 17% for D and 71% for O for 35 four-year-old children. Letter names were considered to be appropriate for *SIP-CCLP Ver. 5* because children are given both the visual (e.g., uppercase letter) and verbal model for each of the letters used as stimulus words. Of the remaining seven stimulus words not listed in Stemach and Williams (1988) or Kolson (1960), two are in both forms (fail, veil), three are in form 1 (spear, Lee, bash) and two are in form 2 (Sue, zap). A familiarization activity for these words was included in the *SIP-CCLP Ver. 5* software to introduce vocabulary to the child participants. In this activity, children are shown the picture, hear the pre-recorded model for each word, and repeat the name of the picture. These productions are recorded but are not judged by listeners. In a future study, these recordings will be used to examine if the intelligibility of children's productions of unfamiliar words changes from the first time they are produced to the second (as a stimulus word).

Neighbourhood density. Neighbourhood density is defined as "the absolute number of words occurring in any given similarity neighbourhood" (Goldinger, Luce & Pisoni, 1989; p. 502). Words are described as phonological neighbors if they differ by a single phoneme (Yates, 2009; Grainger, Muneaux, Farioli & Ziegler, 2005), as is the case with minimal pairs (e.g., sell and fell). Phonological neighbourhoods are composed of all the words that differ from the target by a single phoneme. For example, if *sell* is the target, *fell, soul* and *set* would all be included in its phonological neighbourhood. Research with adults has found that words with high density neighborhoods (i.e., many phonological neighbors) are produced more quickly and accurately than words with low density neighbourhoods (i.e., few phonological neighbours) (Vitevitch, 2002). Sosa and Stoel-Gammon (2012) found that neighbourhood density was a significant predictor of production variability and whole-word proximity (a measure for quantifying how close the child's production of a word is to the target) in spontaneous speech for 15 children with typical language development ranging in age from 2 to 2 years, 5 months. Production was more variable and proximity was lower (e.g., word was less similar to the target) for monosyllabic words from low density neighbourhoods than for words from high density neighbourhoods. While this result suggests that neighbourhood density may be a factor in accuracy of word production in young children, Metsala and Chisholm (2011) found that the relationship between production accuracy and neighbourhood density held only for three and four-syllable non-words. In their sample of 194 children

ranging in age from three to seven years, accuracy did not differ for two-syllable non-words from different density neighbourhoods.

As neighbourhood density may affect accuracy of word production for children, neighbourhood density measures for the *SIP-CCLP Ver. 5* stimulus words were obtained from the Child Corpus Calculator (Storkel & Hoover, 2010). This Calculator was constructed using data from Kolson (1960) and Moe, Hopkins and Rush (1982) for kindergarten and first-grade children's spontaneous speech. The number of neighbours is not significantly different on the two forms ($t(124) = .73$, $p = .465$) and ranges from 4 to 34 (Mean = 16.51, SD = 7.69) on form 1 and from 3 to 31 (Mean = 15.59, SD = 6.35) on form 2. The median number of neighbours is 16 for each form.

***SIP-CCLP Ver. 5* closed-set response task.**

Response options. In the *SIP-CCLP Ver. 3 and 4* closed-set judging task, the listener is instructed to select which of four choices best matches the sound(s) heard in the target position in a word. The four choices are the minimal pair contrast items (e.g., target "b" in "**b**at" and foil "p" in "**p**at"), a "blank" for the listener to identify the highlighted sound in the target position as a English sound that is different from those provided and "can't identify" if the listener is unable to identify the sound as an English phoneme.

Four response options facilitate easy interpretation of the responses to an item, but have a number of limitations. The chance probability of a correct response with a four-option closed-set response task is high ($p = .25$ with four real-word options; Ziegler & Zierdt, 2008). As a result, intelligibility scores

obtained using a four-option closed-set task are higher than those obtained using an open-set task or closed-set tasks with more response options. This can cause a ceiling effect in test scores for those with mild speech disorders. Yorkston and Beukelman (1980) examined the effect of increasing the number of response options on single-word intelligibility scores for a group of nine speakers with dysarthria. For each subject, listeners completed an open-set response task and closed-set response tasks with four, eight or twelve real-word choices of similar sounding words. An inverse relationship was found between the number of response options and intelligibility scores, such that as the number of response options increased the mean intelligibility scores decreased. Intelligibility scores were lowest for the open-set response task (i.e., infinite number of choices) and highest for the four-option closed-set response task. However, all formats resulted in similar rankings of speakers. Yorkston and Beukelman (1980) concluded that intelligibility scores obtained using the open-set response task are "probably a good indicator of functional level" (p. 21) but they are insensitive to differences in severely dysarthric speakers as the range of intelligibility scores obtained using the open-set response task was smaller than the range in scores obtained using a four-option closed-set response task. Therefore, intelligibility scores obtained using the four-option closed-set response task appears to be more sensitive to differences among severely dysarthric speakers.

Another limitation of previous versions of the *SIP-CCLP* closed-set response task is that some target stimulus words are presented multiple times to assess different phonetic contrasts. Listeners may hear and judge some target

stimulus word up to a maximum of four times (e.g., listeners hear "D" and are presented with the minimal pair contrast items: "D" – "E," "D" – "B," "D" – "knee," and "D" – "zee"). Consecutive presentations of the same stimulus words occurs rarely (software randomizes presentation order of items for each listener judging session) but is possible. Coté-Reschny (2007) found that when listeners heard the same stimulus word produced by children with dysarthria repeated consecutively, word identification scores increased on average by approximately 3% from the first to the third presentation. Miller, Heise and Lichten (1951) examined the effect of consecutive repetition on the percentage of words identified correctly for monosyllable words produced by a typical speaker and presented at different signal-to-noise ratios. "Slight" improvement in the percentage of words identified was noted at all signal-to-noise ratios (p. 335, percentage of increase was not reported).

Two additional real-word foil options were added to *SIP-CCLP Ver. 5* to reduce the chance probability of listeners choosing the target and the number of times listeners hear a stimulus word, and to increase the sensitivity and efficiency of the closed-set response task. Listeners choose from four minimally-contrastive words (one target (e.g., "peel") and three foils (e.g., "eel," "heel," and "wheel")), a "blank" to type-in a response different from the provided choices and "can't identify." Because of the challenges associated with identifying sets of minimally contrastive words while maintaining the likelihood that each word would be found in the vocabulary of young children, the number of options was increased by two (total of six). This allows three error patterns to be tested per phonetic

contrast item. This reduces respondent burden by decreasing the number of items judged in the closed-set judging task but may increase the time listeners take to choose which option matches what was heard. Real-word foils differ from the target word in only one consonant that is in the same position for all words (e.g., sail, tail, nail, fail). Thirty stimulus words serve as both targets and foils in form 1 and twenty-two stimulus words serve as both targets and foils in form 2. Proper names are used as foils in three phonetic contrast items in form 1 and six phonetic contrast items in form 2 to provide "real word" options. All three foils target one of the error patterns listed in Appendix A except for one foil for four stimulus words in form 1 (i.e., cow, go, trail, drip) and 2 (i.e., K, guy, trip, dry), as it was not possible to identify three real-word alternatives that targeted *SIP-CCLP Ver. 5* error patterns. In evaluation of *Ver. 5*, listener responses for these eight stimulus words were examined to determine if a different foil is needed for these phonetic contrast items (see Chapter 4).

Response options are presented on the screen in a rectangular matrix with two columns and three rows. The "blank" and "can't identify" choices are located in lower left and lower right positions. The four real-word alternatives are presented in random order to prevent listeners from learning the position of the target word (Ziegler & Zierdt, 2008). Figure 2-1 shows an example of a response screen for the *Ver. 5* closed-set judging task.

If listeners choose one of four real-word alternatives or type a response in the "blank," listeners rate the production of the underlined sound as "clear" or "distorted." In *Ver. 3* and *4*, listeners heard the child's production of the stimulus

word once before choosing one of the response options and then rating the child's production. Hodge and Gotzke (2007) found that the percentage of correct-distorted scores were more variable (and therefore, less reliable) than intelligibility scores across the groups of three listeners. In *Ver. 5*, listeners hear the child's production of the stimulus word once before choosing one of the response options and then choose whether they would like to hear the child's production a second time before selecting a clarity rating. This change was made to determine if it increased the reliability of phonetic accuracy scores among listeners (see Chapter 3).

Lexical variables. Frequency of occurrence of a word in general linguistic use (Howes, 1957; Rosenzweig & Postman, 1957) and neighbourhood density have also been found to affect listeners' ability to identify spoken words. These authors found that word lists constructed using highly familiar or frequently occurring words yielded a higher percentage of words identified correctly than lists constructed using unfamiliar words. Furthermore, Giolas and Epstein (1963) reported that the percentage of words identified correctly obtained using word lists containing highly or extremely familiar words were closest to the percentage of words identified correctly from continuous discourse. Building on previous research by Miller, Heise and Lichten (1951), Howes (1957) concluded that presenting words with a range of frequency of occurrence as alternatives in closed-set tasks "effectively increases the relative frequency of the [presented] words" (p. 302) and would yield similar scores for the percentage of words identified correctly as having a word list composed of all high frequency words

judged using an open-set task. Similarly, Pollack, Rubenstein and Decker (1959) found that word frequency had a "minimal" effect on the signal-to-noise level at which 50% of words were identified correctly when listeners were given a list of the words to be heard, but had a "strong" effect when listeners were not provided with information about the words (p. 275). Frequency of word occurrence has been determined through frequency counts of written material (e.g., Thorndike & Lorge, 1952). Yates (2009) reported that processing time for visual recognition tasks is faster for words with many neighbours than for words with few neighbours. Goldinger, Luce and Pisoni (1989) examined the effect of neighbourhood density on auditory recognition using an open-set response task. The percentage of words identified correctly was higher for words from low density neighborhoods than those from high density neighbourhoods.

In light of possible influences of word frequency and neighbourhood density on adult listeners' responses on the closed-set response task, these variables were analysed for the target and foils words in the two forms of *SIP-CCLP Ver. 5*. Word frequency was obtained from the SUBTLEXus database. It is based on 51 million words obtained from American English subtitles of television and movie scripts (Brysbaert & New, 2009). Two words were not found in the SUBTLEXus database: "As" (plural letter name) (form 1) and "Ks" (plural letter name) (form 2). Word frequency per million words for the stimulus words was similar on the two forms ($t(122) = -.254$, $p = .80$) and ranged from 1.12 to 3793.04 (Mean = 221.5, SD = 734.26) for form 1 and from 1.14 to 5971.55 (Mean = 259.03, SD = 902.69) for form 2. The median word frequency

was 19.92 for form 1 and 27.48 for form 2. Word frequency per million words for the 135 foil words in form 1 ranged from 0.12 to 41857.12 (Mean = 788.87, SD = 3808.11). Word frequency per million words was not found for two foil words in form 2 (i.e., "Kate" and "Nate"). Word frequency for the remaining 130 foil words in form 2 ranged from 0.12 to 41857.12 (Mean = 717.96, SD = 4098.87). The median word frequency was 24.55 for form 1 and 24.57 for form 2. The word frequency per million words was not significantly different on the two forms ($t(263) = .146$, $p = .884$).

Neighbourhood density was determined using the Irvine Phonotactic Online Dictionary (IPhOD) (Vaden, Hickok & Halpin, 2009). Number of neighbours for the stimulus words was not significantly different on the two forms ($t(122) = .624$, $p = .534$) and ranged from 10 to 56 (Mean = 33.1, SD = 11.15) for form 1 and from 8 to 50 (Mean = 31.89, SD = 10.53) for form 2. The median number of neighbours was 36 for form 1 and 33 for form 2. Number of neighbours for the foil words was not significantly different on the two forms ($t(265) = .494$, $p = .622$) and ranged from 12 to 56 (Mean = 35.6, SD = 9.0) for form 1 and from 11 to 56 (Mean = 35.03, SD = 9.34) for form 2. The median number of neighbours was 37 for form 1 and 36 for form 2.

There are no statistical differences in word frequency and phonological neighbourhood size between the two forms. However, other lexical characteristics such as word type (e.g., verb, noun, article) or orthographic neighbourhood (defined as the number of words that differ in a single letter) may affect the equivalency of the two forms. Re-evaluation of form equivalency with

respect to the lexical characteristics of the stimulus words may be required if the two forms are not found to be parallel (see Appendix D).

Software revision. *SIP-CCLP Ver. 4* software has three components: recording, judging and analysis. Each of these components was revised *in Ver. 5* by the programmer who created the *TOCS+* software and previous versions of the *SIP-CCLP* software. The revisions to each component are described in the following sections, followed by the results of pilot testing the software.

Recording. *SIP-CCLP* software was revised to allow the test user to choose form 1 or form 2 on the same screen that provides four other options to select "on" or "off" for administration: pictures, pre-recorded auditory models, animations (provide child with short breaks) and "beep" (cue for child to produce the target word). Recordings of the instructions played to the child at the beginning of the *SIP-CCLP* and all practice and stimulus words were obtained from a young adult male speaker of Western Canadian English with professional voice training. One of the two practice words embedded in the instructions was changed from "pizza" to "coat," as all *SIP-CCLP Ver. 5* stimulus items are single syllable words. After the software presents the task instructions to the child, the examiner chooses either "learn words" or "continue." As mentioned previously, the "learn words," feature was added to provide children with the opportunity to become familiar with the seven words not listed in the frequency usage counts reported by Stemach and Williams (1988) and Kolson (1960) (i.e., bash, fail, Lee, spear, Sue, veil, zap). If the user chooses this feature, the words are presented and the child's productions are recorded and saved as .wav files in the child's

participant folder in a subfolder created by the software. If the examiner chooses "continue" after the instructions, the four new practice words (i.e., "fight", "peel", "rat", and "slow") are presented in random order. Once the practice items have been recorded, the examiner has the option to redo them if the child's needs more practice to learn the task, or to continue with the 63 stimulus words. Original artwork was created and then enhanced using *Macromedia Fireworks 8* (Macromedia, Inc., 2007) for all new practice and stimulus words in *Ver. 5*.

Judging. *SIP-CCLP* software was revised to allow the test user to choose whether listeners judge form 1 or form 2. Several revisions were made to how the judging task is administered. The instructions were revised to reflect the increase in the number of real-word choices in the items and a new feature that allows listeners to hear the child's production a second time prior to rating the production of the underlined sound as "clear" or "distorted." The presentation screen was revised to allow presentation of the six response options in a three-row-by-two-column arrangement. The real-word response alternatives are presented in random order in the four uppermost buttons with the "blank" presented in the bottom leftmost button and "can't identify" presented in the bottom rightmost button, as shown in Figure 2-1. In *Ver. 3* and *4*, listeners saw the response options on the screen 0.5 seconds before they heard the audio recording of the child's word. In *Ver. 5*, listeners hear the recording just before seeing the response choices. Ziegler and Zierdt (2008) suggested that this method helps prevent listeners from "guessing" the stimulus word prior to hearing it. If the listener selects one of the four real-word options or types a response in the "blank," the

listener then rates the underlined sound as "clear" or "distorted." A button labeled "play it again" was added to the bottom of this screen to allow listeners to hear the child's word production a second time prior to rating the underlined sound as "clear" or "distorted." The software documents the number of times listeners hear each word (i.e., "1" or "2").

Analysis. *SIP-CCLP* analysis software was revised to reflect the error patterns tested in *Ver. 5* (see Appendix A). As described for *Ver. 4*, the software first checks the listeners' responses for entries typed in the blank response and presents them on the screen for recoding as the target or foil word. The test administrator recodes the entry if it contains the sound that occurs in the target or foil word(s) in the contrastive position. The test administrator selects "no change" if the entry does not contain the sound that occurs in the target or foil word(s) in the contrastive position. The software then merges responses from three listeners to calculate the child's intelligibility and phonetic accuracy scores and to determine the error patterns represented. It presents this information in Excel files. *SIP-CCLP* software determines the error pattern represented by the listener judges' choice using the following procedure. Items in which a minimum of two of the three listener judges chose the "target" are summarized in the "contrast targets correct" section of the Excel file. Information in this section is organized by the five error categories (manner preference, place preference, sibilant error, voicing error and cluster error). Phonetic contrast items in which a minimum of two of the three listener judges chose the same foil are summarized in the "contrast error profiling (foil)" section of the Excel file. These errors are

also organized by the five error categories. Phonetic contrast items in which no agreement is obtained among the three listener judges, two of the three listeners chose "can't identify," or two of the three listeners typed a response in the "blank" button are listed in the "contrast error profiling (other) section" of the Excel file. Tallies of the number of items in each of the three sections of the file are provided. The results in the analysis file provide a profile of the child's errors by type and frequency.

Pilot testing results. *SIP-CCLP Ver. 5* recordings for form 1 and 2 were collected from four preschool children without cleft palate and judged by 12 listeners. Children and listeners were recruited via convenience sampling. Log notes from the recording and judging sessions, records of comments from child participants about picture stimuli, and from listener participants about the closed-set response task were reviewed to identify additional software revisions. Two final changes to *Ver. 5* software were made based on this review: a) a text line was added below the form 1 and 2 buttons to indicate to the user which form was selected and b) the names of the listener files used to generate the analysis were listed at the top of the analysis Excel file.

The measures calculated by the software (i.e., intelligibility score, phonetic accuracy score) and error analysis were checked and found to be accurate. At this point, development of *SIP-CCLP Ver. 5* software was complete and ready for the next stage: evaluation of its reliability and validity.

Conclusions

Ver. 5 is the result of revisions to improve the efficiency, utility, and sensitivity of *SIP-CCLP* for research and clinical applications. To increase efficiency in *Ver. 5*, the number of words elicited and items judged were decreased, shortening the length of the task for both children (on average 10 minutes) and listeners (on average 10 minutes per listener). Analysis to generate an error profile based on three listeners' responses takes an additional 5-10 minutes. The evaluation of content-related validity using results from Gotzke (2005) and literature authored by recognized experts in the field, and expert assessment of the proposed *Ver. 5* error patterns decreased the number of phonetic contrasts targeted, by focusing on those error patterns that had been identified in the speech of children with cleft palate. This increased the efficiency of *Ver. 5* and its' sensitivity to the error patterns of children with cleft palate. To increase its utility, two forms were developed to reduce the impact of changing child and listener familiarity with the stimulus words if the test is administered multiple times.

Several revisions were made to increase the sensitivity of *Ver. 5* to the speech of young children with cleft palate. Age-appropriateness of the stimulus words for speakers with cleft palate and young children was considered. Guidelines for the phonetic content of single-word speech samples from individuals with cleft palate (e.g., EUROCRAN, 2009) were followed. Stemach and Williams (1988) and Kolson (1960) were used to determine if words are within the vocabulary of kindergarten and/or first grade children. Although all

but seven words were identified as occurring in the vocabulary of English-speaking children in kindergarten or first grade, it is not known if the words are within the vocabulary of younger children. The *SIP-CCLP Ver. 5* software was revised to include a familiarization activity at the beginning of the task to introduce the unfamiliar vocabulary to children. The stimulus words on the two forms were not different with respect to word frequency (obtained from the SUBTLEXus database (Brysbaert & New, 2009)) and neighbourhood density (obtained using the Child Corpus Calculator (Storkel & Hoover, 2010)). The number of response options was increased to six to reduce the chance probability of listeners choosing the target and thereby, increase the sensitivity of the closed-set response task to the error patterns of children with cleft palate. Lexical variables (i.e., word frequency and neighborhood density) for listeners were considered for the stimulus and foil words. Both forms were found to sample words from a wide range of frequencies and densities.

The next step in the development of *Ver. 5* was an evaluation of its reliability and validity, as a discriminative measure of speech intelligibility for young children with cleft palate. This is described in Chapters 3 and 4, respectively.

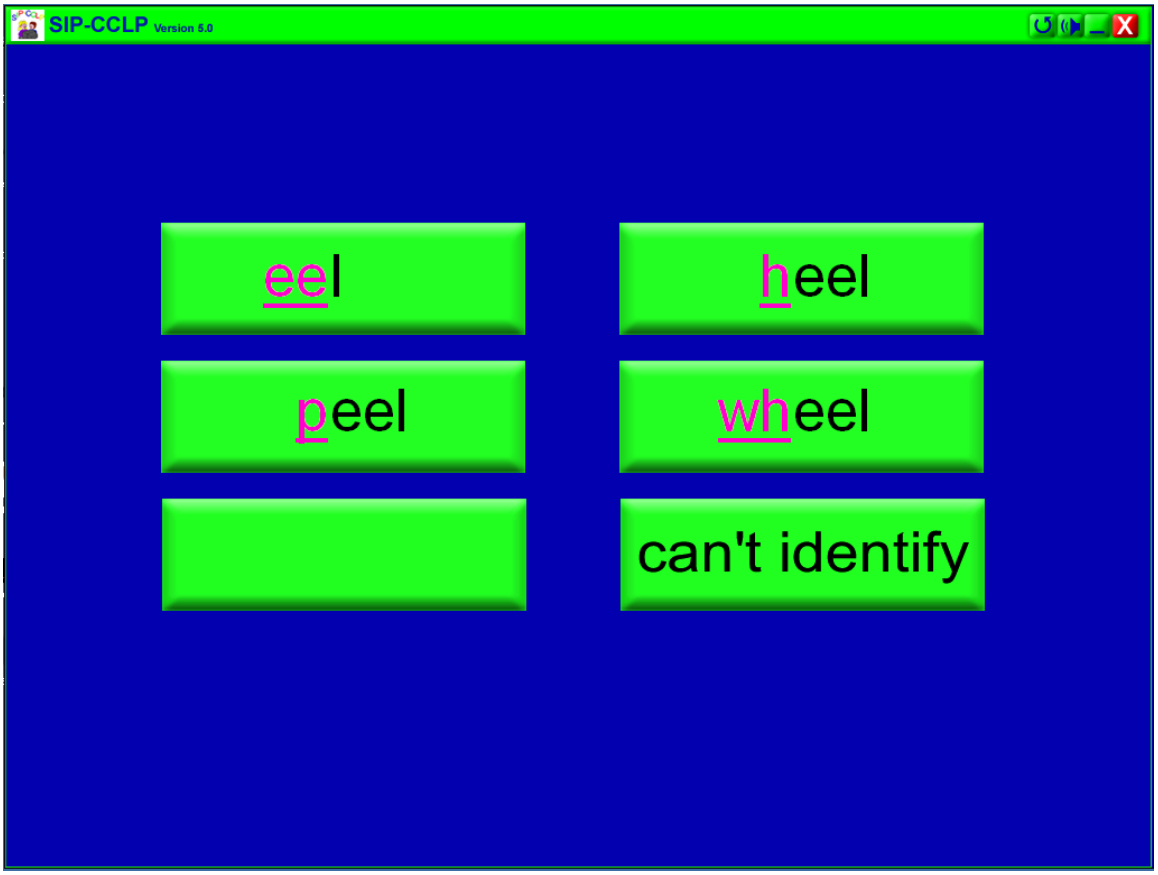

Figure 2-1. *SIP-CCLP Ver. 5* closed-set response task presentation screen.

References

Albery, E., & Grunwell, P. (1993). Consonant articulation in different types of cleft lip and palate. In P. Grunwell (Ed.), *Analysing cleft palate speech* (pp. 83-111). London, UK: Whurr Publishers Ltd.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments & Computers, 33(1)*, 73-79.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41(4)*, 977-990.

Bzoch, K. R. (1965). Articulation proficiency and error patterns of preschool cleft palate and normal children. *Cleft Palate Journal, 2,* 340-349

Chapman, K. L. (1993). Phonologic processes in children with cleft palate. *Cleft Palate-Craniofacial Journal, 30(1)*, 64-72.

Chapman, K. L., & Hardin, M. A. (1992). Phonetic and phonologic skills of two-year-olds with cleft palate. *Cleft Palate-Craniofacial Journal, 29(5)*, 435-443.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers, 36(3)*, 371-383.

Connolly, S. (2001). *A phonetic contrast approach to assessing intelligibility in children with cleft palate*. (Unpublished master's project). University of Alberta, Edmonton, AB.

Coté-Reschny, K. J. (2007). *Effects of talker severity and repeated presentations on listener judgments of the speech intelligibility of young children with dysarthria.* (Unpublished master's thesis). University of Alberta, Edmonton, AB.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Eurocleft Speech Group. (1993). Cleft palate speech in a European perspective: Eurocleft Speech Project. In P. Grunwell (Ed.), *Analysing cleft palate speech* (pp. 142-165). London, UK: Whurr Publishers Ltd.

European Collaboration in Craniofacial Anomalies. (2009). *Single word lists*. Retrieved from http://www.eurocran.org/content.asp?contentID=1387

Evans, M. A., Bell, M., Shaw, D., Moretti, S., & Page, J. (2006). Letter names, letter sounds and phonological awareness: An examination of kindergarten children and of letters across children. *Reading and Writing, 19*, 959-989.

Feltz, C., McClure, K., & O'Hare, J. (2002). *Speech intelligibility probe for children with cleft palate (SIP-CCLP): A preliminary assessment of validity and reliability*. (Unpublished master's project). University of Alberta, Edmonton, AB.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual, 2nd Ed*. Baltimore, MD: Paul H. Brookes.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6,* 349-358.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501-518.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language, 35*, 515-531.

Gotzke, C. L. (2003). *Continuing assessment of the validity and reliability of the speech intelligibility probe for children with cleft palate version 2*. (Unpublished master's project). University of Alberta, Edmonton, AB.

Gotzke, C. L. (2005). *Speech intelligibility probe for children with cleft palate version 3: Assessment of reliability and validity*. (Unpublished master's thesis). University of Alberta, Edmonton, AB.

Gotzke, C. L., & Hodge, M. (2008). S*peech intelligibility probe for children with cleft palate Version 4.0 (SIP-CCLP Ver. 4.0) user's guide*. (Unpublished manuscript). University of Alberta, Edmonton, AB.

Grainger, J., Muneaux, M., Farioli, F., & Ziegler, J. C. (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *The Quarterly Journal of Experimental Psychology, 58A(6)*, 981-998.

Grunwell, P., Brondsted, K., Henningsson, G., Jansonius, K., Karling, J., Meijer, M., . . . Sell, D. (2000). 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. *Scandinavian Journal of Plastic and Reconstructive Hand Surgery, 34*, 219-229.

Hall, W. S., Nagy, W. E., & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Harding, A., & Grunwell, P. (1996). Characteristics of cleft palate speech. *European Journal of Disorders of Communication, 31*, 331-357.

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

Hodge, M. M., Daniels, J., & Gotzke, C. L. (2009). *TOCS+ Intelligibility Measures* [computer software]. Edmonton, AB: University of Alberta.

Hodge, M., & Gotzke, C. L. (2007). Preliminary results of an intelligibility measure for English-speaking children with cleft palate. *Cleft Palate-Craniofacial Journal, 44*(2), 163-174.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrences of English words. *Journal of the Acoustical Society of America, 29*, 296-305.

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. *Cleft Palate-Craniofacial Journal, 43*(3), 272-288.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disorders, 38*(1), 27-36.

Kolson, C. J. (1960). *The Vocabulary of Kindergarten Children*. (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Kummer, A. (2008). *Cleft palate and craniofacial anomalies: Effects on speech and resonance* (2nd Ed.). Clifton Park, NY: Thomson Delmar Learning.

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. *Cleft Palate-Craniofacial Journal, 46*(4), 347-362.

Lynch, J. I., Fox, D. R., & Brookshire, B. L. (1983). Phonological proficiency of two cleft palate toddlers with school-age follow-up. *Journal of Speech and Hearing Disorders, 48*, 274-285.

Macromedia, Inc. (2001). *Macromedia Authorware 6* [computer software]. San Francisco, CA: Macromedia, Inc.

Macromedia, Inc. (2007). *Macromedia Fireworks 8* [computer software]. San Francisco, CA: Macromedia, Inc.

Metsala, J. L., & Chisholm, G. M. (2010). The influence of lexical status and neighborhood density on children's nonword repetition. *Applied Psycholinguistics, 31*, 489-506.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41(5)*, 329-350

Moe, A. J., Hopkins, K. J., & Rush, R. T. (1982). *The vocabulary of first grade children*. Springfield, IL: Thomas.

Monsen, R., Moog, J. S., & Geers, A. E. (1988). *CID Picture SPINE SPeech Intelligibility Evaluation*. St. Louis, MO: Central Institute for the Deaf.

Morris, H., & Ozanne, A. (2003). Phonetic, phonological and language skills of children with a cleft palate. *Cleft Palate-Craniofacial Journal, 40*(5), 460-470.

Peterson-Falzone, S. J. (1990). A cross-sectional analysis of speech results following palatal closure. In J. Bardach & H. L. Morris (Eds.), *Multidisciplinary management of cleft lip and palate* (pp. 750-757). Philadelphia, PA: W.B Saunders Company.

Peterson-Falzone, S., Hardin-Jones, M., & Karnell, M. (2010). *Cleft palate speech* (4th ed.) St. Louis, MO: Mosby, Inc.

Peterson-Falzone, S., Trost-Cardamone, J., Karnell, M., & Hardin-Jones, M. (2006). *The clinician's guide to treating cleft palate speech*. St. Louis, MO: Mosby, Inc.

Pickett, J. M. (1999). *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*. Needham Heights, MA: Allyn & Bacon.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31(3)*, 273-279.

Prins, D., & Bloomer, H. H. (1965). Consonant intelligibility: A procedure for evaluating speech in oral cleft subjects. *Journal of Speech and Hearing Research, 11*, 128-137.

Rosenzweig, M. R., & Postman, L. (1957). Intelligibility as a function of frequency of usage. *Journal of Experimental Psychology, 54*, 412-422.

Ruiter, J. S., Korsten-Meijer, A. G. W., & Goorhuis-Brouwer, S. M. (2009). Communicative abilities in toddlers and in early school age children with cleft palate. *International Journal of Pediatric Otorhinolaryngology, 73*, 693-698.

Scherer, N. J., Williams, A. L., & Proctor-Williams, K. (2008). Early and later vocalization skills in children with and without cleft palate. *International Journal of Pediatric Otorhinolaryngology, 72*, 827-840.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11,* 193-205.

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

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

Sosa, A.V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech, Language and Hearing Research, 55*, 596-608.

Stemach, G., & Williams, W. B. (1988). *Word express: The first 2, 500 words of spoken English illustrated*. Novato, CA: Academic Therapy Publications.

Stokes, S. F., & Whitehill, T. L. (1996). Speech error patterns in Cantonese-speaking children with cleft palate. *European Journal of Disorders of Communication, 31*, 45-64.

Storkel, H. L., & Hoover, J. R. (2010). An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods, 42(2),* 497-506.

Thorndike, E. L., & Lorge, I. (1952). *The teacher's word book of 30,000 words*. New York, NY: Teachers College, Columbia University.

Treiman, R., Tincoff, R., Rodriguez, K., Mouzaki, A., & Francis, D. J. (1998). The foundations of literacy: Learning the sounds of letters. *Child Development, 69*, 1524-1540.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). *Irvine Phonotactic Online Dictionary, Version 2.0*. [Data file]. Retrieved from http://www.iphod.com/

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28(4)*, 735-747.

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes, 21(6)*, 760-770.

Whitehill, T. L., & Chau, C. (2004). Single-word intelligibility in speakers with repaired cleft palate. *Clinical Linguistics & Phonetics, 18*, 341-355.

Weismer, G. (2008). Speech intelligibility. In *The Handbook of Clinical Linguistics* (35). Retrieved from http://www.blackwellreference.com/subscriber/uid=602/tocnode?id=g9781405135221_chunk_g978140513522137

Yates, M. (2009). Phonological neighbourhood spread facilitates lexical decisions. *The Quarterly Journal of Experimental Psychology, 62(7)*, 1304-1314.

Yorkston, K. M., & Beukelman, D. R. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communication Disorders, 13*, 15-31.

Young, T., Hodge, M., & Daniels, J. (2004). *TOCS+ Universal Sound Server.* [computer software]. Edmonton, AB; University of Alberta.

Ziegler, W., Hartmann, E., & von Cramon, D. (1988). Word identification testing in the diagnostic evaluation of dysarthric speech. *Clinical Linguistics & Phonetics, 2*, 291-308.

Ziegler, W., & Zierdt, A. (2008). Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders, 41*, 553-577.

Chapter 3

Reliability of *SIP-CCLP Ver. 5* Intelligibility Measurements

Introduction

Children with cleft lip and/or palate are at risk for developing some type of speech impairment, such as a resonance, articulation or voice disorder (Witzel, 1995). These disorders can affect how well a child with cleft lip and/or palate is understood by peers, caregivers, families and other people in their community. Speech intelligibility has been defined as the degree to which an individual's spoken message is recovered by a listener (Kent, Weismer, Kent & Rosenbek, 1989). When linguistic, morpho-syntactic, environmental and listener variables are controlled, intelligibility is considered to reflect the cumulative effect of a talker's resonance, articulation, voice and prosody during speech production.

Characteristics of the listener, such as familiarity with the speaker, the speech characteristics of the disordered population, and the test stimuli, are recognized as variables that may influence a speaker's intelligibility scores (Walshe, Miller, Leahy & Miller, 2008). These variables may also affect the reliability of intelligibility scores. Reliability of a measure may be assessed in terms of internal consistency or reproducibility (Scientific Advisory Committee of the Medical Outcomes Trust, 2002). Internal consistency examines how consistently examinees performed across items (Crocker & Algina, 1986), while reproducibility examines the stability of an instrument (over time (i.e., test-retest) or forms (i.e., alternate)) or inter-rater and intra-rater agreement (Scientific Advisory Committee of the Medical Outcomes Trust, 2002). For assessment of

test-retest reliability, scores are obtained from the same test form on two occasions. Two designs have been used to assess alternate forms reliability. In one, scores are obtained from two different parallel forms of the same test on two occasions for one group of participants (American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME), 1999). In the other, scores are obtained from two forms of a test in a single occasion (e.g., Zajac, Plante, Lloyd and Haley, 2011). These two designs are sensitive to different sources of error. In the former, the reliability coefficient is sensitive to error due to differences between forms and within examinees (over time); while in the latter, the reliability coefficient is sensitive to error due to differences between forms. Accordingly, two types of alternate forms reliability can be evaluated: coefficient of equivalence and stability (two forms, two occasions) and coefficient of equivalence (two forms, one occasion). Using the same nomenclature, test-retest reliability, which is sensitive to differences within examinees, can be described as a coefficient of stability. Crocker and Algina (1986) recommended that means, standard deviations and standard error of measurement be reported for each testing session when reporting the results of reliability studies in the classical test theory model. Lexell and Downham (2005) described a series of statistical methods that should be used to assess test-retest reliability, which included calculation of test-retest correlation coefficients (e.g., intraclass correlation coefficient (ICC; Type 2, 1)), indices of change (e.g., mean difference between occasions), indices of measurement variability (e.g., standard error of

measurement), and indices of "clinically important" changes (e.g., minimal detectable change; these indices are a statistical construct and do not take into account what constitutes a clinically important change from the perspective of clinicians or clients).

The Standards for Educational and Psychological Testing, developed by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education (1999), outlined the responsibilities of test developers when describing test reliability in a series of twenty standards. While some of the standards are applicable to specific testing conditions (e.g., timed test) or applications (e.g., tests used for program evaluation), other standards are applicable to all tests. Six standards applicable to discriminative speech intelligibility measures are outlined in the first column in Table 3-1. Standard 2.1 states that reliability estimates and standard errors of measurement should be reported for each score that is to be interpreted (e.g., total score, sub-score, and composite score). Standard 2.2 states that the standard error of measurement (SEM) should be reported for all scores used to interpret test results (e.g., raw and derived score). The SEM can then be used to construct confidence intervals for the examinee's true score. Standards 2.4a and b state that each method of determining reliability and the characteristics of the examinees used in the evaluation (e.g., demographics, sampling procedure) should be described. This information helps users assess if the test is appropriate to use with their examinee populations. Standard 2.5 states that as each type of reliability coefficient (and SEM) is sensitive to different sources of error, they

should not be used interchangeably. For example, alternate forms (coefficient of equivalence) reliability coefficients can be used to describe the consistency of examinees over forms, but not to describe the consistency of raters over time. Standard 2.10 states that inter-rater and intra-examiner reliability should be evaluated "when subjective judgment enters into test scoring" (p. 33). The test developer is also advised to clearly state whether inter-rater reliability is based on independent raters, a single panel of raters or an independent panel of raters.

There are two commercially available published measures of speech intelligibility for children: *Central Institute for the Deaf Picture Speech Intelligibility Evaluation (CID Picture SPINE)* (Monsen, Moog & Geers, 1988) and *Children's Speech Intelligibility Measure (CSIM)* (Wilcox & Morris, 1999). The *CID Picture SPINE* was developed to provide "a quantitative index of how intelligible a child's speech is in common communication situations" (p. 11) for children who are severely or profoundly hearing-impaired. For the *CID Picture SPINE*, inter-rater reliability was evaluated by determining the correlation between intelligibility scores from two examiners for 20 children with hearing impairment. The Pearson's correlation coefficient between examiners was .96. Internal consistency, test-retest, alternate forms and intra-rater reliability were not evaluated. Furthermore, standard error of measurement and how examinees were chosen to participate in the study of inter-rater reliability were not reported. Consequently, the *CID Picture SPINE* fails to meet the standards for test reliability described by AERA, APA and NCME (1999) (see Table 3-1).

The *CSIM* (Wilcox & Morris, 1999) was developed to provide an "objective measure of single-word intelligibility of children ages 3 years to 10 years, 11 months whose speech is considered unintelligible" (Wilcox & Morris, 1999; p. 1). Test-retest, alternate forms and inter-rater reliability, and internal consistency, have been established for the *CSIM* (Wilcox & Morris, 1999) using a sample of 148 children identified a priori as exhibiting unintelligible speech. For each analysis, children were divided into three groups classified by age: three years to four years, eleven months; five years to six years, eleven months; and seven years to ten years, eleven months. Results for the two youngest groups are reported in the following section to compare their results with those of similar-aged children on other intelligibility measures (e.g., *SIP-CCLP*). During *CSIM* administration, children repeated 50 target words after the examiner's model. Children's word productions were recorded and later played to a single unfamiliar listener who identified the word perceived from a list of 12 similar sounding words (closed-set response task). The percentage of words identified correctly served as the child's intelligibility score. Confidence intervals for raw scores are reported for each age group. The time interval between administrations of the same form of *CSIM* ranged from one to two weeks. Three unique forms were used in the assessment of test-retest reliability (coefficient of stability). Each child was randomly assigned to one form, which was re-administered at the second session. For the younger group, the mean intelligibility scores were 36.37% (SD = 14.53) at time one and 39.77% (SD = 15.95) at time two. For the older group, the mean intelligibility scores were 52.07% (SD = 20.30) at time one

and 55.85% (SD = 21.30) at time two. The Pearson's correlation coefficients for test-retest reliability for the two groups were .79 and .86, respectively.

One of the 304 different randomly generated forms was administered at time one and a second different randomly generated form was administered at time two to evaluate the alternate forms reliability of *CSIM* (Wilcox & Morris, 1999). Time between administrations ranged from one to two weeks. The mean intelligibility scores for the younger group were 36.18% (SD = 14.70) for the form administered at time one and 35.66% (SD = 16.89) for the form administered at time two. The mean intelligibility scores for the older group were 49.89% (SD = 19.01) at time one and 50.41% (SD = 19.14) at time two. The Pearson's correlation coefficients for alternate forms reliability (coefficients of equivalence and stability) were .64 and .84, respectively.

Inter-rater reliability of the *CSIM* (Wilcox & Morris, 1999) was evaluated by determining the relationship between intelligibility scores obtained from a rater who spoke or was familiar with the regional dialect of the child from the same geographic region and a second rater who was unfamiliar with the regional dialect of the child and from a different geographic region. Both raters were unfamiliar with the child's speech. The mean intelligibility scores for the younger group were 37.05% (SD = 15.83) for the first group of raters and 34.69% (SD = 14.05) for the second group of raters. The mean intelligibility scores for the older group were 48.90% (SD = 17.09) for the first group of raters and 46.98% (SD = 18.17) for the second group of raters. The correlation between raters (Pearson's correlation coefficient) was .80 for each group of children.

Intra-rater reliability of the *CSIM* (Wilcox & Morris, 1999) was evaluated by determining the relationship between intelligibility scores obtained from a group of four raters who judged one child's recordings twice, separated by an interval of one to two weeks. The mean intelligibility scores for the younger group (n = 41) were 32.60% (SD = 13.30) the first time that they were judged and 33.76% (SD = 11.98) the second time that they were judged. The mean intelligibility scores for the older group (n = 36) were 44.99% (SD = 17.36) the first time and 49.44% (SD = 18.97) the second time. The correlations between scores obtained from the same rater at two different points in time (Pearson's correlation coefficients) were .74 for the younger group and .89 for the older group.

Internal consistency of the *CSIM* (Wilcox & Morris, 1999) was evaluated using the coefficient alpha procedure for each of the four administrations (i.e., same form administered each session (2), unique form administered at session one, unique form administered at session two). Alpha coefficients ranged from 0.79 to 0.87 for the younger children and from 0.88 to 0.90 for the older children. The results of the evaluation of test-retest, alternate forms, inter-rater and intra-rater reliability and internal consistency met all of the standards for describing test reliability outlined in Table 3-1. However, the alpha coefficients for the *CSIM* are slightly lower than the value of 0.9-0.95 recommended by the Scientific Advisory Committee of the Medical Outcomes Trust (2002) for measures that will be used to assign scores to individuals. While no minimal standards have been developed for reliability coefficients for test-retest and alternate forms (Crocker & Algina,

1986), the *CSIM* reliability coefficients are less than .9, suggesting that *CSIM* scores may not consistently rank children relative to each other over time and over form and time. As a result, test users should be cautious when interpreting differences in children's scores from administrations over time and over different forms.

The reliability of two measures of speech intelligibility for English-speaking children with cleft palate has been reported. Zajac, Plante, Lloyd and Haley (2011)[3] described results of an evaluation of "parallel forms," inter-rater and intra-rater reliability for a 50-word intelligibility test for 22 children with unilateral or bilateral cleft palate ranging in age from five years to nine years, five months. Children's word productions for two randomly generated forms were recorded in one session and played back to listeners who completed an open-set (orthographic transcription) word identification task. The percentage of words identified correctly serves as the child's intelligibility score. Twenty undergraduate students with English as their primary language and hearing within normal limits served as listeners. Each listener judged both forms from 5 to 6 children with cleft palate in a single session. Five listeners judged each child's recordings. The percentage of words identified correctly served as the child's intelligibility score. The mean intelligibility score was 70.4% (SD = 18.3) on the first form and 69.0% (SD = 19.1) on the second form. The coefficient of

[3] In Zajac, Plante, Lloyd and Haley (2011), results of the evaluation of "parallel forms," inter-rater and intra-rater reliability were reported for a group of 22 children with cleft palate and 16 children without cleft palate. As scores for each listener for each child on each form were reported, it was possible to calculate reliability coefficients for just the children with cleft palate. These results are reported. Results are similar to those reported by Zajac et al for the full group of 38 children. It was not possible to calculate intrajudge reliability for the 6 children with cleft palate who were judged a second time.

equivalence (described by Zajac et al., 2011 as parallel forms reliability) was .97 (Pearson's correlation coefficient). As the same listeners judged the two forms from each child in a single listening session, listeners' familiarity with the speech of children with cleft palate, the speech of specific children in the listening set, and the task may have increased over the course of a listening session, which may influence listener perceptions of speech intelligibility (Walshe et al., 2008). Therefore, the parallel forms reliability coefficient reported by Zajac et al. (2011) is affected by differences due to forms and differences within listeners, as opposed to just differences due to forms. Inter-rater reliability was evaluated using intraclass correlation coefficients (ICC). The ICC (Type 1, 3) for the twenty listeners was .98 on form one and .99 on form two. Intra-rater reliability was reported for three listeners who judged a group of 10 children with and without cleft palate twice. The second listening session was three weeks after the first listening session. Pearson's correlation coefficients for the three listeners' scores for the two sessions ranged from .92 to .95. While Zajac et al. (2011) concluded that "all measures of reliability were adequate" (p. 538), this evaluation failed to meet several of the standards for describing test reliability listed in Table 3-1. Use of the same group of listeners to judge both forms of the test is a limitation. Hustad and Cahill (2003) found that when listeners judged different sets of sentences from a single speaker with dysarthria in a single session, intelligibility scores were consistently higher for the sentences heard second than those heard first. A second limitation of this study is that the reliability

coefficients reported were for the full group of 38 children with and without cleft palate but the test was designed to be used for children with cleft lip and palate.

Gotzke (2005) evaluated the reliability of the *Speech Intelligibility Probe for Children with Cleft Palate Version 3* (*SIP-CCLP Ver. 3)* with 15 children with cleft palate. *SIP-CCLP Ver. 3* is a computer-mediated measure of single-word intelligibility that uses a phonetic contrast approach to target the speech error patterns of children with cleft palate. Children's productions of single words are recorded and played back to listeners who complete a closed-set (multiple choice) response task. Each child's recordings were judged by three independent listeners. Ninety listeners with English as their first language and hearing within normal limits participated. Test-retest reliability was not evaluated. It was not possible to evaluate alternate forms reliability as there was only form of *Ver. 3*. An intraclass correlation coefficient (ICC; Type 1, 3) (Shrout & Fleiss, 1979) was calculated to evaluate the consistency of intelligibility scores across the sets of three listeners (inter-rater reliability). For the children with cleft palate, the ICC was .87 for the listeners' intelligibility scores and .92 for the listeners' phonetic accuracy scores. Phonetic accuracy scores capture all three types of errors characterizing the speech of children with cleft palate (i.e., substitutions, omissions and distortions). In calculating this score, more points are assigned to "correct/clear" responses than to "correct/distorted" responses. Intra-rater agreement was evaluated for twenty repeated items, selected randomly by the *SIP-CCLP* software for each listener. For 12 children with cleft palate, mean intra-rater agreement was 76.4% (range: 58.3 – 96.4) for identifying the same

response for the two presentations. Gotzke (2005) concluded that these results provided initial support that *SIP-CCLP* had potential to be a reliable measure of speech intelligibility for children with cleft palate. However, she recognized that further evaluation was necessary. As is shown in Table 3-1, several standards for describing test reliability had not been met.

Gotzke (2005) also recognized that revisions were needed to improve the sensitivity, efficiency and utility of *SIP-CCLP* for research and clinical applications. *SIP-CCLP Ver. 3* was revised substantially to create *SIP-CCLP Ver. 5*. As described in Chapter 2, revisions included creating a second form, decreasing the number of stimulus items elicited from children and judged by listeners, and increasing the number of response alternatives in the closed-set response task. The purpose of the current study was to use a classical test theory model to evaluate the test-retest, alternate forms, inter-rater and intra-rater reliability and internal consistency of scores obtained using *SIP-CCLP Ver. 5* following the standards described by AERA, APA and NCME (1999).

Method

This study was approved by the ethics review board at the University of Alberta. Children assented to participate. Informed consent was obtained from parents of the children participating and from listeners.

Participants.

Children. A total of twenty-one children with cleft palate, who ranged in age from 3 years, 1 month to 7 years, 0 months (mean age 4 years, 10 months; SD = 1 year, 2 months), were recruited by speech-language pathologists at the Alberta Children's Hospital or in the Edmonton area. These children represent a convenience sample of the children with cleft palate in Alberta. Descriptive information about these children is reported in Table 3-2. Four children with submucous cleft palate, six children with cleft palate only, six children with unilateral cleft lip and palate, and five children with bilateral cleft lip and palate participated. Two children's palates (CP05 and CP15) were not repaired. One child had secondary surgery for velopharyngeal dysfunction (CP20) and one child had secondary surgery to repair a fistula (CP15). All but three children (CP04, CP06 and CP12) had age-appropriate receptive language (i.e., $\geq$ 16th percentile), based on results of the *Fluharty Preschool Speech and Language Screening Test (Fluharty -2*; Fluharty, 2001) on the day of testing or previous assessment by the referring speech-language pathologist. Information about hearing was obtained from all but one child's (CP05) clinic file (file was not available). Hearing was described as within normal limits when aided (bone-anchored hearing aid) for one child (CP12). Hearing was described as within normal limits in at least one ear for the remaining 19 children. Four children had a diagnosed syndrome: CP12 with Goldenhaar syndrome; CP15 with Klippel-Feil syndrome; CP18 with Ectodermal dysplasia Ectodactyly Clefting (EEC) syndrome; and CP20 with Fetal Alcohol syndrome. Six children were adopted from China (CP06, CP11, CP14,

CP17, and CP19) or Hong Kong (CP12). Pollock and Price (2005) suggested that measures developed for monolingual English-speaking children may be used with internationally adopted children who have been exposed to English for two or more years; therefore, data from five of these children were included in this study. Data from one child were not included in this study because time since adoption was 11 months at the time of recording (CP06). To obtain a socioeconomic index for each child, Boyd-NP scores were assigned to each parent based on their occupations (Boyd, 2008). In cases where both parents were employed, the average Boyd-NP score was reported. The highest education level completed or in progress by the mother was Grade 11 for one child, high school for four children, some post-secondary education (i.e., some university or college diploma) for seven children and university (i.e., degree completed) for eight children. Maternal education was not reported by one family (CP20).

Nasalance scores on the picture-cued subtest of the Simplified Nasometric Assessment Procedure (SNAP) (Kummer, 2005) were obtained from the children's clinic file or by the examiner during the session, as additional descriptive information. Two children (CP03 and CP05) refused to wear the headset. For three children (CP02, CP07, CP16), nasalance scores were within two standard deviations of the norm on all four oral subtests. For five children (CP04, CP15, CP17, CP18 and CP20), nasalance scores were within two standard deviations of the norm on one or two of the four subtests. For the remaining 11 children, nasalance scores were more than two standard deviations from the norm on at least three of the four oral subtests.

Listeners. One hundred and fourteen listeners were recruited from the pool of students at the University of Alberta. All listeners had Canadian English as their first language and normal hearing as determined by a hearing screening performed according to Alberta College of Speech-Language Pathologists and Audiologists (2008) guidelines. Each set of recordings for a given form (e.g., *SIP-CCLP Ver. 5* form 1 stimulus words) was judged by two students in a graduate speech-language pathology program and one student in a different course of study. Students in speech-language pathology had not yet completed coursework in resonance or had formal training assessing the speech of children with cleft palate. An honorarium was given to each listener for their time and participation.

Recording. All recordings took place in either a quiet room or a sound booth. All speech samples were recorded directly to computer using an AudioBuddy Dual Mic Preamplifier connected to either a Shure WH20 unidirectional dynamic headset microphone or a Shure SM88 unidirectional hand-held microphone. The headset microphone was used for all but three children: two children (CP01, CP03) refused to wear the headset and one child (CP12) could not wear the headset with his bone-anchored hearing aid. *SIP-CCLP Ver. 5* software uses a sampling rate of 48 kHz and quantization size of 16 bits to record the child's utterances. Each utterance was saved as a separate .wav file in the child's folder. All sessions were also video-recorded using a Panasonic Model AG-DVC30 Digital Video Camera-Recorder and an Audio-Technica AT899 Subminiature Omnidirectional Condenser Microphone worn by the child.

In the first session, both forms of *SIP-CCLP Ver. 5* were administered to each child. Order of administration was counterbalanced among child participants. The software randomized order of item presentation to create a unique order for each child. The child was instructed to "listen for the word that goes with the picture and then say the **same** word" (i.e., repeat the target word after the pre-recorded model was played). The appearance of a frog icon in the upper left-hand corner of the presentation screen was used to cue the child when it was time to say the stimulus item and to signal the examiner that recording had started. Verbal reminders were also used to cue the children to wait for the "frog" before speaking. Familiarization training was conducted during administration of the first form for seven words that were judged to be unfamiliar to young children (i.e., bash, fail, Lee, spear, Sue, veil, zap) as described in chapter 2. Four practice words preceded the presentation of the test words. Short breaks were provided in the form of computer animations that appeared after every 20 stimulus words. If the examiner was unsure about the recording quality of any item or had any concerns about background noise or examiner voiceover, a second imitation was elicited. Administration of each form took between 8 and 10 minutes. The child's attention to the task and the number of times that items had to be repeated to ensure a clean recording affected the amount of time required to complete the task. Fourteen of the 20 children (and parents) included in the study agreed to return for a second session. The number of days between sessions ranged from 5 to 21 (mean: 10.8 days). For these children, order of form presentation was the

opposite of the first session (e.g., first session: form 1 followed by form 2; second session: form 2 followed by form 1).

Preparation of recordings for listening. The researcher listened to each child's recordings of the *SIP-CCLP Ver. 5* words, using *Adobe Audition 1.5* (Adobe Systems Incorporated, 2004) to playback the recordings. Each child's recordings were edited to ensure that any extraneous words and comments made by the child or examiner were removed. If there were multiple productions of the target word, sentence or phrase, the first production without examiner voiceover or environmental noise interference was saved as the .wav file for playback to listener judges.

Judging. All listening sessions took place in a sound booth. During the listening task, the computer hard drive was set up outside the sound booth to improve the signal-to-noise ratio. Speech samples were presented through a Technics Stereo Integrated Amplifier (model SU-V460) connected to ElectroVoice S-40 compact monitor speakers located in the sound booth. Playback volume of the speech sample was standardized to be between 50 – 65 dBA prior to presentation. Each listener independently judged form 1 and 2 from two different children. Order of judging tasks was counterbalanced across listeners.

The software created a unique presentation order of the items for each listener. For each form, the judging task consisted of four practice items and 63 phonetic contrast items. Listeners were instructed to choose which button best matched the sound(s) heard in the underlined position. Listeners were provided

with six choices: four minimally contrastive real words with consonant(s) underlined, a blank for typing in a response if what was heard differed from the choices provided, and "can't identify" if they are unable to identify what sound(s) was heard. Listeners were instructed to focus on what was heard in the underlined position when making their choice. When listeners chose a button with a word on it, they also rated what was heard in the underlined position as "clear" or "distorted." Listeners were given the option of hearing the child's production of the stimulus word a second time before rating it as "clear" or "distorted." For evaluation of intra-rater reliability, one listener for each child's form 1 and 2 recordings was asked to return and complete the same task one week later. Sixty-seven listeners participated; a second judgment was not obtained for one child's form 2 recordings from session two (CP01) because of an administrative error.

Calculation of dependent variables. After three listeners completed the closed-set response task, the *SIP-CCLP Ver. 5* analysis software checked each listener's response file for entries in the "other/blank" response option. These entries were presented, item by item, above the target and foil words, for the test administrator to review and either verify or recode as the "target" or "foil." The test administrator examined each entry typed in by the listeners to see if it actually contained the contrastive sound in the target or foil words. For example, the foils for the target "chew" are "you," "shoe" and "who." If a listener's typed-in response was "U," it would be recoded as the foil "you." If a listener's response was "two," it would not be recoded as it does not match the underlined sound in

the target or foils and the "no change" button would be selected. After checking and any recoding were completed, the software compiled and analyzed the three listeners' responses.

Items in which a minimum of two of the three listeners chose the "target" were given a score of "1." Items in which a minimum of two of the three listeners chose the same foil, no agreement was obtained among the three listeners, two of the three listeners chose "can't identify" or two of the three listeners typed a response in the "blank" button were given a score of "0". This information was used in the calculation of internal consistency of each form.

Two derived scores were obtained: intelligibility and phonetic accuracy. To calculate the intelligibility score**,** the percentage of words identified correctly was determined for each listener. The mean of the percentages for the three listeners, as calculated by the *SIP-CCLP* software, served as the child participant's intelligibility score.

To calculate the phonetic accuracy score, the software assigned a score of two points to each item identified correctly and assigned a "clear" rating and a score of one point to each item identified correctly and assigned a "distorted" rating. All other responses were assigned zero points. Number of points was summed for each listener and divided by the total number of items judged multiplied by the number of listeners multiplied by two and converted to a percentage to yield an individual's phonetic accuracy score (Gotzke & Hodge, 2008). The mean phonetic accuracy score for the three listeners, as calculated by

the *SIP-CCLP* software, served as the child participant's phonetic accuracy score.

Data analysis.

Parallel forms. Statistical testing for parallel forms is a two-step procedure in the classical test theory model (Rogers, 1999). First, the test developer evaluates whether there any effects due to order of administration using paired-samples t-tests for each form's means and variances for the two orders. If results are not significant at $p = .25$[4] (i.e., no order effect), then the test developer evaluates whether the two forms are parallel (i.e., have equal means and variances) using paired-samples t-tests for the means and variances of the combined scores from the two orders. The test developer concludes that the two forms are parallel if the results are not significant at $p = .25$. To test for any effects due to order, paired-samples t-tests were conducted for each form's means and variances for the two orders (i.e., form 1 followed by form 2, form 2 followed by form 1)

Test-retest and alternate forms reliability. Pearson's correlation coefficient and an intraclass correlation coefficient (Type 2, 1) were calculated to evaluate consistency of the scores over time (i.e., test-retest reliability; coefficient of stability), over form (i.e., alternate forms reliability; coefficient of equivalency) and over form and time (i.e., alternate forms reliability; coefficient of equivalency and stability). Although intraclass correlation coefficients are preferred for describing the reliability of measurements (Weir, 2005), Pearson's correlation

[4] In evaluation of parallel forms, a Type 2 error (i.e., accept the null hypothesis that the means and standard deviations of the two forms are equal when the two forms are in fact not parallel) has a more serious consequence; therefore, the probability of Type 1 error is relaxed to 0.25 (Rogers, 1999).

coefficients were also calculated to allow comparison with studies describing the reliability of measurements of intelligibility (e.g., Zajac et al., 2011). For both test-retest and alternate forms reliability, the standard error of measurement (SEM), defined as the square root of the mean square error term from the analysis of variance (ANOVA) used to calculate the ICC (Lexell & Downham, 2005), was calculated. The minimal detectable change (MDC) was calculated using the formula MDC = 1.96 X SEM X √2 (Weir, 2005). To assess if there was a bias in *SIP-CCLP Ver. 5* scores, the limits of agreement method was applied to scores obtained over the two sessions and on the two forms in a single session (Bland & Altman, 1986).

Inter-rater and intra-rater reliability. Inter-rater reliability was estimated using an intraclass correlation coefficient (Type 1, 3; Shrout & Fleiss, 1979) and SEM for the three listener judges' scores. For each set of two listeners, inter-rater reliability was estimated using an intraclass correlation coefficient (Type 1, 2) and SEM. Intra-rater reliability was estimated using an intraclass correlation coefficient (Type 1, 1) and SEM for listeners who judged the same child's recordings *SIP-CCLP Ver. 5* one week later.

Internal consistency. Internal consistency of each form (i.e., "extent to which items within an instrument are related to each other" (p. 25, Wilcox & Morris, 1999)) was assessed using Cronbach's alpha.

Results

Parallel forms (N = 20). The means and variances of the intelligibility scores for the 11 children who were administered form 1 followed by form 2 were significantly lower for form 1 (mean = 64.8%, SD = 22.0, range: 28.0 – 90.5) than form 2 (mean = 71.5%, SD = 19.9, range = 32.2 – 95.8; mean: $t = -2.91$, $p < .25$; variance: $t = 1.97$, $p < .25$). The means and variances of the phonetic accuracy scores for these children were also significantly lower for form 1 (mean = 55.3%, SD = 22.6, range = 20.1 – 85.4) than form 2 (mean = 60.0%, SD = 21.1, range = 22.8 – 87.0; mean: $t = -2.26$, $p < .25$; variance: $t = 1.43$, $p < .25$). For the 9 children who were administered form 2 followed by form 1, there was no significant difference in the means or variances of the intelligibility scores from form 1 (mean = 62.6%, SD = 14.3, range = 36.5 – 75.7) and form 2 (mean = 65.3%, SD = 13.4, range = 41.8 – 78.8) (mean: $t = -0.60$, $p > .25$; variance: $t = 0.56$, $p > .25$). There was also no significant difference in the mean phonetic accuracy scores for form 1 (mean = 52.3%, SD = 12.4, range = 31.0 – 65.3) and form 2 (mean = 54.8%, SD = 11.0, range = 35.7 – 69.8) for these children (mean: $t = -0.35$, $p > .25$; variance: $t = 0.66$, $p > .25$). Because an order effect was detected for the children who were administered form 1 followed by form 2, the two forms are not parallel.

Test-retest reliability (N = 14). Mean intelligibility and phonetic accuracy scores for the 14 children on the two administrations of *SIP-CCLP Ver. 5* form 1 and 2 are reported in Table 3-3. The Pearson's correlation coefficients (i.e., coefficients of stability) ranged from .93 to .98 and ICCs (Type 2, 1) ranged

from .93 to .97 for *SIP-CCLP Ver. 5* scores, shown in Table 3-4. The SEM ranged from 2.95 to 4.85. The minimal detectable change (MDC) ranged from 8.18% for form 1 intelligibility scores to 13.44% for form 2 phonetic accuracy scores.

The signed mean, standard deviation of the differences and 95% limits of agreement in intelligibility and phonetic accuracy scores are reported in Table 3-5. For form 1, the mean differences were greater than zero (i.e., -2.31 and -2.36 for intelligibility and phonetic accuracy scores, respectively) when session 2 scores were subtracted from session 1 scores. Zero appeared in the 95% limits of agreement for both scores. The difference between form 1 intelligibility scores was greater than 10% for two children (CP01, CP04). The difference between form 1 phonetic accuracy scores was greater than 10% for one child (CP04). For form 2, the mean differences were close to zero and zero appeared in the 95% limits of agreement for both scores when session 2 scores were subtracted from session 1 scores. The difference between form 2 intelligibility scores was greater than 10% for two children (CP09, CP11). The difference between form 2 phonetic accuracy scores was greater than 10% for three children (CP09, CP11, and CP14).

Alternate forms reliability – over form (N = 20). Scores for *SIP-CCLP Ver. 5* forms recorded in a single session from 20 children were used to evaluate alternate test reliability (over forms). For form 1, the mean intelligibility and phonetic accuracy scores were 63.81% (SD = 18.48; range: 28.04 – 90.48) and 53.98% (SD = 18.36; range: 20.11 – 85.45), respectively. For form 2, the mean

intelligibility and phonetic accuracy scores were 68.70% (SD = 17.11; range: 32.28 – 95.77) and 57.67% (SD = 17.12; range: 22.75 – 87.04), respectively. The Pearson's correlation coefficient (i.e., coefficient of equivalence) was .95 for intelligibility and .94 for phonetic accuracy scores, shown in Table 3-6. The ICC (Type 2, 1) was .91 (95% CI [.61, .97]) for intelligibility scores and .92 (95% CI [.76, .97]) for phonetic accuracy scores. The SEM was 4.16 and 4.46 for intelligibility and phonetic accuracy scores, respectively. The minimal detectable change was 11.54% for intelligibility scores and 12.37% for phonetic accuracy scores. Results of the limits of agreement method are reported in Table 3-7. When form 2 intelligibility scores were subtracted from form 1 intelligibility scores, the mean difference was -4.89 and zero appeared in the 95% limits of agreement. When form 2 phonetic accuracy scores were subtracted from form 1 phonetic accuracy scores, the mean difference was -3.69 and zero appeared in the 95% limits of agreement.

Alternate forms reliability – over form and time (N = 14). Scores from form 1 obtained at session one and scores from form 2 obtained at session two from 14 children were used to evaluate alternate test reliability (over forms and time). The Pearson's correlation coefficient (i.e., coefficient of equivalence and stability) was .96 for intelligibility and .95 for phonetic accuracy scores, shown in Table 3-6. The ICC (Type 2, 1) was .93 (95% CI [.75, .98]) for intelligibility scores and .94 (95% CI [.82, .98]) for phonetic accuracy scores. The SEM was 4.22 and 4.56 for intelligibility and phonetic accuracy scores, respectively. The

minimal detectable change was 11.69% for intelligibility scores and 12.63% for phonetic accuracy scores.

Inter-rater reliability. As shown in Table 3-8, ICCs (Type 1, 3) ranged from .97 to .98 for session one and from .94 to .98 for session two for the groups of three listeners who judged the recordings from the 14 children who completed two sessions. The SEMs ranged from 5.40 to 6.49 for session one and from 5.00 to 5.91 for session two for the same groups of listeners. For the sets of three listeners who judged the recordings from 20 children, the ICC (Type 1, 3) was .96 (95% CI [.91, .98]) for form 1 and .96 (95% CI [.92, .98]) for form 2 intelligibility scores. The SEM was 6.76 and 5.86 for form 1 and 2 intelligibility scores, respectively. The ICC (Type 1, 3) was .97 (95% CI [.94, .99]) for form 1 and .96 (95% CI [.92, .98]) for form 2 phonetic accuracy scores. The SEM was 5.49 and 5.79 for form 1 and 2 phonetic accuracy scores, respectively. The number of items on which listeners chose to hear the child's production of the target word a second time before rating ranged from 0 to 63 (mean = 20.63, SD = 12.71) for form 1 and from 0 to 62 for form 2 (mean = 23.93, SD = 9.68). To compare listeners who differed in amount of training judging disordered speech, a t-test for independent samples was conducted for two groups: speech-language pathology (n = 40) and other students (n = 20). Scores did not differ significantly for the two groups of listeners for form 1 or 2 (p-value range: 0.85 – 0.98).

As shown in Table 3-9, ICCs (Type 1, 2) ranged from .92 to .95 for intelligibility scores and from .93 to .97 for phonetic accuracy scores for the three sets of two listeners who judged the recordings from 20 children. The SEMs

ranged from 5.57 to 7.55 for intelligibility scores and from 4.75 to 6.40 for phonetic accuracy scores for the same sets of listeners.

Intra-rater reliability. Intra-rater reliability was evaluated for the 67 listeners who judged form 1 and 2 recordings from the 20 children participants and the subset of 14 child participants who returned for a second session (Table 3-10). The time between judging sessions was one week. For the 67 listeners, the ICC (Type 1, 1) was .96 (95% CI [.94, .98]) for intelligibility scores and .96 (95% CI [.94, .98]) for phonetic accuracy scores. The SEM was 2.94 for intelligibility scores and 3.30 for phonetic accuracy scores. To compare listeners who differed in amount of training judging disordered speech, a t-test for independent samples was conducted for two groups: speech-language pathology (n = 48) and other students (n = 19). Scores did not differ significantly for the two groups of listeners for form 1 and 2 intelligibility and phonetic accuracy scores (p-value range: 0.57 – 0.85).

The mean difference in intelligibility scores between the two sessions was -2.81 (SD = 4.15, range: 0 – 14.29) when session 2 scores were subtracted from session 1 scores. Intelligibility scores differed by more than 10% for two listeners. The mean difference in phonetic accuracy scores between the two sessions was -1.91 (SD = 4.67, range: 0 – 18.25) when session 2 scores were subtracted from session 1 scores. Phonetic accuracy scores differed by more than 10% for five listeners. A comparison of the response files for these five listeners revealed that the difference in scores between the two sessions was the result of listeners identifying more words correctly the second time (one listener),

identifying more words correctly and changing ratings from distorted to clear (three listeners), and changing ratings from clear to distorted (one listener).

Intra-rater agreement was examined for ten listeners who judged form 1 session 1 recordings from 10 child participants and ten listeners who judged form 2 session 1 recordings from a different set of 10 child participants. The percentage of exact matches ranged from 58.73% to 95.24% (mean = 80.79, SD = 9.69).

Internal consistency. Cronbach's alpha was 0.93 and 0.92 for form 1 and 2, respectively, for the 20 child participants.

Discussion

The primary purpose of this research was to evaluate the reproducibility and consistency of scores obtained using *SIP-CCLP Ver. 5* following the standards outlined by AERA, APA and NCME (1999) and described in Table 3-1. In the evaluation of test-retest and alternate forms reliability, Pearson's and intraclass correlation coefficients were greater than 0.9, suggesting that *SIP-CCLP Ver. 5* ranked the children in this study relative to each other with consistency. Furthermore, the standard error of measurement was less than 5% for test-retest and alternate forms reliability, suggesting that children's scores were consistent over forms, time, and time and form. Inter-rater and intra-rater reliability coefficients were greater than 0.9, indicating that scores were consistent across the groups of three listeners and within listeners. As expected, the SEM for the inter-rater reliability was larger than the SEM for intra-rater reliability.

Internal consistency reliability was greater than 0.9 for both forms, suggesting that all items are measuring the same construct with similar difficulty.

Test-retest reliability was evaluated using correlation coefficients, standard error of the measurement and the limits of agreement method (Bland & Altman, 1986). Coefficients of stability (Pearson's correlation coefficients) were higher for the *SIP-CCLP Ver. 5* than those reported for the *CSIM* for children in the same age range as those in this study (Wilcox & Morris, 1999). Differences between group means were <2.5% for both *SIP-CCLP Ver. 5* forms and scores, which is lower than the differences in group mean for the two younger groups of children used in the evaluation of the reliability of *CSIM*. However, the limits of agreement method revealed a bias for *SIP-CCLP Ver. 5* form 1 scores, such that scores obtained for recordings from the second administration were higher than scores obtained for recordings from the first administration. Surprisingly, a bias was not noted for form 2 scores. These results suggest that the two forms may differ in difficulty.

When evaluating test-retest reliability, the interval between test administrations should be consistent, as one assumption of test-retest reliability is that the population is stable over the time interval (Scientific Advisory Committee of the Medical Outcomes Trust, 2002). In this study, the time interval between the two sessions ranged from five days to three weeks. Timing of the second session was affected by availability of research space and families. As 18 of the 21 children in this study were from southern Alberta, travel for the test administrator was an additional consideration. Scores for the two children with

the longest interval between administrations (CP09 = 21 days, CP19 = 16 days) were within one standard deviation of the mean difference for both forms and scores, with the exception of form 2 phonetic accuracy scores for CP09. While the results of the evaluation of test-retest reliability suggests that children's intelligibility and phonetic accuracy are stable over time, future studies should ensure that both sessions are booked within a two-week period.

In this study, alternate forms reliability was evaluated in two ways. First, scores obtained from 20 children in one session were used to examine the reliability of forms. The coefficient for equivalence for intelligibility scores (Pearson's correlation coefficient) was slightly lower for this study than the coefficient calculated by this author for the data reported for children with cleft palate by Zajac et al. (2011). However, the average age of children in Zajac et al. (2011) was 86 months, which is more than two years older than the average age of children in this study (58 months). When results for the eight children similar in age to those in this study are examined, the coefficient of equivalence is the same in both studies. The mean difference between *SIP-CCLP Ver. 5* form 1 and 2 intelligibility scores in this study (4.9%) was greater than the mean difference between forms calculated by this author for the children with cleft palate in Zajac et al. (2011; i.e., 1.4%). This result may be somewhat misleading as the same judges listened to both forms in the same session in Zajac et al. (2011). Intelligibility scores for two forms obtained from the same judges would be expected to be more similar than scores obtained from different judges (Wilcox & Morris, 1999). Further analysis of form equivalency in the present study using

the limits of agreement method (Bland & Altman, 1986) revealed a bias in scores such that intelligibility scores on form 2 were higher than scores on form 1 for 18 of the 20 children. The conditions for parallelism were not met for the two forms, as scores on form 1 were significantly different from scores on form 2 for the children who were administered form 1 first. An item analysis has been conducted to identify which items are functioning differently in the two forms. As described in Appendix D, the following changes were recommended to improve the form equivalence with respect to item difficulty: exchange six items between forms and revise three items.

Alternate forms reliability was also evaluated comparing scores obtained from 14 children for form 1 recordings at the first session and form 2 recordings at the second session. This type of reliability evaluation is sensitive to differences due to forms and within examinees (over time). While the coefficient of equivalency and stability (Pearson's correlation coefficient) was higher for the *SIP-CCLP Ver. 5* intelligibility scores than the *CSIM*, the mean difference was higher for *Ver. 5* scores than the *CSIM* mean group difference for both groups of children. These results confirm previous analyses showing a difference between the two forms of *SIP-CCLP Ver. 5*.

As part of the evaluations of test-retest and alternate forms reliability, standard error of measurement was calculated to provide a measure of the precision of scores. Although SEM was consistently lower for intelligibility scores than for phonetic accuracy scores, SEM was less than 5% for each reliability estimate. SEM was used to calculate the minimal detectable change.

The MDC provides the minimal difference needed to be confident, at the 95% level, that there is a real difference in the scores from two children and not a difference consistent with the measurement error of the test. MDC ranged from 8.2% to 11.7% for intelligibility scores and from 12.4% to 13.4% for phonetic accuracy scores. Neither SEM nor MDC were reported by Wilcox and Morris (1999) or Zajac et al. (2011) for their measures of intelligibility. However, 90% confidence intervals for each raw score on *CSIM* (Wilcox & Morris, 1999) were presented to help users interpret whether a change in score for a child represents "actual change in the child's performance, or simply measurement error" (p. 15).

Intraclass correlation coefficients were used to evaluate the inter-rater and intra-rater reliability of *SIP-CCLP Ver. 5*. Intraclass correlation coefficients for inter-rater reliability were high (i.e., greater than .9) and comparable to those calculated by this author for Zajac et al. (2011). Intraclass correlation coefficients were higher than those reported by Gotzke (2005) for *SIP-CCLP Ver. 3*. The option to hear the child's production a second time before rating may have contributed to the higher inter-rater reliability of phonetic accuracy scores. Although the amount of training listeners had judging disordered speech likely differed between the two listeners in speech-language pathology and the third listener, scores did not differ between these two groups of listeners. For the sets of two listeners, intraclass correlation coefficients remained high (i.e., greater than .9). There was less than a one point difference between the SEM for two listeners and three listeners for all scores and sets of two listeners. These results suggest that there is little difference in the stability of scores from two versus three

listeners. Although three listeners are needed to obtain information about the error patterns that may be contributing to the child's intelligibility, *SIP-CCLP Ver. 5* scores calculated from two listeners are reliable.

Intraclass correlation coefficients for intra-rater reliability were also high. Further evaluation of intra-rater reliability revealed that when one individual rates the same child on the same form twice in a one week period, scores tended to be higher the second time than the first. The same trend was noted for both groups of listeners (i.e., students in speech-language pathology and students in other university programs). Wilcox and Morris (1999) also found that scores were higher the second time they were judged than the first time. They suggested that "a "practice effect" or "judge familiarity" factor may have affected the results" (p. 24) as only four judges were involved in the evaluation of intra-rater reliability. Although 67 listeners participated in the evaluation of intra-rater reliability in the current study, the same "practice effect" or "judge familiarity" factor appears to have affected listener results. This result suggests that if listeners judge that same child's recording twice within a one-week period, scores will be higher the second time. However, the average difference is less than the SEM for alternate test reliability (form and time). Therefore, even if listener's scores increase due to practice with the judging task and familiarity with the speaker, the examinee's true score likely falls within the 95% confidence interval. Further research is needed to determine if this increase also occurs when listeners judge a different set of words from the same child in a one-week period.

This assessment of reliability followed a classical test score model. In the classical test score model, reliability is expressed as a correlation coefficient that describes the relationship between the standard deviation of observed and true scores (i.e., the average score that would be obtained if the same test was taken an infinite number of times) on a test. Discrepancy between these scores is the result of error of measurement, a simple random variable. Each type of reliability evaluation identifies and quantifies a single source of measurement error (e.g., different occasions, test forms and combinations of items; Streiner & Norman, 2008). Generalizability theory was developed to provide a way to simultaneously evaluate multiple sources of measurement error (Streiner & Norman, 2008). Using generalizability theory, the test developer calculates the error variance resulting from identified sources of measurement error and from the interaction amongst them, as well as variance due to subjects. These values are used to provide a single overall estimate of dependability (which is equivalent to reliability in the classical test theory model) that describes the "accuracy of generalizing from a person's observed score to the average score that the individual would receive under all the possible conditions that the test user would be willing to accept" (i.e., the universe of generalization; Shavelson & Webb, 1991, p. 1). For readers who are interested, an evaluation of the dependability of *SIP-CCLP Ver. 5* scores using generalizability theory is outlined in Appendix E.

Item response theory is an alternative statistical framework for evaluating test reliability. In item response theory, reliability is expressed in terms of item and test information functions (Hambleton, Swaminathan, & Rogers, 1991). The

test information function describes where and how well items are working on the score scale and is calculated by summing the item information functions. The reciprocal of the square root of the test information function yields the standard error of the ability estimate. Using the test information function or standard error of the ability estimate, the relative efficiency of two forms of a test can be compared. The purpose of relative efficiency is to compare the precision of two different tests (or forms) measuring the same construct. An evaluation of the reliability of *SIP-CCLP Ver. 5* scores using item response theory is outlined in Appendix F for readers who are interested.

In conclusion, the results of this study provide support for the reliability of *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility for young English-speaking children with cleft palate. This evaluation is consistent with the standards for reporting the results of reliability studies developed by AERA, APA & NCME (1999). Reliability coefficients, standard error of measurement and minimal detectable change are reported. The results also identified that the current two forms of *SIP-CCLP Ver. 5* do not meet the conditions for being parallel. An item analysis has been conducted to identify items to be exchanged between the two forms and items to be revised to improve form equivalence.

Table 3-1

Comparison of Intelligibility Measures with Standards for Describing Test Reliability (AERA, APA, & NCME, 1999)

Standard	Published Measures of Speech Intelligibility		Intelligibility Measures for English-speaking Children with Cleft Palate		
	CID Picture SPINE (Monsen, Moog & Geers, 1988)	*CSIM* (Wilcox & Morris, 1999)	Zajac et al. (2011)	*SIP-CCLP Ver. 3* (Gotzke, 2005)	*SIP-CCLP Ver. 5* (current study)
2.1 Report reliability estimates for each score to be interpreted.	✗	✓	alternate forms only	✗	✓
2.2 Report standard error of measurement for each score used in test interpretation	✗	✗	✗	✗	✓
2.4a Describe methods for evaluating reliability. Use statistics appropriate to each method.	Inter-rater only	✓	✓	Inter-rater only	✓
2.4b Report how examinees in a reliability study were sampled. Describe examinees.	✗	✓	✓	✓	✓

Table 3-1 continued

Standard	Published Measures of Speech Intelligibility		Intelligibility Measures for English-speaking Children with Cleft Palate		
	CID Picture SPINE (Monsen, Moog & Geers, 1988)	*CSIM* (Wilcox & Morris, 1999)	Zajac et al. (2011)	*SIP-CCLP Ver. 3* (Gotzke, 2005)	*SIP-CCLP Ver. 5* (current study)
2.5 Do not interchange reliability coefficients or SEMs from different methods.	n/a	✓	✗	✓	✓
2.10 Report inter-rater and intra-examinee reliability when test scoring is subjective.	Inter-rater only	✓	Inter-rater only	Inter-rater only	✓

Note. ✓ = consistent with standard; ✗ = not consistent with standard.

Table 3-2

Characteristics of the Child Participants

Participant	Age (months)	Gender	Cleft Type	Receptive Language	Boyd-NP Score	Nasalance (%)					Number of days between sessions
						Bilabial	Alveolar	Velar	Sibilants	Nasal	
CP01	37	F	CPO	WNL	87.5	52	48	38	70	67	14
CP02	39	F	SMCP	WNL	89	15	17	14	14	54	7
CP03	40	M	UCLP	WNL	89	n/a	n/a	n/a	n/a	n/a	9
CP04	42	M	BCLP	$\leq 16^{th}$ %ile	48	13	15	14	39	55	7
CP05	44	M	SMCP[1]	WNL	79.5	n/a	n/a	n/a	n/a	n/a	
CP06	44	F	UCLP	$\leq 16^{th}$ %ile	87	22	26	27	35	59	
CP07	49	F	UCLP	WNL	44	20	22	26	25	56	10
CP08	50	F	CPO	WNL	16	30	30	21	26	59	
CP09	53	M	UCLP	WNL	71	31	32	39	42	53	21
CP10	55	M	UCLP	WNL	30	28	25	21	39	44	9

Participant	Age (months)	Gender	Cleft Type	Receptive Language	Boyd-NP Score	Nasalance (%)					Number of days between sessions
						Bilabial	Alveolar	Velar	Sibilants	Nasal	
CP11	60	M	BCLP	WNL	90	38	40	32	48	59	7
CP12	63	M	CPO	$\leq 16^{th}$ %ile	70	64	50	63	74	66	
CP13	64	M	UCLP	WNL	34	48	40	45	50	64	
CP14	66	F	BCLP	WNL	55	11	13	11	16	50	14
CP15	70	F	SMCP[1]	WNL	30	23	19	17	56	62	
CP16	72	F	CPO	WNL	45.5	15	17	7	10	58	9
CP17	72	M	BCLP	WNL	82	21	18	27	20	46	7
CP18	74	M	BCLP	WNL	55	18	16	14	30	35	5
CP19	74	F	SMCP	WNL	n/a[2]	55	53	53	59	69	16
CP20	77	F	CPO	WNL	62	16	14	11	45	62	
CP21	84	M	CPO	WNL	92.5	27	28	36	47	66	14

Notes. [1]The cleft was unrepaired for these two children. [2]Both parents were self-employed but did not provide any additional information about their occupation; therefore, it was not possible to assign a Boyd-NP score. M = male; F = female. CPO = cleft palate only; SMCP = submucous cleft palate; UCLP = unilateral cleft lip and palate; BCLP = bilateral cleft lip and palate; WNL = within normal limits; n/a = not available. A grey box in the "number of days between sessions" column indicates that the child did not complete a second session.

Table 3-3

SIP-CCLP Ver. 5 Intelligibility and Phonetic Accuracy Scores for Forms 1 and 2 (N = 14)

	Time 1			Time 2		
	Mean	SD	Range	Mean	SD	Range
Form 1						
Intelligibility (%)	65.31	19.84	28.0 - 90.5	67.61	19.04	29.6 - 91.0
Phonetic Accuracy (%)	55.35	19.89	20.1 - 85.5	57.71	19.67	22.8 - 87.8
Form 2						
Intelligibility (%)	68.59	19.12	32.3 - 95.8	69.05	16.97	42.9 - 94.7
Phonetic Accuracy (%)	57.28	18.55	22.8 - 87.0	57.79	18.00	33.6 - 89.7

Table 3-4

Reliability Coefficients (r, ICC (2,1)) and Error Estimates (SEM, MDC) for SIP-CCLP Ver. 5 Test-Retest Reliability

	r	ICC (2, 1)	95% CI for ICC	SEM	MDC (%)
Form 1					
Intelligibility	0.98	.97	[.90, .99]	2.95	8.18
Phonetic Accuracy	0.97	.96	[.88, .99]	3.71	10.29
Form 2					
Intelligibility	0.97	.96	[.89, .99]	3.67	10.17
Phonetic Accuracy	0.93	.93	[.81, .98]	4.85	13.44

Note. r = Pearson's correlation coefficient; ICC = intraclass correlation coefficient; SEM = standard error of the measurement; MDC = minimal detectable change.

Table 3-5

Bias and Limits of Agreement for Measurements Obtained from Two Administrations (first – second) of SIP-CCLP Ver. 5 (N = 14)

		Mean Difference (diff)	SD_{diff}	95% Limits of Agreement
Form 1				
	Intelligibility (%)	-2.31	4.17	[-10.48, 5.87]
	Phonetic Accuracy (%)	-2.36	5.25	[-12.65, 7.93]
Form 2				
	Intelligibility (%)	-0.45	5.19	[-10.69, 9.72]
	Phonetic Accuracy (%)	-0.51	6.86	[-13.95, 12.92]

Note. SD = standard deviation.

Table 3-6

Reliability Coefficients (r, ICC (2,1)) and Error Estimates (SEM, MDC) for SIP-CCLP Ver. 5 Alternate Forms Reliability

	r	ICC (2, 1)	95% CI for ICC	SEM	MDC (%)
Over Forms (N = 20)					
Intelligibility	0.95	.91	[.61, .97]	4.16	11.54
Phonetic Accuracy	0.94	.92	[.76, .97]	4.46	12.37
Over Time and Form (N = 14)					
Intelligibility	0.96	.93	[.75, .98]	4.22	11.69
Phonetic Accuracy	0.95	.94	[.82, .98]	4.56	12.63

Note. r = Pearson's correlation coefficient; ICC = intraclass correlation coefficient; SEM = standard error of the measurement; MDC = minimal detectable change.

Table 3-7

Bias and Limits of Agreement for Measurements Obtained from SIP-CCLP Ver. 5 Forms 1 and 2 (Time 1 Form 1-Time 1 Form 2; N = 20)

	Mean Difference (diff)	SD_{diff}	95% Limits of Agreement
Intelligibility (%)	-4.89	5.89	[-16.43, 6.64]
Phonetic Accuracy (%)	-3.69	6.31	[-16.06, 8.68]

Note. SD = standard deviation.

Table 3-8

Inter-rater Reliability for Time 1 and Time 2 for SIP-CCLP Ver. 5 (N = 14)

	Time 1			Time 2		
	ICC (1, 3)	95% CI for ICC	SEM	ICC (1, 3)	95% CI for ICC	SEM
Form 1						
Intelligibility	.97	[.92, .99]	6.49	.97	[.94, .99]	5.54
Phonetic Accuracy	.97	[.93, .99]	6.02	.98	[.95, .99]	5.00
Form 2						
Intelligibility	.97	[.94, .99]	5.40	.97	[.92, .99]	5.58
Phonetic Accuracy	.97	[.93, .99]	5.78	.97	[.92, .99]	5.91

Note. ICC = intraclass correlation coefficient, CI = confidence interval, SEM = standard error of measurement.

Table 3-9

Inter-rater Reliability for Two Listeners for SIP-CCLP Ver. 5 (N = 20)

	Listener 1 – Listener 2			Listener 1 – Listener 3			Listener 2 – Listener 3		
	ICC (1, 2)	95% CI for ICC	SEM	ICC (1, 2)	95% CI for ICC	SEM	ICC (1, 2)	95% CI for ICC	SEM
Form 1									
Intelligibility	.94	[.85, .98]	6.63	.92	[.80, .97]	7.54	.95	[.88, .98]	6.02
Phonetic Accuracy	.95	[.88, .98]	5.86	.97	[.92, .99]	4.75	.95	[.87, .98]	5.81
Form 2									
Intelligibility	.95	[.87, .97	5.81	.94	[.86, .98]	5.57	.94	[.86, .98]	6.17
Phonetic Accuracy	.97	[.92, .99]	4.90	.94	[.84, .98]	5.96	.93	[.83, .97]	6.40

Note. ICC = intraclass correlation coefficient, CI = confidence interval, SEM = standard error of measurement.

Table 3-10

Intra-rater Reliability for SIP-CCLP Ver. 5 Form 1 and 2 Scores

	Time 1			Time 2		
	ICC (1, 1)	95% CI for ICC	SEM	ICC (1, 1)	95% CI for ICC	SEM
Form 1						
Intelligibility	.96	[.90, .98]	3.38	.96	[.88, .99]	3.49
Phonetic Accuracy	.96	[.91, .99]	3.24	.98	[.93, .99]	3.11
Form 2						
Intelligibility	.98	[.94, .99]	2.37	.96	[.89, .99]	2.57
Phonetic Accuracy	.97	[.93, .99]	2.43	.92	[.76, .97]	4.64

Notes. ICC = intraclass correlation coefficient; CI = confidence interval, SEM = standard error of measurement. Data is based on 20 independent listeners for Form 1 Time 1 recordings, 20 independent listeners for Form 2 Time 2 recordings, 14 independent listeners for Form 1 Time 2 recordings and 13 independent listeners for Form 2 Time 2 recordings.

References

Alberta College of Speech-Language Pathologists and Audiologists. (2008). *Hearing screening guidelines*. Retrieved from http://www.acslpa.ab.ca/public/data/documents/ACFC3D5.pdf

Adobe Systems Incorporated. (2004). *Adobe Audition 1.5* [computer software]. San Jose, CA; Adobe Systems Incorporated.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bland, J. M., & Altman, D G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, i,* 207-310.

Boyd, M. (2008). A socioeconomic scale for Canada: Measuring occupational status from the census. *Canadian Review of Sociology, 45(1)*, 51-91.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Fluharty, N. (2001). *Fluharty Preschool Speech and Language Screening Test (Second edition)*. Austin, TX, Pro-Ed.

Gotzke, C. L. (2005). *Speech intelligibility probe for children with cleft palate version 3: Assessment of reliability and validity*. (Unpublished master's thesis.) University of Alberta, Edmonton, AB.

Gotzke, C. L., & Hodge, M. (2008). S*peech intelligibility probe for children with cleft palate Version 4.0 (SIP-CCLP Ver. 4.0) user's guide*. (Unpublished manuscript). University of Alberta, Edmonton, AB.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*, 198-208.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Kummer, A. W. (2005). *The MacKay-Kummer SNAP Test-R: Simplified Nasometric Assessment Procedures*. KayPENTAX. Retrieved from http://www.kayelemetrics.com/index.php?option=com_product&view=product&Itemid=3&controller=product_innerpage&rec_id=50&no_id=2

Lexell, J. E., & Downham, D. Y. (2005). How to assess the reliability of measurements in rehabilitation. *American Journal of Physical Medicine and Rehabilitation, 84*, 719-723.

Monsen, R., Moog, J. S., & Geers, A. E. (1988). *CID Picture SPINE SPeech Intelligibility Evaluation*. St. Louis, MO: Central Institute for the Deaf.

Pollock, K. E., & Price, J. R. (2005). Phonological skills of children adopted from China: Implications for assessment. *Seminars in Speech and Language, 26(1)*, 54-63.

Rogers, T. (1999). *Error of measurement and validity.* (Unpublished class notes for Educational Psychology 507: Test Theory). University of Alberta, Edmonton, AB.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11,* 193-205.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications, Inc.

Shrout, P. E.. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2)*, 420-428.

Streiner, D.L., & Norman, G.R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford, UK: Oxford University Press.

Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language & Communication Disorders, 43(6)*, 633-648.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the *SEM*. *Journal of Strength and Conditioning Research, 19(1)*, 231-240.

Wilcox, K., & Morris. S. (1999). *Children's Speech Intelligibility Measure*. San Antonio, TX: The Psychological Corporation.

Witzel, M. A. (1995). Communicative impairment associated with clefting. In R. J. Shprintzen & J. Bardach (Eds.), *Cleft palate speech management: A multidisciplinary approach* (pp. 137-166). St. Louis, MO: Mosby.

Zajac, D. J., Plante, C., Lloyd, A., & Haley, K. L. (2011). Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate. *Cleft Palate-Craniofacial Journal, 48(5)*, 538-549.

Chapter 4

Evaluation of the Validity of the *Speech Intelligibility Probe for Children with Cleft Palate Version 5 (SIP-CCLP Ver. 5)*

Introduction

Intelligibility has been defined as the degree to which an individual's spoken message is recovered by a listener (Kent, Weismer, Kent & Rosenbek, 1989). Intelligibility scores reflect interactions among speaker, listener, and communication situation. As a result, the clarity of the speech signal produced by the speaker, the characteristics of the listener and the method used to estimate intelligibility all influence intelligibility scores. Clinically, intelligibility is used as a measure of speech disorder severity and is often estimated through the use of equal-appearing interval scales (Whitehill, 2002). However, Schiavetti (1992) questioned the validity of using rating scales for intelligibility because he found that listeners are unable to divide intelligibility into equal intervals.

The validity of a test has been defined as "the degree to which the instrument measures what it purports to measure" (Scientific Advisory Committee of the Medical Outcomes Trust, 2002, p. 200). Evidence of a measure's validity can be content-related, criterion-related and/or construct-related. A common means of assessing content-related validity is to have items in a measure evaluated by a panel of experts familiar with the subject matter of the measure and/or the population for whom the measure is intended for (Crocker & Algina, 1986; Scientific Advisory Committee of the Medical Outcomes Trust, 2002).

This was addressed in Chapter 2. The current chapter focuses on the evaluation of criterion and construct-related validity. Evidence for criterion-related validity is obtained by examining the relationship of test scores to scores obtained from a scaled, valid measure of the test construct obtained at the same (concurrent) or a later (predictive) time. Common methods for assessing construct-related validity include examining relationships between test scores and constructs that are and are not expected to be related (convergent and divergent validity), comparing groups known to differ on the construct, and conducting factor analysis to identify the variable(s) accounting for variation in the construct (Crocker & Algina, 1986). Validity evidence is collected to "support the intended interpretation of the test scores and their relevance to the proposed use" (p. 9; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Assessment of validity is an ongoing process, as no one study can address all aspects of validity. Furthermore, if test items are modified or different scores are derived from the test, new validity studies are needed (Streiner & Norman, 2008).

There are two commercially available measures of speech intelligibility for English-speaking children: *Central Institute for the Deaf Picture Speech Intelligibility Evaluation (CID Picture SPINE)* (Monsen, Moog & Geers, 1988) and *Children's Speech Intelligibility Measure (CSIM)* (Wilcox & Morris, 1999). The *CID Picture SPINE* was developed to provide "a quantitative index of how intelligible a child's speech is in common communication situations" (p. 11) for children who are severely or profoundly hearing-impaired. Predictive criterion-

related validity was evaluated for a sample of 20 children with profound hearing impairment ranging in age from 6 to 13 years. Intelligibility scores obtained from two independent examiners on the *CID Picture SPINE* were correlated with scores obtained from an independent measure of intelligibility obtained at a later time. To obtain the latter measure, children's productions of 10 sentences were transcribed orthographically by 12 listeners and mean intelligibility scores (percentage of words identified correctly) calculated. The correlation between scores was r = .96 for one examiner and r = .91 for the other examiner. Studies of construct-related validity were not reported in the test manual.

The *CSIM* (Wilcox & Morris, 1999) was developed to provide an "objective measure of single-word intelligibility of children ages 3 years to 10 years, 11 months whose speech is considered unintelligible" (Wilcox & Morris, 1999; p. 1) to monitor progress during treatment. Concurrent criterion-related and convergent construct-related validity were evaluated for the *CSIM* using a sample of children identified by speech-language pathologists as exhibiting unintelligible speech in three age groups: three years to four years, eleven months; five years to six years, eleven months; and seven years to ten years, eleven months. Results for the two younger groups of children are reported here as they cover the age range of interest in the current study. Concurrent criterion-related validity was evaluated by determining the correlation between *CSIM* intelligibility scores and speech language pathologists' ratings of intelligibility. Each child's speech-language pathologist provided a clinical rating estimating the child's intelligibility in connected speech in percentage points. A 20-point descriptive rating scale with

each point described by an intelligibility range in percentage points was used (e.g., rating of 7 represents 31-35% intelligible in connected speech). The mean *CSIM* intelligibility score was 35.95%[5] (SD = 15.14), and the mean clinician rating was 8.00 (SD = 4.15; equivalent to 36-40% intelligible in connected speech) for the younger group. The mean *CSIM* intelligibility score was 55.23% (SD = 18.08) and the mean clinician rating was 10.05 (SD = 5.20; equivalent to 46-50% intelligible in connected speech) for the older group. Moderate positive correlations were reported between *CSIM* intelligibility scores and clinical ratings for the two groups (r = .52 and r = .46, respectively). Convergent construct-related validity was evaluated by correlating *CSIM* intelligibility scores with the number of errors on the *Goldman-Fristoe Test of Articulation* (*GFTA*; Goldman & Fristoe, 1986). The mean *CSIM* intelligibility score was 38.61% (SD = 17.26) and the mean number of errors on the *GFTA* was 38.87 (SD = 13.40) for the younger group. The mean *CSIM* intelligibility score was 52.52% (SD = 17.76) and the mean number of errors on the *GFTA* was 30.70 (SD = 13.40) for the older group. Moderate negative correlations between *CSIM* intelligibility scores and the number of errors on the *GFTA* were reported for both groups of children (r = -.63 and r = -.55, respectively), suggesting a tendency for intelligibility to decrease as the number of articulation errors increased.

[5] There were a different number of children in the evaluations of construct and criterion-related validity. In the examination of construct-related validity, the number of children in both groups was 23. In the examination of criterion-related validity, there were 38 children in the younger group and 39 children in the older group. As a result, the mean *CSIM* score is different in each evaluation.

Validity has been examined for two measures of speech intelligibility for young English-speaking children with cleft palate. Zajac, Plante, Lloyd and Haley (2011) evaluated construct-related validity of a single-word intelligibility test "designed to be a global measure of severity of speech disability in children with repaired cleft lip and palate" (p. 540). Convergent construct-related validity was examined by correlating the mean intelligibility score obtained on two alternate test forms with a measure of articulation accuracy (percentage of consonants correct score (PCC)). The mean intelligibility score for the 21 children with unilateral cleft lip and palate was 68.9% (SD = 18.63) and their mean PCC score was 84.7% (SD = 15.59). A moderate positive correlation was reported between intelligibility and PCC scores (r = .78). Construct validity was also examined by comparing mean intelligibility scores on two alternate test forms for children with (n = 22) and without (n = 16) cleft lip and palate. The mean intelligibility score on the two forms was 70% (SD = 18) for the children with cleft lip and palate and 81% (SD = 13%) for the children without cleft lip and palate. A significant group difference in intelligibility scores was reported, such that intelligibility scores for the children with cleft lip and palate were lower than those for the children without cleft lip and palate. Criterion-related validity was not examined.

Gotzke (2005) evaluated criterion and construct-related validity of the *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP) Ver. 3* with 30 children with and without cleft palate (15 children per group), ranging in age from three to seven years. Concurrent criterion-related validity was examined

by correlating both intelligibility and phonetic accuracy scores from *SIP-CCLP Ver. 3* with intelligibility scores from a 100-word conversational speech sample. For the children with cleft palate, the mean intelligibility score was 66.07% (SD = 15.07) for the conversational speech sample. Positive correlations between intelligibility scores (closed-set response task) for *SIP-CCLP* and the conversational speech sample (r = .39) and between phonetic accuracy scores and intelligibility scores for the conversational speech sample (r = .56) were reported for the children with cleft palate. Construct-related validity was examined by comparing *Ver. 3* scores on the closed-set response task for children with and without cleft palate. The mean intelligibility score was 81.58% (SD = 10.83) for the children with cleft palate and 95.43% (SD = 2.93) for the children without cleft palate. The mean phonetic accuracy score was 68.83% (SD = 13.89) for the children with cleft palate and 89.82% (SD = 4.61) for the children without cleft palate. As predicted, compared to the children without cleft palate, those with cleft palate had significantly lower intelligibility and phonetic accuracy scores. Construct-related validity was also examined by describing the error patterns identified using *Ver. 3* for the children with and without cleft palate. More errors were identified in each of the six error categories (i.e., manner preference error, place preference error, glottal error, voicing error, sibilant error, cluster error) for the children with cleft palate than for the children without cleft palate. The majority of errors made by the children with cleft palate were of two types: manner preference errors (n = 58) and place preference errors (n = 48). Gotzke (2005) concluded that the results of this initial evaluation provided support for the

validity of *SIP-CCLP Ver. 3* as a measure of speech intelligibility for young children with cleft palate. Table 4-1 summarizes the approaches and results that have been reported for evaluating the criterion and construct-related validity of the *CID Picture Spine*, the *CSIM*, the measure developed by Zajac et al. (2011) and *SIP-CCLP Ver. 3*.

As described in Chapter 2, *Ver. 3* was revised substantially to create *SIP-CCLP Ver. 5*. Therefore, new validity studies were needed. This study evaluated the criterion and construct-related validity of *Ver. 5* as a discriminative measure of speech intelligibility that provides an index of severity of speech disorder for young English-speaking children with cleft palate. Spontaneous (or conversational) speech samples are preferred for evaluating children's speech intelligibility (Gordon-Brannan & Hodson, 2000) as this kind of speech sample has the highest face validity. Therefore, an intelligibility score based on word identification of a spontaneous speech sample was identified as a criterion variable in this study. Both concurrent (same session) and predictive (second session) criterion-related validity of *SIP-CCLP Ver.5* were assessed by examining relationships of its scores to intelligibility scores obtained from 100-word spontaneous speech samples. Moderate positive correlations were expected. Concurrent criterion-related validity of *Ver. 5* was also evaluated by examining relationships of its scores to intelligibility scores obtained from an imitative sentence measure (*Test of Children's Speech Plus* (*TOCS+*) sentence intelligibility measure; Hodge, Daniels & Gotzke, 2009), as children's *TOCS+* sentence and *SIP-CCLP* word productions are obtained using the same method

(i.e., delayed imitation). Moderate positive correlations were expected. Alternate forms reliability has been evaluated for the *TOCS+* sentence intelligibility measure for a group of 18 children with cerebral palsy and developmental dysarthria (Hodge & Gotzke, 2010). A coefficient of stability and equivalence (Pearson's correlation coefficient) of .93 was reported.

Construct-related validity was examined by describing the error patterns identified using *SIP-CCLP Ver. 5*, examining the relationships of measures of hypernasality, voice severity, and articulation accuracy to *Ver. 5* scores, and modeling the relationships of speech variables (i.e., hypernasality and voice severity ratings) and *SIP-CCLP Ver. 5* scores to intelligibility scores obtained from the two criterion measures. It was expected that the majority of errors would be among the eleven error patterns identified by experts in the content-review (i.e., nasals for obstruents, stopping, gliding of liquids, velar stops for obstruents, glottal stops for oral stops and affricates, glottal fricatives for oral fricatives, palatal fricatives for alveolar fricatives, fronting, deletion of an obstruent from an obstruent-sonorant cluster, and backing and cluster reduction). Based on previous research examining the relationships of speech variables to intelligibility for children with cleft palate, the following was expected: 1) a weak-to-moderate negative correlation between hypernasality ratings and *Ver. 5* scores (Keuning, Wieneke, van Wijngaarden & Dejonckere, 2002; McWilliams, 1954); 2) a weak-to-moderate negative correlation between voice severity ratings and *Ver. 5* scores (Whitehill & Wong, 2006); and 3) a moderate-to-strong positive correlation between a measure of articulation accuracy and *Ver. 5* scores (Moller

& Starr, 1984; Whitehill & Chun, 2002, Zajac et al., 2011). Construct-related validity was also examined by determining the relative contribution of predictor variables (i.e., hypernasality ratings, voice severity ratings, *SIP-CCLP* scores) to intelligibility scores obtained from the criterion variables. Konst, Rietveld, Peters & Weersink-Braks (2003) reported that 93% of the variance in intelligibility ratings for two-and-a-half year old children with (n = 20) and without (n = 8) unilateral cleft lip and palate was explained by correctness of articulation, lateralization and backing. An R^2 of at least 0.6 was expected for each model in the current study as it was expected that at least two of the predictor variables would contribute significantly to each model. The approaches taken to evaluate the validity of *SIP-CCLP Ver. 5* are summarized in Table 4-1.

Method

Participants.

Children and listeners. The 20 children with cleft palate described in Chapter 3 participated in this study. Listeners were a subset of the listeners described in Chapter 3.

Expert raters. One speech-language pathologist with expert knowledge about resonance and one speech-language pathologist with expert knowledge about voice disorders were recruited from the community. Each expert rater had more than 10 years experience analyzing the speech of children with resonance or voice disorders and normal hearing as determined by a hearing screening performed on the day of their participation according to Alberta College of

Speech-Language Pathologists and Audiologists (2008) guidelines. An honorarium was given to each expert rater.

Recordings. All recordings took place in either a quiet room or a sound booth. All speech samples were recorded directly to computer using an AudioBuddy Dual Mic Preamplifier and software (i.e., *SIP-CCLP Ver. 5*, *TOCS+ Recorder-Player Ver. 2.0* (Hodge, Gotzke & Daniels, 2009), *TOCS+ Intelligibility Measures* (Hodge, Daniels & Gotzke, 2009) and *Zoo Passage Recorder* (Gotzke & Hodge, 2011)) with a sampling rate of 48 kHz and a quantization size of 16 bits. A Shure WH20 unidirectional dynamic headset microphone was used for all but three children: two children (CP01, CP03) who refused to wear the headset and one child (CP11) who could not wear the headset with his bone-anchored hearing aid. For these three children, a Shure unidirectional SM88 microphone was used. It was held by the examiner close (range: 1 – 4 inches) to the child during production of the *SIP-CCLP Ver. 5* words, *TOCS+* Sentences and *Zoo Passage* (Fletcher, 1978) and placed on a microphone stand on the table during collection of the spontaneous speech sample. All sessions were also video-recorded using a Panasonic Model AG-DVC30 Digital Video Camera-Recorder and an Audio-Technica AT899 Subminiature Omnidirectional Condenser Microphone worn by the child.

A fifteen minute spontaneous speech sample was collected while the child and examiner were playing with playdough (e.g., a parallel play and an interactive play scenario; Shriberg & Kwiatkowski, 1985). The examiner commented about the play materials and asked open-ended questions on a topic of interest to the

child to encourage conversation. The *TOCS+ Recorder-Player Ver. 2.0* (Hodge, Gotzke, & Daniels, 2009) software was used to record the speech sample. A second spontaneous sample was recorded using the same procedures from the 14 children who returned for a second session. *Adobe Audition 1.5* (Adobe Systems Incorporated, 2004) was used to playback the recording for orthographic transcription of each child's spontaneous sample. Words that could not be understood by the examiner after five attempts were indicated in the transcript with an asterisk. The orthographic transcription was used to identify the 100-word sample that would be played back to listeners and to provide a "key" to use to obtain the intelligibility score for the sample. After utterances in the first minute of the recording were omitted, the examiner counted words until 100 consecutive words were identified. Word-for-word repetitions of previous utterances, fillers (e.g., "uh", "um"), or exclamations ("wow") were not counted. Conventions developed by Shriberg, Kwiatkowski and Rasmussen (1990) were used to determine utterance boundaries. Adobe Audition was used to create digital .wav files for each utterance in the 100-word sample to be used in the listening task.

Children were administered the two forms of *SIP-CCLP Ver. 5* as described in Chapter 2. An 80-word form of the *TOCS+* sentence intelligibility measure (Hodge, Daniels, & Gotzke, 2009) was administered to each child. Maximum sentence length of the items was adjusted for each child's estimated MLU. For children with age-appropriate expressive language, maximum item length in words was determined by adding one to the child's age up to a

maximum of seven (e.g., child age: 3 years, 2 months, maximum sentence length in test: 4 words). The child was instructed to look at the picture, listen for the pre-recorded model of the utterance and then say the same utterance. A "beep" and the appearance of a frog icon in the upper left-hand corner of the presentation screen cued the child when it was time to say the stimulus item and to signal the examiner that recording has started. Verbal reminders were also used to cue the children to wait for the beep before speaking. Two practice items preceded administration and recording of the test items. Short breaks were provided in the form of computer animations that appeared after every 10 stimulus items. If the examiner was unsure about the recording quality of any item or had any concerns about background noise or examiner voiceover, a second imitation was elicited. Administration of the *TOCS+* sentence intelligibility measure took between 8 and 10 minutes.

Children's productions of the first four sentences of the *Zoo Passage* (Fletcher, 1978) were recorded directly to computer using the *Zoo Passage Recorder* (Gotzke & Hodge, 2011) developed for this study. To reduce the memory load for children, the first four sentences of the *Zoo Passage* were divided into six phrases as shown in Appendix G. Recordings were obtained as described for the *TOCS+* software using pictures and pre-recorded models.

Each child's recordings of the *SIP-CCLP Ver. 5* words, *TOCS+* sentences and *Zoo Passage* phrases were reviewed using Adobe Audition and edited to ensure that any extraneous words and comments made by the child or examiner were removed. If there were multiple productions of the target word, sentence or

phrase, the first production without examiner voiceover or environmental noise interference was saved as the .wav file for playback to listener judges.

Listener tasks.

Intelligibility. The *TOCS+ Recorder-Player Ver. 2.0* (Hodge, Gotzke & Daniels, 2009) software uses "C" files to present the list of sound files for judging and to write listener output files. "C" files were created for the utterances and corresponding sound files in each child's 100-word spontaneous sample. The software presented the utterances to listeners in the same order as in the original sample. The maximum number of times each utterance was played was set to two. Listeners were instructed to type the words that they heard. These appeared on the response screen on the computer monitor. Two practice items (i.e., first two sentences of the *Zoo Passage* (Fletcher, 1978)) were presented at the beginning of the task to familiarize listeners with the task and child's voice. The intelligibility score for the 100-word spontaneous sample was calculated by comparing the words that the listener typed in for each utterance to the orthographic gloss of the utterance in the corresponding "C" file and determining the number of words identified correctly. This value was converted to a percentage. The mean percentage of words identified correctly by the three listeners served as the child's intelligibility score.

The *SIP-CCLP Ver. 5* closed-set response task described in Chapter 2 was administered to listeners. The *SIP-CCLP Ver. 5* software calculated an intelligibility and phonetic accuracy score for each listener and group of three listeners as described in Chapter 3.

The *TOCS+* sentence intelligibility test software (Hodge, Daniels, & Gotzke, 2009) was used to play sentence recordings to listener judges. The software allowed listener judges to hear each utterance a maximum of two times. Listeners were instructed to type the words that they heard. These appeared on the response screen. As with the spontaneous sample open-set response task, listeners judged two practice items before judging the test items. The *TOCS+* software determined the number of words identified correctly for each listener for the *TOCS+* sentence recordings by comparing the listener's responses with the test key. As needed, the researcher hand-corrected the test key for each sentence to match the words that the child actually said for each utterance, compared the test key and listener's responses on the corrected sentences, and recalculated the number of words identified correctly. The percentage of words identified correctly for each listener was recalculated. The mean percentage of words identified correctly by the three listeners served as the child's intelligibility score.

All recordings were presented to listeners using ElectroVoice S-40 compact monitor speakers located in a sound booth. To improve the signal-to-noise ratio, the computer hard drive and Technics Stereo Integrated Amplifier (model SU-V460) were located outside the sound booth. Each listener independently judged the *SIP-CCLP Ver. 5* form 1, *SIP-CCLP Ver. 5* form 2, spontaneous speech sample (session one) and *TOCS+* sentence recordings from four different children in a single one-hour session. Listeners never judged two sets of *SIP-CCLP Ver. 5* recordings consecutively. Order of judging tasks was counterbalanced across listeners. Each set of recordings for a child (e.g., *TOCS+*

sentence recordings, *SIP-CCLP Ver. 5* form 1 and 2 recordings, and 100-word spontaneous speech sample) was judged by three listeners; therefore, a total of 12 different listeners judged each child's recordings. Three independent listeners judged also each child's 100-word spontaneous speech sample obtained from session two. For each set of recordings, playback volume was standardized to 50 – 65 dBA prior to presentation to listeners. Listeners were asked if the playback volume was adequate after the practice items for each task and adjustments were made to the amplification as requested.

SIP-CCLP errors. *SIP-CCLP Ver. 5* analysis software was used to compile and analyze responses from the three listeners for each child's recordings. Prior to analyzing the three listeners' responses, the software checked each listener's response file for entries in the "other/blank" response option. As described in Chapter 2, these entries were presented, item by item, above the target and foil words, for the test administrator to review and either verify or recode as the "target" or "foil." The test administrator examined each entry typed in by the listeners to see if it contained the sound in the contrastive position in the target or foil words. After checking, the software compiled and analyzed the three listener's responses.

SIP-CCLP software determined the error pattern represented by the listeners' choices. Items in which a minimum of two of the three listeners chose the same foil were summarized in the contrast error profiling (foil) section of the analysis print-out. These errors were organized by the five categories of error patterns (i.e., manner preference error (MPE), place preference error (PPE),

voicing error (VE), sibilant error (SE) and cluster error (CE)). Items for which no agreement was obtained among the three listeners, two of the three listeners chose "can't identify" or two of the three listeners typed a response in the "blank" button were listed in the contrast error profiling (other) section of the analysis print-out. The test administrator analyzed items in which two of three listeners typed a response in the "blank" button to determine if the error matched one of the error patterns included in *Ver. 5*. Items which could not be matched were described as "unclassified."

Ratings of hypernasality and voice severity. The edited phrase .wav files for each child's recording of the *Zoo Passage* (Fletcher, 1978) were copied into a single .wav file using *Adobe Audition 1.5* (Adobe Systems Incorporated, 2004). Each passage .wav file was normalized to 80% using the Adobe Audition software to achieve consistent amplitude among *Zoo Passage* samples from the 20 children (Chapman, Hardin-Jones, Goldstein, Halter, Havlik & Schulte, 2008).

A modulus for the hypernasality ratings task and a modulus for the voice severity ratings task were chosen from a clinical database of *Zoo Passage* recordings maintained in the Glenrose Rehabilitation Hospital's Resonance Clinic. The modulus sample chosen was considered the "best example of the midpoint" for each of these speech characteristics (Chapman et al., 2008; p. 301) by the researcher and a speech-language pathologist with over 20 years of experience assessing children's voice and resonance.

Practice samples were obtained from the set of recordings at the Glenrose Rehabilitation Hospital's Resonance Clinic and recordings from four children

without cleft palate, which were obtained during pilot-testing of *SIP-CCLP Ver. 5*. Practice samples for each rating task were chosen to represent the range of hypernasality and voice severity. The samples from the Resonance Clinic (i.e., modulus for hypernasality, modulus for voice severity, and practice samples for both rating tasks) were digitized using a sampling rate of 48 kHz and quantization of 16 bits and normalized to 80% using Adobe Audition. A 5-second interval of silence and the normalized modulus *Zoo Passage* .wav file was appended to each normalized *Zoo Passage* .wav file from the practice children and the 20 children with cleft palate for presentation to the expert raters.

Using direct magnitude estimation (DME) with a modulus, each expert rater (speech-language pathologist) independently rated hypernasality (Chapman et al., 2008) or voice severity (Eadie & Doyle, 2002) of each child's productions of the first four sentences of the *Zoo Passage* (Fletcher, 1978). At the beginning of the rating task, the expert rater assigning DME values for hypernasality was instructed to ignore articulation/phonology and voice severity. The expert rater assigning DME values for voice severity was instructed to ignore articulation/phonology and hypernasality. The expert raters were advised that the modulus represented a value of 100 on the scale and that all samples should be rated relative to the modulus sample. The expert rater assessing hypernasality was instructed that if a sample is judged to be twice as hypernasal as the modulus, it should be rated as 200 and if the sample is judged to be half as hypernasal as the modulus, it should be rated as 50. Similar comparisons were provided in the

instructions for the expert rater assessing voice severity (e.g., if judged to be twice as severe as modulus, rate as 200).

Audio files were played back using *TOCS+ Recorder-Playback Ver. 2.0* software (Hodge, Gotzke & Daniels, 2009) in a sound booth through ElectroVoice S-40 compact monitor speakers. Order of playback of the samples was randomized for each expert rater. Each rater judged the samples independently. The ratings task consisted of six practice samples and 20 "test" samples. The sample to be rated was followed by the modulus sample with an inter-stimulus interval of five seconds. The software allowed expert raters to hear each sample a maximum of two times. The ratings task took about 30 minutes to complete.

For examination of intra-rater reliability, each expert returned two weeks after the initial rating session to rate the recordings a second time. The intraclass correlation coefficients (Shrout & Fleiss, 1979; ICC (Type 1, 1)) were .778 (95% Confidence Interval (CI): .529 - .905) for voice severity and .914 for hypernasality (95% CI [.80, .97]).

Articulation accuracy: Percentage of consonants correct. Each child's 100-word spontaneous speech sample from session one (N = 20) was transcribed phonetically to determine the percentage of consonants correct. In transcribing the samples, the researcher followed guidelines outlined by Shriberg (1986). Diacritics for dentalization, palatalization, lateralization, nasalization and nasal emission were used. The diacritics for devoiced, backed and frictionalized were also used exclusively to transcribe active nasal fricatives (i.e., devoiced /m/ or /n/

with nasal air emission) and pharyngeal fricatives (i.e., backed, frictionalized /k/). Each sample was transcribed independently by a second trained transcriber to determine inter-transcriber agreement for phonetic transcription. Agreement was based on segment-by-segment comparison of the consonants only for the two transcripts. Point-by-point transcription agreement ranged from 61.5% to 82.8% (mean = 71.8, SD =6.3) for narrow transcription (included agreement on diacritics) and from 76.9% to 94.8% (mean = 84.3, SD = 5.2) for agreement that a phoneme was correct or incorrect. The majority of disagreements occurred for instances of transcribing nasal air emission, active nasal fricatives, pharyngeal fricatives, and glottal stops. To resolve disagreements, the two transcribers listened to the audio files a maximum of three times and reviewed the differing transcriptions. If the transcribers agreed on a transcription after the review, the consensus transcription was recorded. If agreement was not obtained, the "benefit of the doubt" procedure was followed whereby the transcription closest to the adult model of the target word was selected (i.e., distortion chosen before substitution). Disagreements were rarely resolved when one transcriber marked an active nasal fricative and the other marked a pharyngeal fricative. As both of these non-standard substitutions are treated as errors in calculation of the percentage of consonants correct, the first transcriber's symbol was used in the consensus transcript. *Programs to Examine Phonetic and Phonologic Evaluation Records* (*P.E.P.P.E.R.*; Shriberg, 1986) were used to determine the percentage of consonants correct (PCC) for the consensus transcript of the 100-word spontaneous speech sample.

Results

Group results for *SIP-CCLP Ver. 5* form 1 and form 2 scores, spontaneous speech sample and *TOCS+* sentence mean intelligibility scores, percentage of consonants correct, hypernasality ratings and voice severity ratings are reported in Table 4-2. Results for each of the 20 child participants on the aforementioned variables are reported in Appendix H.

Concurrent criterion-related validity. Spontaneous speech sample intelligibility scores ranged from 30.4 to 86.3% (mean = 64.0, SD = 17.4) for the 20 children. Inter-rater reliability for the three listeners' intelligibility scores for the 20 children in session one was evaluated by calculating an intraclass correlation coefficient (ICC; Type 1, 3) and standard error of measurement (SEM). The ICC was .98 (95% CI [.96, .99]) and the SEM was 4.32. Moderate positive correlations were found between intelligibility scores on *SIP-CCLP Ver. 5* and the spontaneous speech sample (F1: $r = .61$, $p < .01$; F2: $r = .51$, $p < .05$) and between *SIP-CCLP Ver. 5* phonetic accuracy scores and spontaneous speech intelligibility scores (F1: $r = .64$, $p < .01$; F2: $r = .50$, $p < .05$). Appendix I contains plots of the relationships between *SIP-CCLP* scores and spontaneous speech intelligibility scores.

TOCS+ sentence intelligibility scores ranged from 29.1 to 97.5% (mean = 68.2, SD = 20.3). Inter-rater reliability for the three listeners' intelligibility scores for the 20 children in session one was evaluated by calculating an ICC (Type 1, 3) and SEM. The ICC was .98 (95% CI [.97, .99]) and the SEM was 4.52. Strong positive correlations were found between intelligibility scores on *SIP-CCLP Ver.*

5 and the *TOCS+* sentence intelligibility test (F1: r = .90, p < .01; F2: r = .89, p < .01) and between *SIP-CCLP Ver. 5* phonetic accuracy scores and *TOCS+* sentence test intelligibility scores (F1: r = .88, p < .01; F2: r = .86, p < .01). Appendix J contains plots of the relationships between *SIP-CCLP* scores and *TOCS+* sentence intelligibility scores.

Predictive criterion-related validity. Predictive criterion-related validity was examined for the 14 children who participated in two sessions. For the 14 children, session one *SIP-CCLP* intelligibility scores ranged from 28.0 to 90.5% (mean = 65.31, SD = 19.84) for form 1 and from 32.3 to 95.8% (mean = 68.59, SD = 19.12) for form 2. Session one *SIP-CCLP* phonetic accuracy scores ranged from 20.1 to 85.5% (mean = 55.35, SD = 19.89) for form 1 and from 22.8 to 87.0% (mean = 57.28, SD = 18.55) for form 2. The mean intelligibility score on the 100-word spontaneous speech sample recorded during session two was 65.77% (SD = 13.42; range: 43.14 – 88.14). Moderate positive correlations were found between intelligibility scores on *SIP-CCLP Ver. 5* and the spontaneous speech sample (F1: r = .54, p < .05; F2: r = .52, p < .05) and between *SIP-CCLP Ver. 5* phonetic accuracy scores and spontaneous speech intelligibility scores (F1: r = .61, p < .05; F2: r = .58, p < .05). Appendix K contains plots of the relationships between *SIP-CCLP* scores and session two spontaneous speech intelligibility scores.

Alternate forms reliability was examined for intelligibility scores obtained from the 100-word spontaneous speech sample over the two sessions. This was necessary as previous reports of the stability of this measure could not be located.

For the 14 children who attended two sessions, the mean intelligibility score on the 100-word spontaneous speech sample recorded during session one was 68.39% ((SD = 15.72; range: 40.33 – 86.27); (session two: mean = 65.77%; SD = 13. 42; range: 43.14 – 88.14)). The Pearson's correlation coefficient was .82 ($p < .01$) and the ICC (Type 2, 1) was .81 (95% CI [.51, .94]). The standard error of measurement was 6.33.

Construct-related validity – error patterns. A total of 456 and 384 errors on phonetic contrast items were identified for form 1 and 2, respectively. Number of errors per child ranged from 6 to 49 for form 1 (mean = 22.8, SD = 12.4) and from 2 to 42 for form 2 (mean = 19.2, SD = 11.1). The difference in number of errors between the two forms was four or less for 12 of the 20 children. A minimum of two of three listeners chose the same foil for 249 phonetic contrast items identified as errors in form 1 and 227 items in form 2. A minimum of two of three listeners typed the same response in the "blank" for 83 errors in form 1 and 75 errors in form 2. A minimum of two of three listeners chose "can't identify" for 17 errors in form 1 and 10 errors in form 2. No consensus was reached among listeners on the error response for 107 errors in form 1 and 72 errors in form 2. A breakdown of errors into these four categories for each form is shown in Figure 4-1.

Listener-generated errors could be classified using the error patterns described for *Ver. 5* for 73 errors for form 1 and 68 errors for form 2. The remaining errors identified using the "blank" could not be classified according to the *Ver. 5* error patterns (F1: 10 errors; F2: 7 errors). The unclassified listener-

generated errors for form 1 and 2 are listed in Table 4-3. For form 1, there were two instances where a stop target was identified as fricative, seven instances where a fricative target was identified as a consonant cluster, and one sonorant error. For form 2, there were one instance where a stop target was identified as fricative, three instances where a stop or fricative target was identified as a consonant cluster, one instance where a consonant cluster was identified as a sonorant, and two sonorant errors.

A breakdown of errors into the five error categories for each form is shown in Figure 4-2. A total of 122 manner preference errors (MPE) were identified for form 1, compared to 98 for form 2. Number of manner preference errors ranged from 0 to 21 per child (mean = 6.1, SD = 5.4) for form 1 and from 0 to 23 (mean = 4.9, SD = 5.1) for form 2. Manner preference errors organized by error subtype are listed in Table 4-4. A total of 91 place preference errors (PPE) were identified for form 1, compared to 78 for form 2. Number of place preference errors per child ranged from 0 to 22 (mean = 4.7, SD = 5.3) for form 1 and from 0 to 19 (mean = 3.9, SD = 4.9) for form 2. Place preference errors organized by error subtype are listed in Table 4-5. A total of 37 voicing errors (VE) were identified for form 1, compared to 45 for form 2. Number of voicing errors per child ranged from 0 to 6 (mean = 1.9, SD = 1.4) for form 1 and from 0 to 7 (mean = 2.3, SD = 1.7) for form 2. Voicing errors organized by error subtype are listed in Table 4-6. A total of 46 sibilant errors (SE) were identified for form 1, compared to 52 for form 2. Number of sibilant errors per child ranged from 0 to 5 (mean = 2.3, SD = 1.5) for form 1 and from 0 to 6 (mean = 2.6, SD = 1.9) for

form 2. A total of 23 cluster errors (CE) were identified for form 1, compared to 20 for form 2. Number of cluster errors per child ranged from 0 to 5 (mean = 1.2, SD = 1.4) for form 1 and from 0 to 5 (mean = 1.1, SD = 1.5) for form 2. Sibilant and cluster errors organized by error subtype are listed in Tables 4-7 and 4-8, respectively.

Construct-related validity – speech variables. *SIP-CCLP Ver. 5* intelligibility scores ranged from 28.0 to 90.5% for form 1 (mean = 63.8, SD = 18.5) and from 32.3 to 95.8% for form 2 (mean = 68.7, SD = 17.1). *SIP-CCLP Ver. 5* phonetic accuracy scores ranged from 20.1 to 85.5% for form 1 (mean = 54.0, SD = 18.4) and from 22.8 to 87.0% for form 2 (mean = 57.7, SD = 17.1). DME ratings for hypernasality ranged from 40 to 300 (mean = 158.8, SD = 77.8). Moderate negative correlations were found between hypernasality ratings and *SIP-CCLP Ver. 5* intelligibility scores (F1: $r = -.71$, $p < .01$; F2: $r = -.65$, $p < .01$) and *SIP-CCLP Ver. 5* phonetic accuracy scores (F1: $r = -.76$, $p < .01$; F2: $r = -.70$, $p < .01$). Appendix L shows the relationships between hypernasality ratings and *SIP-CCLP* scores. DME ratings for voice severity ranged from 10 to 100 (mean = 47.9, SD = 46.2). Weak correlations were found between voice severity ratings and *SIP-CCLP Ver. 5* intelligibility scores (F1: $r = .007$, $p > .05$; F2: $r = .05$, $p > .05$) and *SIP-CCLP Ver. 5* phonetic accuracy scores (F1: $r = .004$, $p > .05$; F2: $r = .10$, $p > .05$). The percentage of consonants correct for the 100-word spontaneous sample ranged from 27.7 to 82.0% (mean = 50.4, SD = 14.2). Moderately strong positive correlations were found between percentage of consonants correct and *SIP-CCLP Ver. 5* intelligibility scores (F1: $r = .72$, $p< .01$; F2: $r = .73$, $p < .01$)

and *SIP-CCLP Ver. 5* phonetic accuracy scores (F1 and F2: $r = .76$, $p < .01$). Appendix M contains plots of the relationships between percentage of consonants correct and *SIP-CCLP* scores.

Standardized and unstandardized coefficients for form 1 and 2 are reported in Tables 4-9 and 4-10 for intelligibility and phonetic accuracy scores, respectively. Using the stepwise method, a significant model with two predictors (percentage of consonants correct, hypernasality ratings) emerged for intelligibility scores for form 1 (Adjusted $R^2 = .60$, $F(2, 17) = 15.12$) and for phonetic accuracy scores for both forms (F1: Adjusted $R^2 = .69$, $F(2, 17) = 22.08$, $p < .01$; F2: Adjusted $R^2 = .64$, $F(2, 17) = 17.50$, $p < .01$). A significant model with one predictor (percentage of consonants correct) emerged for form 2 intelligibility scores (Adjusted $R^2 = .51$, $F(1, 18) = 20.54$, $p < .01$). Voice severity was never a significant predictor of *SIP-CCLP Ver. 5* scores.

Zero-order correlations for all session one variables are reported in Table 4-11. In all models, *SIP-CCLP* score, hypernasality ratings and voice severity ratings were possible predictors. Using stepwise multiple regression, significant models with one predictor (*SIP-CCLP Ver. 5* score) emerged for session one spontaneous speech intelligibility scores, regardless of which form or *SIP-CCLP Ver.* 5 score was used. For form 1, the adjusted R^2 was .34 ($F(1, 18) = 10.70$, $p < .01$) when *SIP-CCLP Ver. 5* intelligibility score was the dependent variable and the adjusted R^2 was .37 ($F(1,18) = 12.29$, $p < .01$) when *SIP-CCLP Ver. 5* phonetic accuracy score was the dependent variable. For form 2, the adjusted R^2 was .22 ($F(1, 18) = 6.22$, $p < .05$) when *SIP-CCLP Ver. 5* intelligibility score was

the dependent variable. A significant model with two predictors (*SIP-CCLP Ver. 5* form 2 phonetic accuracy score, voice severity ratings) emerged for session one spontaneous speech intelligibility scores (Adjusted $R^2 = .34$, $F(2,17) = 5.82$, $p < .05$). Hypernasality ratings were never significant predictors of spontaneous speech intelligibility scores.

Using stepwise multiple regression, significant models with one predictor (*SIP-CCLP Ver. 5* form 1 score) emerged for *TOCS+* sentence intelligibility scores (intelligibility: Adjusted $R^2 = .80$, $F(1,18) = 79.04$, $p < .01$; phonetic accuracy: Adjusted $R^2 = .77$, $F(1, 18) = 64.38$, $p < .01$). A significant model with one predictor (*SIP-CCLP Ver. 5* form 2 intelligibility score) emerged for *TOCS+* sentence intelligibility scores (Adjusted $R^2 = .79$, $F(1,18) = 70.68$, $p < .01$). A significant model with two predictors (*SIP-CCLP Ver. 5* form 2 phonetic accuracy score, voice severity ratings) emerged for *TOCS+* sentence intelligibility scores (Adjusted $R^2 = .78$, $F(2,17) = 34.56$, $p < .01$). Hypernasality ratings were never significant predictors of *TOCS+* sentence intelligibility scores.

Discussion

The purpose of this study was to evaluate the validity (criterion and construct-related) of *SIP-CCLP Ver. 5.* In the evaluation of both concurrent and predictive criterion-related validity, Pearson's correlation coefficients were greater than 0.6 for intelligibility scores obtained from *SIP-CCLP* and a 100-word spontaneous speech sample and greater than 0.85 for *SIP-CCLP* scores and intelligibility scores from an imitative sentence measure (*TOCS+*). These results suggest that while *SIP-CCLP* is measuring the same construct (i.e., intelligibility)

as both criterion measures, there is a difference in intelligibility scores when the speech sample is elicited imitatively (i.e., *TOCS+*, *SIP-CCLP*) or spontaneously. Listeners identified errors in all five error categories in *SIP-CCLP Ver. 5*. In the evaluation of construct-related validity, moderately positive significant correlations were found between *SIP-CCLP* scores and hypernasality ratings, and between *SIP-CCLP* scores and PCC, suggesting that both hypernasality and articulation contribute to, but are not the same as, intelligibility. *SIP-CCLP* scores were the only significant predictor of intelligibility measured by the criterion variables (100-word spontaneous sample, *TOCS+* sentence) for all but two of the stepwise regression models, confirming that all three are measuring the same construct.

Criterion-related validity was examined by evaluating the relationships between *SIP-CCLP* scores and intelligibility scores obtained from a 100-word spontaneous speech sample recorded on the same day (concurrent validity) and on a different day (predictive validity). In all cases, moderate positive correlations were found between *SIP-CCLP Ver. 5* form 1 and 2 scores and intelligibility scores obtained from a 100-word spontaneous speech sample. In Gotzke (2005), the correlation between intelligibility scores from *Ver. 3* and the 100-word spontaneous speech sample recorded on the same day was lower than the correlation between phonetic accuracy scores and intelligibility scores from the 100-word spontaneous sample ($r = .39$ and $r = .56$, respectively). The stronger correlation for *Ver. 5* and criterion variable intelligibility scores in this study suggests that the changes to the closed-set response task outlined in Chapter 2

made *Ver. 5* more sensitive to differences in intelligibility among children. The correlation between intelligibility scores on imitated words (*SIP-CCLP*) and spontaneous speech is lower in this study than the correlation between intelligibility scores on the *CSIM* (Wilcox & Morris, 1999) and 100-word spontaneous speech samples (r = .79) reported by Gordon-Brannan and Hodson (2000) for 48 children ranging in age from four years to five years, six months. In the current study, the mean difference between intelligibility scores obtained using imitated words and spontaneous speech was smaller for the nine children with spontaneous sample intelligibility scores greater than 68% (mean difference = 7.29 (form 1); 9.28 (form 2)) than for the eleven children with spontaneous speech intelligibility scores less than 68% (mean difference = 15.14 (form 1); 18.52 (form 2)). In Gordon-Brannan & Hodson (2000), 36 of the 48 children had spontaneous speech intelligibility scores greater than 68%. It is possible that intelligibility scores obtained using imitated words and intelligibility scores obtained using a spontaneous speech sample were more similar for these 36 children than for the 12 children with spontaneous speech intelligibility scores less than 68%, which would have led to an higher correlation between scores than was found in the current study.

Criterion-related validity was also examined by evaluating the relationships between *SIP-CCLP* scores and *TOCS+* sentence intelligibility scores. The strong positive correlations found between *SIP-CCLP Ver. 5* intelligibility scores and *TOCS+* sentence intelligibility scores (r = .9) were slightly higher than the correlation between imitated words and sentences reported

by Gordon-Brannan and Hodson (2000, r = .81). The correlations were also higher than those found between *SIP-CCLP* scores and intelligibility scores obtained from the spontaneous speech sample, suggesting that children may use strategies (e.g., speak louder, more precise articulation) to improve the intelligibility of their message when imitating models, that they don't use in conversation.

The error patterns identified using both forms of *SIP-CCLP Ver. 5* were described as part of the evaluation of construct-related validity. The majority of errors were identified when a minimum of two of the three listeners chose the same foil. However, listeners frequently used the "blank" to type in an alternative response. Listener-generated errors accounted for 18.2% and 19.5% of the error patterns identified for the children with cleft palate on form 1 and 2, respectively. In Gotzke (2005), listener-generated errors accounted for 42.9% of the error patterns identified for the 15 children with cleft palate. The increased number of real-word response alternatives (i.e., one target and three foils) and error patterns tested in *Ver. 5* appears to have captured more the of the error patterns present in the speech of children with cleft palate, reducing listeners' use of the "blank" to indicate what they heard. The majority of the listener-generated errors were additional instances of error patterns tested in *Ver. 5*. Evaluation of the reliability of the classification of the listener-generated errors is straightforward for a user familiar with the error patterns sampled in *Ver. 5*. However, inter-rater reliability of the classification of the listener generated errors for this study will be conducted prior to submitting this chapter for publication. The listener-generated

errors included four instances in form 2 in which a stop was identified as a glottal fricative (2 children) and one instance in form 2 in which a sonorant was deleted from an obstruent-sonorant consonant cluster. Listeners used foils to identify an additional two instances in form 1 of the former (2 children) and one instance of the latter. Both of these patterns were considered for exclusion from *Ver. 5* after expert review of the content-related validity of the error patterns. Based on these results, it is recommended that the place preference error pattern in which a stop is identified as a glottal fricative be included in the next revision of *SIP-CCLP* and a foil targeting a different error pattern be identified for two stimulus items in each form where deletion of a sonorant from an obstruent-sonorant consonant cluster was targeted (i.e., F1: trail, drip; F2: trip, dry). There were 17 errors in form 1 and 10 errors in form 2 in which a minimum of two of the three listeners chose "can't identify." The majority of these errors in both forms were identified for two children (i.e., CP11: 8 errors, CP12: 13 errors). These children had the lowest *SIP-CCLP* intelligibility scores on form 1 of all the children in this study. Phonetic transcription of these items would reveal whether listeners were using the "can't identify" option for words where the child substituted a non-English (e.g., active nasal fricative) sound for the target. No consensus was obtained for 23.5% and 18.8% of the errors in form 1 and 2, respectively. These items may be additional instances of non-English substitutions and distortions that were difficult to identify.

The majority of errors identified with both forms were in the manner preference error category (F1: 38.2%; F2: 33.4%). Re-examination of the errors identified for the 15 children with cleft palate in the evaluation of *Ver. 3* (Gotzke, 2005) using the *Ver. 5* error categories revealed that the majority of errors were in the manner preference error (28.9%) and place preference error categories (27.9%). Whitehill and Chau (2004) reported that three error patterns (i.e., oral stop versus nasal, stop versus fricative, stop versus affricate) accounted for 42.4% of the total errors for their group of 15 Cantonese speakers with repaired cleft lip and palate who ranged in age from five to 44 years. These three patterns are represented in the manner preference error category in *SIP-CCLP Ver. 5*. Overall, more errors were identified using form 1 than form 2, suggesting that the two forms are not equivalent.

The error patterns identified for the children with cleft palate in this study were compared to the error patterns identified by experts as occurring in more than 10% of children with cleft palate who also have a speech sound disorder (described in Chapter 2 and highlighted in Tables 4-4 to 4-8). Of the eleven error patterns identified by experts, nine in form 1 (Nasals for obstruents (MPE), Stopping (MPE), Gliding of liquids (MPE), Velar stops for obstruents (PPE), Glottal stops for oral stops (PPE), Glottal fricatives for oral fricatives (PPE), Palatal fricatives for alveolar fricatives (SE), Fronting (SE), Deletion of an obstruent from an obstruent-obstruent cluster (CE)) and seven in form 2 (as for form 1 except Glottal stops for oral stops (PPE) and Glottal fricatives for oral fricatives (PPE)) were identified in more than 10% of the children in this study.

Two error patterns in form 1 and four error patterns in form 2 were identified less frequently than expected based on results of the review of content-related validity (i.e., F1 and F2: glottal stops for affricates (PPE), backing and cluster reduction (CE); F2 only: glottal stops for oral stops (PPE), glottal fricatives for oral fricatives (PPE)). Glottal stops for affricates and backing and cluster reduction were not identified for any of the 15 children with cleft palate in the evaluation of *Ver. 3* (Gotzke, 2005). These results suggest that these two patterns are not common in the speech of children with cleft palate in Alberta. A multi-center international study is needed to evaluate if there are regional differences in the occurrence of different error patterns related to variation in surgical timing and treatment and access to services in different countries. The small sample size may also account for the differences between this study and experts' ratings. Of the 35 error patterns included in *SIP-CCLP Ver. 5*, 24 patterns in form 1 and 20 patterns in form 2 were identified for more than two children in this study, indicating that *Ver. 5* is sampling the error patterns of young English-speaking children with cleft palate.

To examine construct-related validity, the relationships between *SIP-CCLP* scores and measures of other speech variables (resonance, voice, and articulation) were evaluated. As expected, moderate negative relationships between hypernasality ratings and *SIP-CCLP* scores were found. Investigations into the relationship of resonance to intelligibility for speakers with cleft palate has yielded mixed results with some researchers finding significant correlations between resonance and intelligibility (e.g., Keuning et al., 2002) and others

finding nonsignificant correlations (e.g., Whitehill & Chun, 2002). Like this study, intelligibility scores in Whitehill and Chun (2002) were obtained using closed-set response task developed using a phonetic contrast approach (Kent et al., 1989). Differences in how hypernasality was rated may account for the conflicting results found in this study and Whitehill and Chun (2002). In Whitehill and Chun (2002), nasality ratings were obtained using a 7-point equal-appearing interval (EAI) scale from three speech-language pathologists who participated in a training session focusing on judging resonance prior to completing their ratings. EAI scaling is not considered to be a valid method of evaluating hypernasality (Whitehill, Lee & Chun, 2002). The correlations between hypernasality ratings and *Ver. 5* phonetic accuracy scores for both forms were slightly higher than the correlations between hypernasality ratings and *Ver. 5* intelligibility scores for both forms, suggesting that listeners may be using the "distorted" rating to capture the effects of hypernasality (e.g., muffled production of oral sounds) on children's word productions.

The relationships of voice severity ratings to *SIP-CCLP* scores were weak and not significant, suggesting that voice severity does not account for the variance in speech intelligibility obtained using an imitative word measure. The majority of children in this study did not appear to have a voice disorder as voice severity was rated as more severe than the modulus for only two of the 20 children and the median voice severity rating was 30.

As expected, strong positive relationships between percentage of consonants correct obtained from the spontaneous speech sample and *SIP-CCLP*

scores were found. This result is consistent with previous examinations of the relationship between articulation accuracy and intelligibility for children with cleft palate (e.g., Zajac et al., 2011; Whitehill & Chun, 2002). The correlations between percentage of consonants correct and *Ver. 5* intelligibility scores were similar to the correlations between percentage of consonants correct and *Ver. 5* phonetic accuracy scores.

The results of the stepwise regression model of *SIP-CCLP Ver. 5* scores indicated that percentage of consonants correct contributed consistently to the variance in speech intelligibility measured using an imitative single word measure for young children with cleft palate. Except for the model for form 2 intelligibility scores, hypernasality rating also contributed significant unique variance to *Ver. 5* scores. Magnus, Hodson and Schommer-Aikins (2011) found that a measure of articulation (i.e., phonological deviation average) and resonance were significant predictors of intelligibility ratings of spontaneous speech. In the examination of the predictive relationship of nine speech variables (i.e., palatalization, lateralization of /s/, backing, glottal articulation, hyperkinetic voice, hypernasality, nasal emission, nasal fricative and correctness of articulation) to intelligibility of spontaneous speech for toddlers with and without cleft palate by Konst et al. (2003), three measures of articulation (i.e., correctness of articulation, lateralization and backing) accounted for 93% of the variance in intelligibility scores. The results of these three studies confirm that articulation is a key correlate of intelligibility for children with cleft palate. The varying role of hypernasality (or resonance) in the three models may be related to age. The

children in this study and Magnus et al. (2011) were three years and older, while the children in Konst et al. (2003) were between 30 and 32 months.

Construct validity was also examined by modeling the relationships of speech variables (i.e., hypernasality and voice severity) and *SIP-CCLP Ver. 5* scores to intelligibility scores obtained from the two criterion measures. *SIP-CCLP Ver. 5* score was the only significant predictor of spontaneous speech and imitative sentence intelligibility for all but one regression model. Unexpectedly, when form 2 phonetic accuracy scores were used as a predictor, voice severity became a significant predictor of *TOCS+* sentence intelligibility scores. Between 25 and 41% of the variance in spontaneous speech intelligibility scores was explained by *SIP-CCLP Ver. 5* scores., while between 77 and 81% of the variance in *TOCS+* sentence intelligibility scores was explained by *SIP-CCLP Ver. 5* scores. This supports the results of previous studies that concluded that intelligibility is affected by how speech samples are obtained (e.g., Gordon-Brannan & Hodson, 2000).

The results of this evaluation are interpreted as support for the criterion and construct-related validity of *SIP-CCLP Ver. 5*. Comparison of the group of children in this study with age-similar children without cleft palate would provide additional support for the construct-related validity of *SIP-CCLP Ver. 5*.

Table 4-1

Comparison of Validity Evidence for Children's Intelligibility Measures

	Published Measures of Speech Intelligibility		Intelligibility Measures for English-speaking Children with Cleft Palate		
Validity Evidence	*CID Picture SPINE* (Monsen, Moog & Geers, 1988)	*CSIM* (Wilcox & Morris, 1999)	Zajac et al. (2011)	*SIP-CCLP Ver. 3* (Gotzke, 2005)	*SIP-CCLP Ver. 5* (current study)
Criterion-related					
Concurrent	✕	intelligibility ratings (conversational speech) r = .52 (younger) r = .46 (older)	✕	intelligibility scores (conversational speech) r = .39 (intelligibility score) r = .56 (phonetic accuracy score)	intelligibility scores 1. conversational speech 2. imitated sentences
Predictive	intelligibility scores (imitated sentences) r = .91 (examiner a) r = .96 (examiner b)	✕	✕	✕	intelligibility scores (conversational speech)

Table 4-1 continued

Validity Evidence	Published Measures of Speech Intelligibility		Intelligibility Measures for English-speaking Children with Cleft Palate		
	CID Picture SPINE (Monsen, Moog & Geers, 1988)	*CSIM* (Wilcox & Morris, 1999)	Zajac et al. (2011)	*SIP-CCLP Ver. 3* (Gotzke, 2005)	*SIP-CCLP Ver. 5* (current study)
Construct-related					
Convergent	✗	Number of errors on *Goldman-Fristoe Test of Articulation* r = -.63 (younger) r = -.55 (older)	Percentage of consonants correct (single words) r = .78	*Ver. 3* error patterns	*Ver. 5* error patterns Percentage of consonants correct (conversational speech) Hypernasality ratings (imitated passage) Voice severity ratings (imitated passage)
Group Comparison	✗	✗	✓	✓	✗

Note. ✗ = not completed; ✓ = completed.

Table 4-2

Mean, Standard Deviation, Minimum and Maximum Scores for SIP-CCLP Ver. 5 Forms 1 and 2, 100-word Spontaneous Speech Sample and TOCS+ Sentence Intelligibility Test, Hypernasality, Voice Severity, and Percentage of Consonants Correct (N = 20)

Measure		Mean	SD	Minimum	Maximum
SIP-CCLP Ver. 5 Intelligibility Score (%)	F1	63.8	18.5	28.0	90.5
	F2	68.7	17.1	32.3	95.8
SIP-CCLP Ver. 5 Phonetic Accuracy Score	F1	54.0	18.4	20.1	85.5
(%)	F2	57.7	17.1	22.8	87.0
Spontaneous Speech Intelligibility Score (%)		64.0	17.4	30.4	86.3
TOCS+ Sentence Intelligibility Score (%)		68.2	20.3	29.1	97.5
Hypernasality Rating		158.8	77.8	40	300
Voice Severity Rating		47.9	46.2	10	200
Percentage of Consonants Correct (%)		50.4	14.2	27.7	82.0

Table 4-3

Unclassified Errors Identified for the Children with Cleft Palate

Form	Description of Error Pattern	Listener Identified Error Pattern	Child ID
1	Stop identified as a fricative	bat – that	CP17
		trail – swail	CP17
	Fricative identified as a consonant cluster	thick – slick	CP03
		V – free	CP03
		zee – bree	CP03
		sick – stick	CP08
		she – ski	CP15
		shop – slop	CP15
		zip – skip	CP15
	Sonorant error	lock – rap	CP03
2	Stop identified as a fricative	trip – srip	CP17
	Stop identified as a consonant cluster	guy – sky	CP14
	Fricative identified as a consonant cluster	zap – slap	CP02
		shy – try	CP09
	Consonant cluster identified as a sonorant	slap - rap	CP04
	Sonorant Error	yell – lell	CP20
		low – though	CP17

Table 4-4

Manner Preference Errors Identified for the Children with Cleft Palate

	Form 1			Form 2		
	Number of Times Identified		Number of Children	Number of Times Identified		Number of Children
	Foil	Listener-Generated		Foil	Listener-Generated	
Glides for obstruents	4	0	2	5	1	5
Liquids for obstruents	2	1	1	2	1	3
Nasals for obstruents[1]	22	5	11	16	6	6
Nasals for liquids	4	2	4	1	0	1
Stopping[1]	19	11	9	17	12	7
Deaffrication	6	1	3	7	0	5
Gliding of liquids[1]	30	1	14	20	0	10
Oral fricatives for liquids and glides	3	0	1	0	0	0
Oral stops for nasals	2	0	2	1	0	1
Affricates for oral stops	1	2	3	3	0	3
Affricates for fricatives	6	0	5	5	1	6
TOTAL	99	23		77	21	

Note. [1]Error identified by experts as occurring in more than 10% of children with cleft palate with a speech sound disorder.

Table 4-5

Place Preference Errors Identified for the Children with Cleft Palate

	Form 1			Form 2		
	Number of Times Identified		Number of Children	Number of Times Identified		Number of Children
	Foil	Listener-Generated		Foil	Listener-Generated	
Bilabial stops for alveolar stops	3	4	5	1	3	3
Velar stops for obstruents[1]	15	13	11	13	8	8
Glottal stops for oral sounds						
a. stops[1]	13	0	8	7	0	1
b. fricative	0	3	1	4	0	1
c. affricates[1]	0	0	0	0	0	0
d. sonorants	8	0	8	0	0	0
Glottal fricatives for oral sounds						
a. fricatives[1]	6	13	3	5	10	2
b. affricates	2	1	3	2	1	2
Alveolar stops for velar stops	7	1	5	7	4	6
Alveolar stops for bilabial stops	0	1	1	0	1	1
Alveolar fricatives for labiodental and interdental fricatives	1	0	1	5	1	5
TOTAL	55	36		46	32	

Note. [1]Error identified by experts as occurring in more than 10% of children with cleft palate with a speech sound disorder.

Table 4-6

Voicing Errors Identified for the Children with Cleft Palate

	Form 1			Form 2		
	Number of Times Identified		Number of Children	Number of Times Identified		Number of Children
	Foil	Listener-Generated		Foil	Listener-Generated	
Voiced for voiceless						
a. stops	3	0	3	1	0	1
b. fricative	2	0	2	3	0	3
c. affricates	0	0	0	0	0	0
Voiceless for voiced						
a. stops	17	0	12	19	0	14
b. fricative	11	0	9	19	1	13
c. affricates	4	0	4	2	0	2
TOTAL	37	0		44	1	

Table 4-7

Sibilant Errors Identified for the Children with Cleft Palate

	Form 1			Form 2		
	Number of Times Identified		Number of Children	Number of Times Identified		Number of Children
	Foil	Listener-Generated		Foil	Listener-Generated	
Palatal fricative for alveolar fricatives[1]	4	2	4	5	3	6
Labiodental fricatives for alveolar fricatives	10	1	8	17	0	11
Addition of a nasal following a sibilant	3	0	3	1	1	2
Fronting[1]	23	3	16	22	3	14
TOTAL	40	6		45	7	

Note. [1]Error identified by experts as occurring in more than 10% of children with cleft palate with a speech sound disorder.

Table 4-8

Cluster Errors Identified for the Children with Cleft Palate

	Form 1			Form 2		
	Number of Times Identified		Number of Children	Number of Times Identified		Number of Children
	Foil	Listener-Generated		Foil	Listener-Generated	
Deletion of an obstruent from an obstruent-obstruent cluster[1]	8	5	8	7	2	5
Deletion of an obstruent from an obstruent-sonorant cluster	6	2	5	6	4	6
Backing and cluster reduction[1]	1	1	2	2	0	2
TOTAL	15	8		15	5	

Note. [1]Error identified by experts as occurring in more than 10% of children with cleft palate with a speech sound disorder.

Table 4-9

Unstandardized and Standardized Coefficients for Predicting SIP-CCLP Ver. 5 Intelligibility Scores from Percentage of Consonants Correct, Hypernasality Ratings, and Voice Severity Ratings

		Unstandardized Coefficients		Standardized Coefficient
Form	Predictor Variables	B	SE B	β
1	Step 1			
	Constant	16.76	11.12	
	PCC	0.93	0.21	.719**
	Step 2			
	Constant	49.63	16.82	
	PCC	0.605	0.233	.466*
	Hypernasality ratings	-0.103	0.04	-.433*
2	Step 1			
	Constant	24.48	10.12	
	PCC	0.877	0.194	.73**

Note. PCC = percentage of consonants correct; * = significant at 0.05; ** = significant at 0.01.

Table 4-10

Unstandardized and Standardized Coefficients for Predicting SIP-CCLP Ver. 5 Phonetic Accuracy Scores from Percentage of Consonants Correct, Hypernasality Ratings, and Voice Severity Ratings

		Unstandardized Coefficients		Standardized Coefficient
Form	Predictor Variables	B	SE B	β
1	Step 1			
	Constant	82.36	6.40	
	Hypernasality ratings	-0.18	0.036	-.757**
	Step 2			
	Constant	41.00	14.69	
	Hypernasality ratings	-0.11	0.04	-.479**
	PCC	0.61	0.20	.476**
2	Step 1			
	Constant	11.65	9.64	
	PCC	0.913	0.184	.759**
	Step 2			
	Constant	38.57	14.85	
	PCC	0.64	0.21	.536**
	Hypernasality ratings	-0.08	0.04	-.38*

Note. PCC = percentage of consonants correct; * = significant at 0.05; ** = significant at 0.01.

Table 4-11

Zero Order Correlation Coefficients for Speech Variables

		SIP-CCLP Intelligibility (%)		*SIP-CCLP* PA (%)		Spontaneous Intelligibility (%)	*TOCS+* Intelligibility (%)	Hypernasality	Voice Severity
		F1	F2	F1	F2				
SIP-CCLP Intelligibility (%)	F2	.948**							
SIP-CCLP PA (%)	F1	.985**	.938**						
	F2	.932**	.975**	.939**					
Spontaneous Intelligibility (%)		.611**	.507*	.637**	.500*				
TOCS+ Intelligibility (%)		.902**	.893**	.884**	.864**	.756**			
Hypernasality		-.705**	-.650**	-.757**	-.696**	-.449*	-.606**		
Voice Severity		.007	.045	.004	.100	-.344	-.150	.231	
Percentage of Consonants Correct (%)		.719**	.730**	.756**	.759**	.697**	.746**	-.584**	-.113

Note. * = significant at 0.05 (1-tailed); ** = significant at 0.01 (1-tailed).

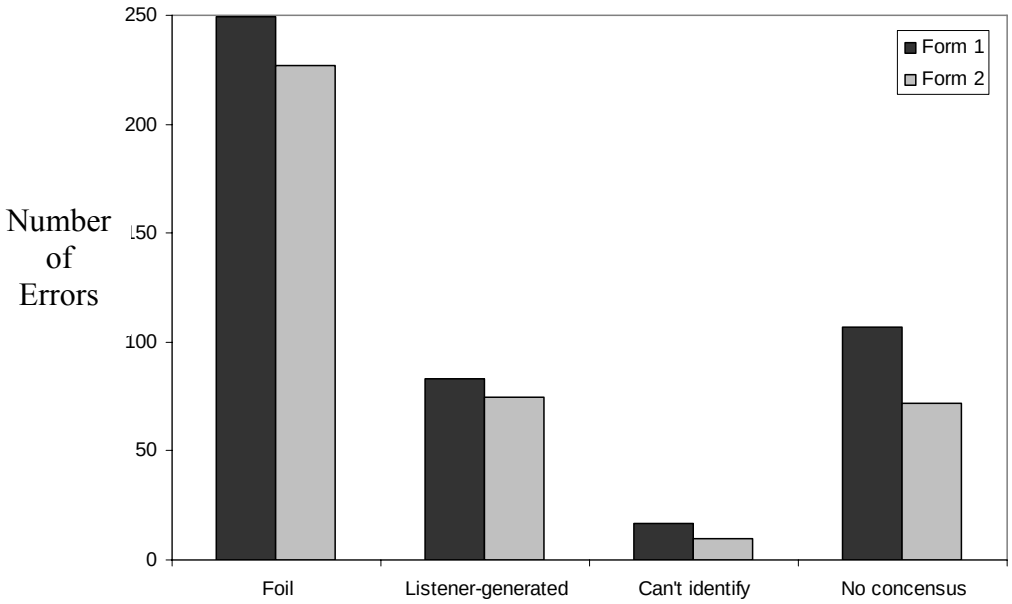

Figure 4-1. Frequency of errors by listener response.

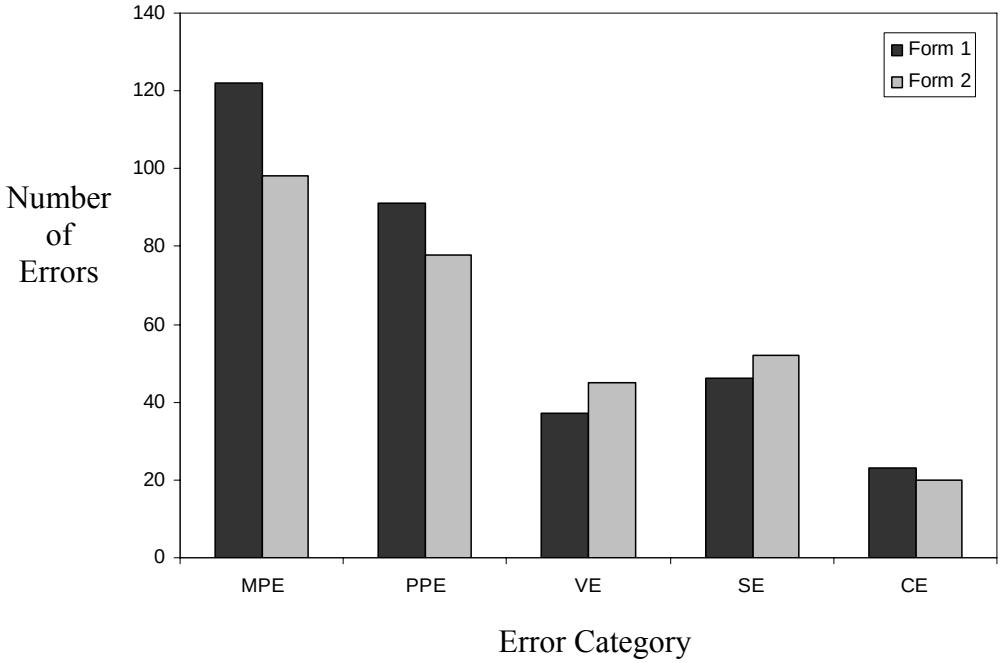

Figure 4-2. Frequency of errors by error category.

References

Adobe Systems Incorporated. (2004). *Adobe Audition 1.5*. San Jose, CA; Adobe Systems Incorporated.

Alberta College of Speech-Language Pathologists and Audiologists. (2008). *Hearing screening guidelines*. Retrieved from http://www.acslpa.ab.ca/public/data/documents/ACFC3D5.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America, 112(6)*, 3014-3021.

Fletcher, S. G. (1978). *Diagnosing speech disorders from cleft palate.* New York, NY: Grune & Stratton.

Goldman, R., & Fristoe, M. (1986). *Goldman-Fristoe Test of Articulation* . Circle Pines, MN: American Guidance Service.

Gordon-Brannan, M., & Hodson, B. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology, 9*, 141-150.

Gotzke, C. L. (2005). *Speech intelligibility probe for children with cleft palate version 3: Assessment of reliability and validity*. (Unpublished master's thesis). University of Alberta, Edmonton, AB.

Gotzke, C. L., & Hodge, M. M. (2011). *Zoo Passage Recorder* [computer software]. Edmonton, AB: University of Alberta.

Hodge, M. M., Daniels, J., & Gotzke, C.L. (2009). *TOCS+ Intelligibility Measures* [computer software]. Edmonton, AB: University of Alberta.

Hodge, M., & Gotzke, C. (2010). Stability of intelligibility measures for children with dysarthria and cerebral palsy. *Journal of Medical Speech Language Pathology, 18(4)*, 61-65

Hodge, M. M., Gotzke, C. L., & Daniels, J. (2009). *TOCS+ Recorder-Player Ver. 2.0* [computer software]. Edmonton, AB: University of Alberta.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Keuning, K. H., Wieneke, G. H., Van Wijngaarden, H. A., & Dejonckere, P. H. (2002). The correlation between nasalance and a differentiated perceptual rating of speech in Dutch patients with velopharyngeal insufficiency. *Cleft Palate-Craniofacial Journal, 39(3)*, 277-284.

Konst, E. M., Rietveld, T., Peters, H. F. M., & Weersink-Braks, H. (2003). Use of a perceptual evaluation instrument to assess the effects of infant orthopedics on the speech of toddlers with cleft lip and palate. *Cleft Palate-Craniofacial Journal, 40*(6), 597-605.

Magnus, L. C., Hodson, B. W., & Schommer-Aikins, M. (2011). Relationships of speech-related and nonspeech variables to speech intelligibility of children with palatal and lip anomalies. *Canadian Journal of Speech-Language Pathology and Audiology, 35(1)*, 32-39.

McWilliams, B. J. (1954). Some factors in the intelligibility of cleft-palate speech. *Journal of Speech and Hearing Disorders, 19*, 524-527.

Moller, K. T., & Starr, C. D. (1984). The effects of listening conditions on speech ratings obtained in a clinical setting. *Cleft Palate Journal*, *21(2)*, 65-69.

Monsen, R., Moog, J. S., & Geers, A. E. (1988). *CID Picture SPINE SPeech Intelligibility Evaluation*. St. Louis, MO: Central Institute for the Deaf.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In: R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement and management* (pp. 119-155). Amsterdam, NL: John Benjamins.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11,* 193-205.

Shriberg, L. (1986). *Programs to Examine Phonetic and Phonological Evaluation Records (P.E.P.P.E.R.) Version 4.0.* Madison, WI: University of Wisconsin-Madison.

Shriberg, L. D., & Kwiatkowski, J. (1985). Continuous speech sampling for phonologic analyses of speech-delayed children. *Journal of Speech and Hearing Disorders, 50*, 323-334.

Shriberg, L. D., Kwiatkowski, J., & Rasmussen, C. (1990). *Prosody-Voice Screening Profile (PVSP): Scoring forms and training manual*. Tucson, AZ: Communication Skill Builders.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2)*, 420-428.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford, UK: Oxford University Press.

Whitehill, T. (2002). Assessing intelligibility in speakers with cleft palate: A critical review of the literature. *Cleft Palate-Craniofacial Journal, 39(1)*, 50-58.

Whitehill, T. L., & Chau, C. H. (2004). Single-word intelligibility in speakers with repaired cleft palate. *Clinical Linguistics and Phonetics, 18(4-5)*, 341-355.

Whitehill, T., & Chun, J. C. (2002). Intelligibility and acceptability of speakers with cleft palate. In: F. Windsor, M. L. Kelly, & N. Hewlett (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 405-415). Mahwah, NJ: Lawrence Erlbaum Associates.

Whitehill, T. L., Lee, A.S, & Chun, J. C. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language, and Hearing Research, 45*, 80-88.

Whitehill, T. L., & Wong, C.Y. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14(4)*, 335-341.

Wilcox, K., & Morris. S. (1999). *Children's Speech Intelligibility Measure*. San Antonio, TX: The Psychological Corporation.

Zajac, D. J., Plante, C., Lloyd, A., & Haley, K. L. (2011). Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate. *Cleft Palate-Craniofacial Journal, 48(5)*, 538-549.

Chapter 5

Effects of Repeated Exposure to *SIP-CCLP* Stimuli Spoken by Children with Cleft Palate

Introduction

Listener gender, training, age, and familiarity with the speaker, the speech characteristics of the disordered population, and the test stimuli, are all variables that may influence listener perceptions of intelligibility (Walshe, Miller, Leahy & Murray, 2008). Barreto and Ortiz (2008) reviewed past research on the effect of listener gender, listener familiarity with speakers and with impaired speech in general on intelligibility scores. They concluded that gender and listener familiarity with speakers did not influence intelligibility scores. However, there was a lack of agreement among the studies reviewed about the effect of listener familiarity with impaired speech on intelligibility measurements. Factors associated with the listening task, such as listening environment (e.g., Pennington & Miller, 2007) and number of presentations of stimuli (e.g., Coté-Reschny, 2007), may also affect intelligibility scores. Consequently, it has been recommended that intelligibility scores be interpreted relative to listener and listening task variables associated with their measurement (Barreto & Ortiz, 2008; Kent, Weismer, Kent & Rosenbek, 1989). How these variables are manipulated and controlled may also influence the reliability and validity of the intelligibility scores obtained. The current study addressed the effect of listener familiarity with speaker, test stimuli, and judging task on *Ver. 5* intelligibility scores.

Familiarization has been defined as "the process by which a listener's percepts are facilitated by prior exposure to a given signal" (Spitzer, Liss, Caviness & Adler, 2000, p. 285). Listener familiarity or experience with a particular group of speakers ranges on a continuum from experienced listeners, who are familiar because of daily exposure to the speakers, to inexperienced listeners, who may have heard speakers from a particular group "in the past but not on a daily basis or not for a long time" (Monsen, 1983; p. 290). The use of listeners who have frequent exposure to a speaker, such as spouses or parents, or to a group of speakers, such as speech-language pathologists involved in the care of persons with a speech disorder, has been found to yield higher intelligibility scores than using listeners with little familiarity or experience. Dagenais, Watts, Turnage and Kennedy (1999) compared intelligibility scores of speakers with and without dysarthria obtained from three listener groups with normal hearing: young adult listeners (19 - 30 years), older adult listeners (61 - 71 years) and speech-language pathologists with experience working with speakers with dysarthria. The percentage of words identified correctly served as the speaker's intelligibility score and was obtained using open-set word identification of sentences from the *Assessment of Intelligibility of Dysarthric Speech* (*AIDS*) (Yorkston & Beukelman, 1981). While intelligibility scores for the speakers without dysarthria were similar for the three groups of listeners, intelligibility scores for the speakers with dysarthria were significantly higher (i.e., > 5%) for the speech-language pathologists, suggesting that familiarity with the speech characteristics of a group of disordered speakers influences intelligibility scores. Comparisons of listeners

with and without exposure to non-native speakers and of listeners with and without experience listening to hearing impaired speakers has also found that intelligibility scores are higher using listeners with more experience (Kennedy & Trofimovich, 2008: Monsen, 1983). However, Finizia, Lindstrom & Dotevall (1998) found no difference in intelligibility scores for speakers with tracheoesophageal or irradiated laryngeal speech for listeners unfamiliar with this group and speech-language pathologists (i.e., experienced listeners). While this result suggests that experience or familiarity was not a factor, Finizia et al. (1998) did not describe the familiarity of the speech-language pathologists with the group of speakers, therefore, their result may have been a consequence of comparing two groups of inexperienced listeners. These results have implications for how researchers and clinicians describe listeners when reporting the results of intelligibility assessment and for how they interpret intelligibility scores.

Repeated exposure to the same speaker may also increase intelligibility scores. Hustad and Cahill (2003) found that when listeners judged recordings of four different sets of sentences from a single speaker with dysarthria in a single session, intelligibility scores were consistently lower for the sentences heard first than for the sentence set heard second, third and fourth. Scores were also lower for the second presentation than the fourth. No difference in scores was found between the third and fourth presentation. Hustad and Cahill (2003) concluded that familiarization, leading to increased intelligibility scores, can occur as listeners are exposed to more utterances of a speaker within a listening session,

but acknowledged that increased intelligibility scores may also be the result of listeners becoming familiar with the listening task.

Repeated exposure to the same speaker has been found to increase listeners' word identification scores for synthetic speech using a six-choice response task. However, the difference in means for the two listening occasions was less than 3% for both studies (Greenspan, Nusbaum & Pisoni, 1988; Schwab, Nusbaum & Pisoni, 1985). In these studies, listeners heard recordings produced by "a speaker" with fixed severity. Each listener judged a different set of recordings twice, six or ten days apart. These results suggest that familiarity with the listening task may have a small effect on intelligibility scores.

Repeated exposure to the same stimuli may also have an effect on intelligibility scores. Pollack, Rubenstein and Decker (1959) examined the effect of successive judging of word stimuli on intelligibility, determined using an open-set word identification task. Listeners heard either an eight word set presented 15 times in a row or a 144 word set presented three times in a row in a different order each time. For the eight word set, increases in the percentage of words identified correctly were noted after each listen up to three or four successive presentations at positive signal-to-noise ratios. After three or four successive presentations, minimal change in the percentage of words identified correctly was noted. For the 144 word set, a "minimal" increase in the percentage of words identified correctly was found after three presentations at positive signal-to-noise ratios (p. 277, actual increase in percentage points was not described). Ellis and Beltyukova (2008) reported a mean increase of 2.1% for eleven listeners who

orthographically transcribed the same set of 50-word narrative samples from eight children with severe-to-profound hearing loss at sessions one week apart.

For children with cleft palate, it is not known how intelligibility scores obtained using a closed-set response task are influenced by speaker severity, exposure to different sets of words or number of repeated exposures. Research is needed to provide guidelines to test users about how these factors affect intelligibility scores and their interpretation. These findings have implications for how listeners are selected when conducting intelligibility assessments in clinical and research settings with a restricted set of listeners.

The purpose of this study was to develop guidelines for listener participation by examining the effects of listener familiarity with speaker, test stimuli, and listening task on scores obtained using the *Speech Intelligibility Probe for Children with Cleft Palate Version 5* (*SIP-CCLP Ver. 5*) closed-set response task for recordings from children with cleft palate with differing severity of speech disorder. A small but consistent effect (i.e., < 3%) of repeated exposure to the same child was expected, based on previous literature. Recommended guidelines for recruiting listeners for intelligibility assessment are provided based on the results.

Method

Participants

Twenty-seven listeners with no experience judging the speech of children with cleft palate or speech intelligibility were recruited from students at the University of Alberta. Each listener had Canadian English as their first language,

normal hearing as determined by a hearing screening performed according to Alberta College of Speech-Language Pathologists and Audiologists (2008) guidelines and reported that they listen to young children speaking on average once a month or less. Written consent was obtained at the beginning of the listening session. Each listener received an honorarium for their time and participation.

Judging

Selection of child subjects. Recordings of the *SIP-CCLP Ver. 5* words for form 1 and 2 were obtained from 20 children with cleft palate as part of a related study (see chapters 2, 3 and 4) evaluating the reliability and validity of *SIP-CCLP Ver. 5*. Three children were selected randomly from this pool based on their severity classification. Severity classifications were assigned to the 20 children on the basis of the percentage of consonants correct scores (PCC) obtained from phonetic transcription of a 100-word conversational speech sample as described by Shriberg and Kwiatkowski (1982) for 60 children with a developmental phonological disorder ranging in age from three to nine years of age (i.e., mild = 85-100%, mild-moderate = 65-85%, moderate-severe = 50-65%, and severe= <50%). The procedure for phonetic transcription is described in Chapter 3.

One child's *SIP-CCLP* recordings were selected randomly for use in the judging task from the group of children assigned a classification of mild-moderate (n = 2) and two children's *SIP-CCLP* recordings were selected randomly from the group of children assigned a classification of severe (n =12). Two children classified as severe were selected because the majority of recordings available

were from children classified as severe. Furthermore, *SIP-CCLP* will likely be used more frequently to assess children with moderate-to-severe speech disorders than children with mild-to-moderate speech disorders. The child selected with a classification of mild-moderate ((RL01; PCC = 80.33%) was female, 3 years, 3 months of age and had a repaired submucous cleft palate. The first child selected with a classification of severe ((RL02; PCC = 45.11%) was female, 6 years, 2 months of age and had a repaired submucous cleft palate. The second child selected with a classification of severe (RL03; PCC = 39.04%) was male, 3 years, 6 months of age, and had a repaired unilateral cleft lip and palate. Two of the three children had age-appropriate receptive language based on results of the *Fluharty Preschool Speech and Language Screening Test (Fluharty -2)* (Fluharty, 2001). The third child (RL03) scored below the 16th percentile on this screening test. Descriptive information about the three child subjects is provided in Table 5-1.

SIP-CCLP Ver. 5 judging task. All listening sessions took place in a sound booth. During the listening task, the computer hard drive was set up outside the sound booth to improve the signal-to-noise ratio. Speech samples were presented through a Technics Stereo Integrated Amplifier (model SU-V460) connected to ElectroVoice S-40 compact monitor speakers located in the sound booth. Amplification of the speech sample was standardized prior to presentation, with playback volume ranging between 50 – 65 dBA, based on listener comfort level.

Each listener judge was assigned randomly to one child's recordings and to one of two possible form orders: 1) order 1: form 1 – form 2 – form 1 – form 2 ; and 2) order 2: form 2 – form 1 – form 2 – form 1. Nine listeners judged each child's recordings. Order 1 was assigned to five of the nine listeners who judged RL01, four of the nine listeners who judged RL02 and five of the nine listeners who judged RL03. Each listener completed the *SIP-CCLP Ver. 5* closed-set judging task for one form at each session. The *SIP-CCLP Ver. 5* closed-set response task was administered to listeners as described in Chapter 3. The task took between 7 and 15 minutes to complete in a session. Sessions were scheduled one week apart.

Dependent variables. The researcher checked each listener judge's response file for each entry in the "blank" response option that allowed listeners to enter a response that did not correspond to one of the choices provided to see if it contained the sound that occurred in the target word in the contrastive position. If the typed-in response contained the sound that occurred in the target word, it was rescored as correct. The number of phonetic contrast items in which the listener identified the target word was determined, divided by the number of items judged (63) and converted to a percentage to yield an intelligibility score. The phonetic accuracy score was calculated by assigning two points to each item identified correctly with a "clear" rating and one point to each item identified correctly with a "distorted" rating. All other responses received zero points. Number of points was summed for each listener, divided by the total possible

points (number of items judged multiplied by two = 126), and converted to a percentage.

Analysis. Interjudge reliability was estimated via intraclass correlation coefficient (ICC; Type 1, 9; Shrout & Fleiss, 1979) and standard error of measurement (SEM) for each exposure for *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores, where SEM = $\sqrt{MS_{error}}$ of the ICC Analysis of Variance model.

A multivariate analysis of variance (MANOVA) with repeated measures on number of exposures (4 levels) and a between-subjects presentation order factor (2 levels) was conducted to assess the effects of form and presentation order on *Ver. 5* scores. A MANOVA with repeated measures on number of exposures was conducted for each child to assess the effect of repeated exposure on *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores.

Results

Interjudge reliability. Intraclass correlation coefficients with 95% confidence intervals and standard errors of measurement are reported in Table 5-2 for each exposure. ICCs (1, 9) ranged from .993 to .995 for intelligibility scores and .988 to .991 for phonetic accuracy scores. SEMs ranged from 3.98 to 5.52 for intelligibility scores and from 5.32 to 6.65 for phonetic accuracy scores.

Presentation order. For the 14 listeners who judged form 1 first, mean intelligibility scores across the three children for each exposure (exp) were: 65.19% (SD = 21.03) for exp 1, 68.93% (SD = 17.73) for exp 2, 69.16% (SD = 18.48) for exp 3 and 73.01% (SD = 17.22) for exp 4. For the 13 listeners who

judged form 2 first, mean intelligibility scores across the three children for each exposure were as follows: 69.96% (SD = 25.57) for exp 1, 69.96% (SD = 26.03) for exp 2, 72.53% (SD = 25.23) for exp 3 and 74.72% (SD = 26.17) for exp 4. Intelligibility scores over time are shown in Figure 5-1 for the two presentation orders. There was a significant main effect of exposure ($F_{(3, 75)} = 18.17$, $p = .000$). The main effect of presentation order was not statistically significant ($F_{(1, 25)} = .153$, $p = .699$). The interaction between exposure and presentation order was also not statistically significant ($F_{(3, 75)} = 1.83$, $p = .149$).

For the 14 listeners who judged form 1 first, mean phonetic accuracy scores across the three children for each exposure were: 57.43% (SD = 20.54) for exp 1, 60.54% (SD = 17.16) for exp 2, 61.05% (SD = 19.08) for exp 3 and 64.57% (SD = 16.39) for exp 4. For the 13 listeners who judged form 2 first, mean phonetic accuracy scores across the three children for each exposure were as follows: 57.51% (SD = 19.84) for exp 1, 58.55% (SD = 22.74) for exp 2, 61.72% (SD = 21.08) for exp 3 and 63.98% (SD = 23.48) for exp 4. Phonetic accuracy scores over time are shown in Figure 5-2 for the two presentation orders. A significant main effect of exposure was found ($F_{(3, 75)} = 18.17$, $p = .000$). The main effect of presentation order was not statistically significant ($F_{(1, 25)} = .005$, $p = .944$). The interaction between exposure and presentation order was also not statistically significant ($F_{(3, 75)} = .716$, $p = .545$)

Mean difference in intelligibility scores and standard deviation of the differences when the same form was judged twice with two weeks between judgments (i.e., exp 1 and exp 3; exp 2 and exp 4), different forms were judged

with one week between judgments (i.e., exp 1 and exp 2; exp 2 and exp 3; exp 3 and exp 4), and different forms were judged with three weeks between judgments (i.e., exp 1 and exp 4) for each order and child are shown in Table 5-3. For all three children (collapsed across order), the greatest mean difference was between exp 1 and exp 4. Between exp 1 and exp 4, the mean difference was -4.23 (SD = 3.07) for RL01, -4.94 (SD = 5.41) for RL02, and -9.88 (SD = 7.56) for RL03. Regardless of order, the mean difference between the same forms judged at exp 1 and exp 3 was lowest for RL01 (order 1: -1.27 (SD = 3.79); order 2: 1.19 (SD = 1.99)) and highest for RL03 (order 1: -7.30 (SD = 4.91); order 2: -4.37 (SD = 4.91)). The first time form 2 was judged before form 1 (i.e., order 1: exp 2 and exp 3; order 2: exp 1 and exp 2) the difference before scores tended to be smaller than other comparisons. For five of the six comparisons, the mean difference in scores was less than one, with the only exception being the comparison for RL02 Order 1 (Mean difference = -1.98, SD = 1.52).

Mean difference in phonetic accuracy scores and standard deviation of the differences when the same form was judged twice with two weeks between judgments (i.e., exp 1 and exp 3; exp 2 and exp 4), different forms were judged with one week between judgments (i.e., exp 1 and exp 2; exp 2 and exp 3; exp 3 and exp 4), and different forms were judged with three weeks between judgments (i.e., exp 1 and exp 4) for each order and child are shown in Table 5-4. For all three children (collapsed across order), the greatest mean difference was between exp 1 and exp 4. Between exp 1 and exp 4, the mean difference was -7.14 (SD = 5.84) for RL01, -6.47 (SD = 5.53) for RL02, and -6.82 (SD = 5.56) for RL03.

Regardless of order, the mean difference between the same forms judged at exp 1 and exp 3 was lowest for RL01 (order 1: -1.43 (SD = 5.06); order 2: -2.58 (SD = 1.99)).

Mean intelligibility and phonetic accuracy scores for each child at each exposure are displayed graphically in Figures 5-3 and 5-4, respectively. For the child with a speech severity classification of mild-moderate (RL01), a significant effect of exposure was found for both intelligibility ($F_{(3, 24)} = 6.84$, $p = .002$) and phonetic accuracy scores ($F_{(3, 24)} = 11.61$, $p = .000$). The mean intelligibility score was 84.48% (SD = 4.47) for exp 1, 85.18% (SD = 3.64) for exp 2, 84.66% (SD = 4.56) for exp 3 and 88.71% (SD = 3.12) for exp 4. Post-hoc testing revealed that mean intelligibility scores were significantly different between exp 1 and exp 4 ($p < .05/6$) and exp 2 and exp 4 ($p < .05/6$). The mean phonetic accuracy score was 73.90% (SD = 6.19) for exp 1, 75.48% (SD = 5.82) for exp 2, 75.84% (SD = 5.82) for exp 3 and 79.89% (SD = 6.02) for exp 4. Post-hoc testing revealed that mean phonetic accuracy scores were significantly different between exp 1 and exp 4 ($p < .05/6$) and exp 2 and exp 4 ($p < .05/6$).

For RL02 (speech severity classification of severe), a significant effect of exposure was found for both intelligibility ($F_{(3, 24)} = 5.63$, $p = .005$) and phonetic accuracy scores ($F_{(3, 24)} = 6.128$, $p = .003$). The mean intelligibility score was 76.37% (SD = 5.41) for exp 1, 77.43% (SD = 3.44) for exp 2, 80.07% (SD = 3.73) for exp 3 and 81.31% (SD = 4.11) for exp 4. Post-hoc testing revealed that mean intelligibility scores were significantly different between exp 2 and exp 3 ($p < .05/6$). The mean phonetic accuracy score was 61.99% (SD = 7.53) for exp 1,

65.08% (SD = 4.48) for exp 2, 67.99% (SD = 5.07) for exp 3 and 69.49% (SD = 7.68) for exp 4. Post-hoc testing did not reveal any significant differences between the exposures.

For the second child with a speech severity classification of severe (RL03), a significant effect of exposure was found for both intelligibility ($F_{(3, 24)}$ = 9.418, p = .000) and phonetic accuracy scores ($F_{(3, 24)}$ = 4.80, p = .009). The mean intelligibility score was 41.62% (SD = 4.03) for exp 1, 45.68% (SD = 5.77) for exp 2, 47.62% (SD = 6.05) for exp 3 and 51.50% (SD = 7.06) for exp 4. Post-hoc testing revealed that mean intelligibility scores were significantly different between exp 1 and exp 4 ($p < .05/6$) and exp 3 and exp 4 ($p < .05/6$). The mean phonetic accuracy score was 36.51% (SD = 5.05) for exp 1, 38.18% (SD = 6.09) for exp 2, 40.30% (SD = 5.24) for exp 3 and 43.47% (SD = 5.03) for exp 4. Post-hoc testing revealed that mean phonetic accuracy scores were significantly different between exp 2 and exp 4 ($p < .05/6$).

Discussion

This study examined the effect of listener familiarity with speaker, test stimuli, and judging task on *SIP-CCLP Ver. 5* scores for three children with cleft palate. One child with a speech severity classification of mild-moderate and two children with a speech severity classification of severe were selected for judging to represent the range of speech severity for children with cleft palate. The mean increase from the first to the fourth exposure was 6.4% for intelligibility scores and 6.9% for phonetic accuracy scores.

Examination of the results for each child revealed a consistent effect of familiarization on *Ver. 5* scores, such that scores became higher as the number of exposures increased. For all children, the mean intelligibility and phonetic accuracy scores at exposure one were always lower than the mean intelligibility and phonetic accuracy scores at exposure four. This difference was statistically significant for two of the three children (RL01, RL03). Hustad and Cahill (2003) also found that when listeners judged recordings of four different sets of sentences from a single speaker with dysarthria in a single session, intelligibility scores were consistently lower for the set heard first than for the set heard fourth. This difference was statistically significant for all five speakers in Hustad and Cahill (2003).

When listeners judged the same form two weeks after the initial exposure, the mean difference was -3.85 (SD = 4.42) and -4.31 (SD = 4.35) for intelligibility and phonetic accuracy scores, respectively. This difference is higher than the mean difference reported in chapter 3 for listeners who judged the same form one week after the initial exposure (i.e., intelligibility scores: -2.81 (SD = 4.15); phonetic accuracy scores: -1.91 (SD = 4.67)). Listeners in the current study judged a different form from the same child before they judged the same form a second time. This additional experience appears to result in higher *Ver. 5* scores. Further research examining the effects of repeated exposure with a two-week interval between judging either the same form from the same child or a different form from the same child is needed.

Intelligibility and phonetic accuracy scores were higher at exposure two than exposure one (mean increase 3.3% and 3.9%, respectively). This is higher than the mean increase of 1.1% reported by Greenspan, Nusbaum and Pisoni (1988) and mean decrease of 2.8% reported by Schwab, Nusbaum and Pisoni (1985). In these studies, listeners judged synthetic speech using a six-choice response task six or ten days after the initial exposure. A greater difference between scores obtained at the two times might be expected in this study as there is more variability in children's word productions (e.g., loudness) than in synthetically produced "speech."

Closer examination of the results at exposure one and two revealed different patterns depending on which form listeners judged first. When listeners judged form 1 the first week and form 2 the second week (order one), intelligibility scores were on average 3.7% higher for form 2. However, when listeners judged form 2 the first week and form 1 the second week (order two), the mean difference in intelligibility scores was 0%. These results suggest that form 1 and 2 may not be equivalent. An item analysis has been conducted to identify which items are functioning differently in the two forms. As described in Appendix D, the two forms are not equivalent as form 2 has more easy items (i.e., high difficulty index) than form 1.

The effect of repeated exposure on intelligibility scores was different for all three children in this study. For the child classified as mild-moderate severity; the largest differences in mean intelligibility score were between the first and fourth exposure, the second and fourth exposure and the third and fourth

exposure, suggesting a gradual cumulative effect of repeated exposure. Hustad and Cahill (2003) reported the opposite pattern for their two speakers with mild dysarthria, such that intelligibility scores obtained in the first trial were significantly different from scores obtained in the other three trials. In Hustad and Cahill (2003), listener judgments were collected in a single session, while in this study listener judgments were collected over a four-week period. Repeated exposure may have a more immediate effect (i.e., significant difference between exposure one and other three exposures) when there is less time between judgments.

For the two children with a speech severity classification of severe, the largest differences in mean intelligibility score were between the first and fourth exposure, the second and fourth exposure and the first and third exposure. Examination of Figures 5-3 and 5-4 revealed a stair-step pattern of increasing scores for these two children that is similar to results reported by Hustad and Cahill (2003) for three adult speakers with severe dysarthria. For intelligibility scores, the mean difference between the first and second exposure was 4% for RL03 but less than 1.5% for RL02 and RL01. This result suggests that repeated exposure had a greater effect on listeners judging RL03 than on listeners judging RL02, such that listeners' ability to identify RL03's productions of target words improved with each exposure. For phonetic accuracy scores, a smaller mean difference was found for RL03 (i.e., 1.7%) than for RL02 (i.e., 3.1%). This result suggests that although listeners were identifying more words correctly after repeated exposure to the RL03's voice, more words were rated as distorted. For

RL02, it appears that listeners were identifying a similar number of words but rating more of them as clear after the second exposure to the child's voice. These results suggest that repeated exposure may not have the same effect on *Ver. 5* scores for children with similar speech severity, as defined by the percentage of consonants correct. The results also demonstrate that PCC scores and *SIP-CCLP* scores rank children in the same order but differ in their sensitivity, with the latter showing a greater difference between scores than the former.

In summary, these results suggest the following effects on *SIP-CCLP Ver. 5* scores if listeners are recruited to judge the same child twice within a two-week period: 1) If the same form is judged, scores are biased to be higher on the second exposure, 2) If form 1 is judged first, scores are expected to be higher on form 2 and 3) If form 2 is judged first, scores are expected to be similar for the two forms/exposures. In general, the difference between scores is expected to be greater for children with cleft palate with a speech severity classification of severe than for children with cleft palate with a speech severity classification of mild-moderate.

Intelligibility and phonetic accuracy scores were collected for each of these children as part of an evaluation of the reliability and validity of *SIP-CCLP Ver. 5.* Mean scores and the 95% confidence interval for each child's intelligibility and phonetic accuracy scores are reported in Table 5-5. Comparison of listeners' scores at each exposure to the 95% confidence interval for each form and child revealed that at exposure two all but two listeners' scores for RL02 were within this interval for intelligibility scores. Similarly, all but three listeners'

scores for RL02 and one listener's score for RL03 were within this interval for phonetic accuracy scores. As a result, it is recommended that, while it is better to use listeners who have not been exposed previously to the child's speech for each administration of the judging task, users of *SIP-CCLP Ver. 5* can recruit the same listener to judge the same child (on a different form) as long as there is at least a week between administrations.

In this study, listeners judged the same child four times within a four-week period. A small but significant effect of repeated exposure to a child's speech was noted. Severity of the child's speech disorder appears to affect the number of exposures at which change in scores is noted. Further research into the effects of increased time between exposures on *SIP-CCLP Ver. 5* scores is needed to develop additional recommendations about listener recruitment. Investigation into processes listeners use when responding to items in the S*IP-CCLP Ver. 5* closed-set response task would provide additional insight into differences among listeners, as well as the construct validity of *Ver. 5*.

Table 5-1

Descriptive Characteristics of the Children whose Recordings were Judged

	Child Subjects		
	RL01	RL02	RL03
Age (months)	39	74	42
Gender	Female	Female	Male
Cleft Type	SMCP	SMCP	UCLP
Receptive Language	$\geq 16^{th}$ %ile	$\geq 16^{th}$ %ile	$\leq 16^{th}$ %ile
Speech Severity Classification	mild-moderate (PCC = 80.33)	severe (PCC = 45.11)	severe (PCC = 39.04)
Hypernasality Rating	40	180	100
Voice Severity Rating	80	45	25
Additional Information	n/a	Adopted from China	n/a

Note. SMCP = submucous cleft palate; UCLP = unilateral cleft lip and palate. In Chapter 3, child RL01 is identified as CP02, child RL02 is identified as CP19, and child RL03 is identified as CP04.

Table 5-2

Inter-rater Reliability for SIP-CCLP Ver. 5 Scores at Each Exposure

		ICC	95% Confidence Interval for ICC	SEM
Exposure 1	Intelligibility	.995	.980 – 1.0	3.98
	Phonetic Accuracy	.988	.947 – 1.0	5.95
Exposure 2	Intelligibility	.995	.979 – 1.0	4.31
	Phonetic Accuracy	.991	.961 – 1.0	5.54
Exposure 3	Intelligibility	.994	.972 – 1.0	4.45
	Phonetic Accuracy	.991	.960 – 1.0	5.32
Exposure 4	Intelligibility	.993	.969 – 1.0	5.52
	Phonetic Accuracy	.987	.945 – 1.0	6.65

Note. ICC = intraclass correlation coefficient; SEM = standard error of measurement.

Table 5-3

Mean and Standard Deviation of the Differences in Intelligibility Scores for Each Child and Order

	Exp1 – Exp3	Exp2 – Exp4	Exp1 - Exp2	Exp2 – Exp3	Exp3 – Exp4	Exp1 – Exp4
Order 1	F1 – F1	F2 – F2	F1 – F2	F2 – F1	F1 – F2	F1 – F2
RL01	-1.27 (3.79)	-4.13 (2.40)	-1.59 (1.12)	0.32 (3.06)	-4.44 (5.07)	-5.71 (2.13)*
RL02	-3.18 (5.94)	-2.78 (3.00)	-1.19 (6.90)	-1.98 (1.52)	-0.80 (2.04)	-3.97 (6.54)
RL03	-7.30 (4.91)*	-5.08 (6.78)	-7.94 (5.87)	0.63 (5.08)	-5.71 (2.90)*	-13.02 (7.60)
01, 02 & 03	-3.97 (5.49)	-4.08 (3.72)*	-3.74 (4.89)	-0.23 (3.17)	-3.85 (3.87)*	-7.82 (6.42)*
Order 2	F2 – F2	F1 – F1	F2 – F1	F1 – F2	F2 – F1	F2 – F1
RL01	1.19 (1.99)	-2.78 (2.00)	0.40 (5.24)	0.80 (3.99)	-3.57 (2.38)	-2.38 (3.30)
RL02	-4.13 (2.13)	-4.76 (4.05)	-0.95 (1.42)	-3.17 (2.51)	-1.59 (5.94)	-5.71 (4.97)
RL03	-4.37 (4.91)	-6.75 (6.78)	0.79 (5.87)	-5.16 (5.08)	-1.59 (2.90)	-5.95 (7.60)
01, 02 & 03	-2.56 (3.83)	-4.76 (4.54)*	0 (3.94)	-2.56 (4.21)	-2.20 (3.91)	-4.76 (5.25)*
ALL	-3.29 (4.76)*	-4.41 (4.07)*	-1.94 (4.83)	-1.35 (3.88)	-3.06 (3.96)*	-6.35 (6.00)*

Note. Exp = exposure; F1 = form 1; F2 = form 2. Standard deviation of the differences is in brackets beside the mean. * = significant at p = .05/6.

Table 5-4

Mean and Standard Deviation of the Differences in Phonetic Accuracy Scores for Each Child and Order

	Exp1 – Exp3	Exp2 – Exp4	Exp1 - Exp2	Exp2 – Exp3	Exp3 – Exp4	Exp1 – Exp4
Order 1	F1 – F1	F2 – F2	F1 – F2	F2 – F1	F1 – F2	F1 – F2
RL01	-1.43 (5.06)	-4.76 (2.51)	-0.47 (2.54)	-0.96 (3.81)	-3.81 (5.65)	-5.24 (1.44)*
RL02	-5.75 (9.37)	-2.58 (3.91)	-2.18 (11.19)	-3.57 (3.70)	0.99 (1.00)	-4.76 (8.86)
RL03	-4.13 (2.62)*	-4.44 (4.81)	-6.51 (5.04)	2.38 (4.01)	-6.83 (2.28)*	-10.95 (4.87)*
01, 02 & 03	-3.63 (6.61)	-4.03 (3.58)*	-3.12 (6.36)	-0.51 (4.80)	-3.51 (4.71)	-7.14 (5.84)*
Order 2	F2 – F2	F1 – F1	F2 – F1	F1 – F2	F2 – F1	F2 – F1
RL01	-2.58 (1.99)	-3.97 (1.72)	-2.98 (2.37)	0.39 (3.58)	-4.36 (2.29)	-6.94 (1.98)*
RL02	-6.19 (1.97)*	-5.88 (4.29)	-3.81 (1.89)	-2.38 (2.92)	-3.49 (6.56)	-9.68 (5.97)
RL03	-3.37 (2.62)	-6.35 (4.81)	4.37 (5.04)	-7.74 (4.01)	1.39 (2.28)	-1.98 (4.87)
01, 02 & 03	-4.21 (2.74)*	-5.43 (3.85)*	-1.04 (4.63)	-3.18 (4.54)	-2.26 (4.70)	-6.47 (5.53)*
ALL	-3.91 (5.00)*	-4.70 (3.64)*	-2.12 (5.66)	-1.79 (4.83)	-2.91 (4.73)*	-6.82 (5.56)*

Notes. Exp = exposure; F1 = form 1; F2 = form 2. Standard deviation of the differences is in brackets beside the mean. * = significant at p = .05/6.

Table 5-5

SIP-CCLP Ver. 5 Results for Child Subjects from Chapter 3

		Child Subjects		
		RL01	RL02	RL03
Form 1				
Intelligibility	Mean	84.66	69.84	42.86
	95% CI	[76.39, 92.93]	[61.57, 78.11]	[34.59, 51.13]
Phonetic Accuracy	Mean	76.46	55.29	35.19
	95% CI	[67.52, 85.40]	[46.35, 64.23]	[26.25, 44.13]
Form 2				
Intelligibility	Mean	85.19	69.84	44.44
	95% CI	[76.92, 93.46]	[61.57, 78.11]	[36.17, 52.71]
Phonetic Accuracy	Mean	76.72	58.99	40.74
	95% CI	[67.78, 85.66]	[50.05, 67.93]	[31.80, 49.68]

Note. CI = confidence interval.

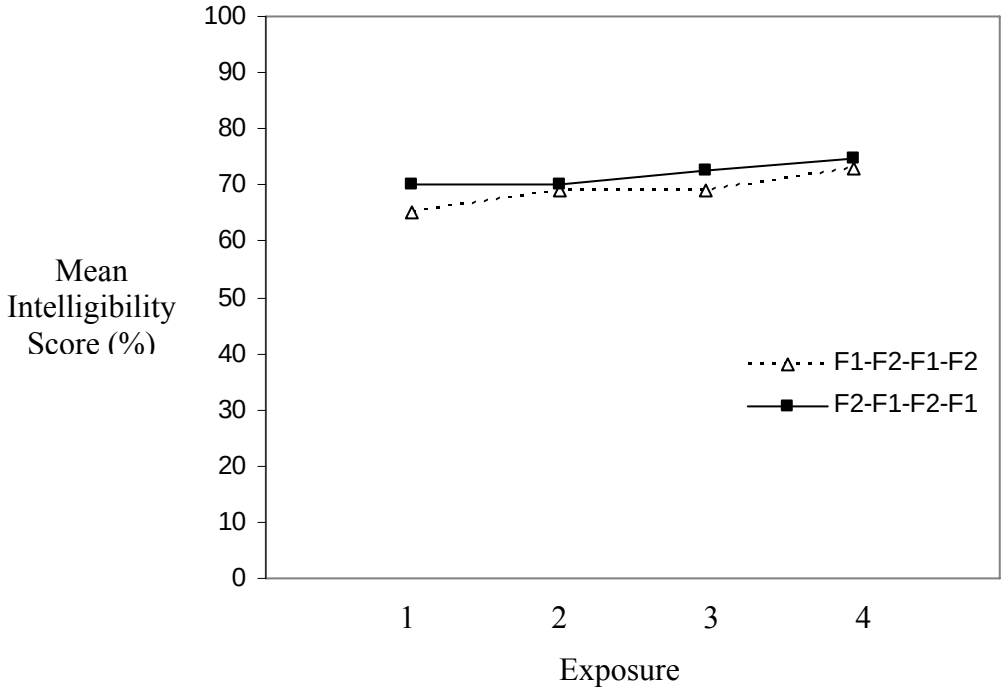

Figure 5-1. Mean intelligibility scores over time for each order (collapsed across children)

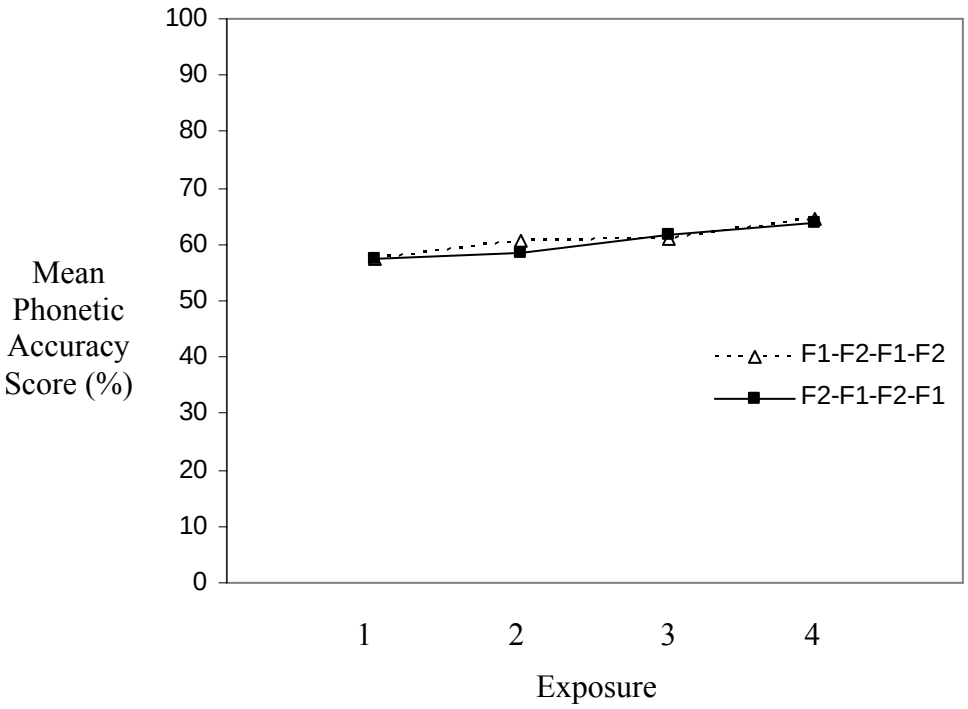

Figure 5-2. Mean phonetic accuracy scores over time for each order (collapsed across children)

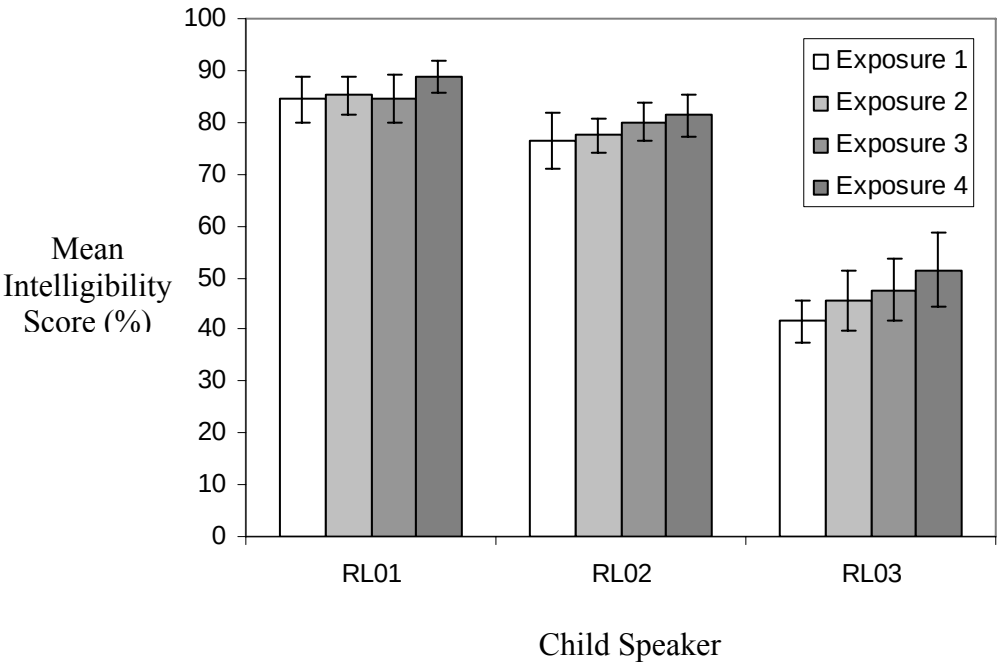

Figure 5-3. Mean intelligibility scores (±1 SD) by exposure and child speaker

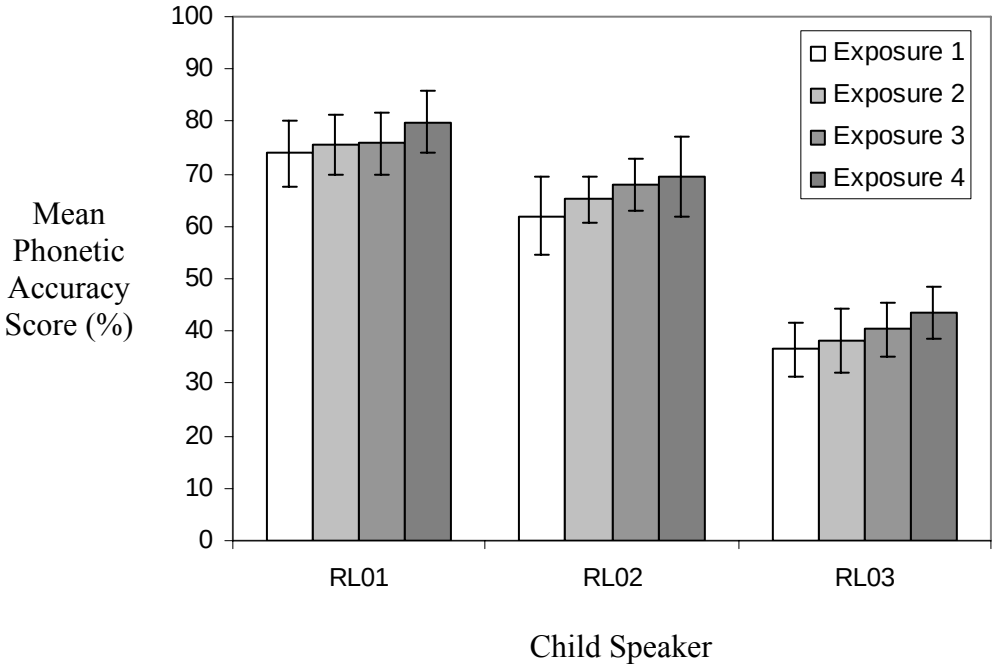

Figure 5-4. Mean phonetic accuracy scores (±1 SD) by exposure and child speaker

References

Alberta College of Speech-Language Pathologists and Audiologists. (2008). *Hearing screening guidelines*. Retrieved from http://www.acslpa.ab.ca/public/data/documents/ACFC3D5.pdf

Barreto, S. S., & Ortiz, K. Z. (2008). Intelligibility measurements in speech disorders: A critical review of the literature. *Pro-Fono Revista de Atualização Cientifica, 20*(3), 201-206.

Coté-Reschny, K. J. (2007). *Effects of talker severity and repeated presentations on listener judgments of the speech intelligibility of young children with dysarthria.* (Unpublished master's thesis). University of Alberta, Edmonton, AB.

Dagenais, P. A., Watts, C. R., Turnage, L. M., & Kennedy, S. (1999). Intelligibility and acceptability of moderately dysarthric speech by three types of listeners. *Journal of Medical Speech-Language Pathology, 7(2)*, 91-96.

Ellis, L. W., & Belyukova, S. A. (2008). Effects of training on naïve listeners' judgments of the speech intelligibility of children with severe-to-profound hearing loss. *Journal of Speech, Language and Hearing Research, 51*, 1114-1123.

Finizia, C., Lindstrom, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: Laryngectomy versus radiotherapy. *Laryngoscope, 108(1)*, 138-143.

Fluharty, N. (2001). *Fluharty Preschool Speech and Language Screening Test (Second edition)*. Austin, TX, Pro-Ed.

Greenspan, S. L., Nusbaum, H.C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3)*, 421-433.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*, 198-208.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64(3)*, 459-489.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Monsen, R. B. (1983). The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders, 48*, 286-296.

Pennington, L., & Miller, N. (2007). Influence of listening conditions and listener characteristics on intelligibility of dysarthric speech. *Clinical Linguistics & Phonetics, 21(5)*, 393-403.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of the Acoustical Society of America, 31(3)*, 273-279.

Shriberg, L. D., & Kwiatkowski, J. (1982). Continuous speech sampling for phonologic analyses of speech-delayed children. *Journal of Speech and Hearing Disorders, 50*, 323-334.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2)*, 420-428.

Spitzer, S. M., Liss, J. M., Caviness, J. N., & Adler, C. (2000). An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech. *Journal of Medical Speech-Language Pathology, 8(4)*, 285-293.

Schwab, E.C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27(4)*, 395-408.

Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language & Communication Disorders, 43(6)*, 633-648.

Yorkston, K. M., & Beukelman, D. R. (1981). *Assessment of Intelligibility of Dysarthric Speech*. Tigard, OR; C.C. Publications, Inc.

Chapter 6

General Discussion and Conclusions

Overview

Articulation and resonance disorders can affect how well children with cleft palate are understood by peers, caregivers, families and other people in their community. Measures of speech intelligibility reflect the integrated impact of a talker's resonance, articulation, voice and prosody on speech ability. Intelligibility, defined as the degree to which an individual's spoken message is recovered by a listener (Kent, Weismer, Kent & Rosenbek, 1989), is used by researchers and clinicians to describe speech disorder severity. Protocols developed for speech assessment of children with cleft palate recommend the inclusion of a reliable and valid global measure of communicative activity, such as intelligibility (e.g., Lohmander, Willadsen, Persson, Henningsson, Bowden & Hutters, 2009; Henningsson, Kuehn, Sell, Sweeney, Trost-Cardamone, & Whitehill, 2008). The *Speech Intelligibility Probe for Children with Cleft Palate* (*SIP-CCLP*) is a computer-mediated word imitation measure of intelligibility that targets the speech error patterns of English-speaking children with cleft palate (Hodge & Gotzke, 2007). The purpose of the current work was two-fold. First, *SIP-CCLP* was revised substantially to improve its' sensitivity, utility and efficiency for clinical and research applications. With each change to a measure, the test developer must reevaluate reliability and validity with samples of subjects from the population for whom the test is intended. The second purpose was to evaluate the reliability and validity of the revised *SIP-CCLP*, *Version 5*, with a

sample of 20 English-speaking children with cleft palate, ranging in age from 37 to 84 months. The construction (item selection and item scaling) and results of the evaluation of *SIP-CCLP Ver. 5,* as a discriminative index of intelligibility, and recommendations for research and clinical use of this measure are discussed in the following sections.

Construction

Item selection. According to Kirshner and Guyatt (1985), three criteria should be used when selecting items for a discriminative measure: a focus on features that are influenced by the condition, applicability to all examinees and stability over time. To ensure that the *SIP-CCLP Ver. 5* items focused on the speech error patterns of children with cleft palate, results from Gotzke (2005), guidelines for sampling cleft palate speech (e.g., European Collaboration in Craniofacial Anomalies (EUROCRAN), 2009), and literature authored by recognized experts in the field were used to develop its content. To assess content relevance and representativeness of these error patterns, eight international experts in the speech of children with cleft palate were invited to participate in the evaluation of the content-related validity of *SIP-CCLP Ver. 5*. For the content review, each expert was asked to identify which of those error patterns tested in *SIP-CCLP* occur rarely (i.e., <10%) in the speech of young children with cleft palate who also have a speech disorder and to provide comments about their ratings. Then, each expert was asked to list any error patterns that should be added to provide adequate representation of the speech error patterns of children with cleft palate. Experts agreed that 11 error patterns targeted in *Ver. 5* occur in

more than 10% of these children and 26 error patterns occur in fewer than 10% of these children. A limitation of this method of rating is that it required experts to rely on their recall to judge if each error pattern occurs in more or less than 10% of the young children with cleft palate who also have a speech disorder. Errors identified by listeners for the 15 children with cleft palate who participated in the evaluation of *SIP-CCLP Ver. 3* (Gotzke, 2005) were compared to the list of 26 error patterns rated as occurring in fewer than 10% of children with cleft palate and a speech disorder by the expert raters. Listeners identified errors for all but four of these patterns for the children in the previous study. These four error patterns were removed from the list of error patterns tested in *SIP-CCLP Ver. 5* and the other 22 were retained.

The error patterns identified by listeners for the 20 children with cleft palate in the current study were examined as part of the evaluation of construct-related validity of *Ver. 5*. The majority of errors identified with both forms were in the manner preference error category (F1: 38.2%; F2: 33.4%), confirming the results obtained by Gotzke (2005) and Whitehill and Chau (2004). All but two error patterns (i.e., glottal stops for affricates, voiced for voiceless affricates) were identified for at least one child (5% of subjects). This result for glottal stops for affricates is surprising as experts identified this pattern as one of the eleven error patterns targeted in *Ver. 5* that occur in more than 10% of children with cleft palate with a speech disorder. The error pattern glottal stops for oral stops, also identified as occurring in more than 10% of children with cleft palate by the expert panel, was identified in fewer than 10% of children in this study (form 2).

The remaining nine error patterns were identified in more than 10% of children in this study. Three error patterns identified by experts as occurring in less than 10% of children with cleft palate occurred for more than five children for both forms: labiodental fricatives for alveolar sibilants (sibilant error), voiceless stops for voiced (voicing error), and voiceless fricatives for voiced (voicing error). One error pattern that was not included in *Ver. 5* (i.e., glottal fricative for stops) was identified by listeners for 15% of the children using the type-in "blank" option. High agreement of expert raters with results of children's performance on the *SIP-CCLP* evaluation for errors that occurred in 10% or more of children (9 agreements; 2 disagreements) and for errors that occurred in less than 10% of children (23 agreements, 3 disagreements) support the selection of the *Ver. 5* error patterns for a discriminative measure of speech intelligibility, as well as the construct-related validity of the error patterns sampled in *Ver. 5*. A multi-center international study is needed to determine if there are regional differences in the occurrence of some of the *SIP-CCLP* error patterns related to variation in surgical timing and procedures and access to services in different countries. The small sample size may also account for the differences between this study and experts' ratings. In addition, different methods of error identification may account for some disagreements. The expert raters likely based their ratings on online phonetic transcription of children's speech where the transcriber has a view of the child's face and the stimulus words are known to the transcriber, as in an articulation test, whereas in *SIP-CCLP Ver. 5*, error patterns are based on phonetic identification using a forced-choice task and the auditory signal of the

child's production. On the basis of the results from the current study, one error pattern will be added (i.e., glottal fricative for stops) and two error patterns will be deleted (i.e., glottal stops for affricates and voiced for voiceless affricates) from the next version of *SIP-CCLP*.

Frequency of usage counts by Stemach and Williams (1988) and Kolson (1960) were used to determine if the *Ver. 5* stimulus words were appropriate for kindergarten and/or first grade children. All but seven of the 114 words were identified as occurring in the vocabulary of English-speaking children in kindergarten or first grade. The *SIP-CCLP Ver. 5* software was revised to include a familiarization activity at the beginning of the task to introduce the unfamiliar vocabulary to child participants. While recordings of the child's production of these words in the familiarization activity were obtained, they were not judged by listeners. Future research is needed to investigate if there are differences in the accuracy of children's productions in the familiarization activity and during test administration. The *Ver. 5* stimulus words were determined to be appropriate for children in kindergarten or first grade. However, half of the children in this study were less than five years of age. While all of the children produced all of the words, it is not known if familiarity of the words to the younger children affected the intelligibility of their imitative productions.

The stability of children's productions of the *SIP-CCLP Ver. 5* stimulus words can also be examined by evaluating whether there are differences in listeners' responses to the 12 stimulus words that are in both forms. The number of "common" stimulus words that were scored as incorrect by a minimum of two

of the three listeners in one form and as correct in the other form ranged from 1 to 8 words (median = 3 words) for the 20 children with cleft palate, suggesting that the accuracy with which children produce a word within a one hour session can vary. Neither age nor severity of speech intelligibility impairment appears to be related to the number of variable productions. Because two different sets of three listeners judged the children's productions in each form, it is possible that differences in listeners' abilities to identify children's productions contributed to the lack of stability. Phonetic transcription of the *SIP-CCLP* words would provide additional insight into whether it was the child's productions or the listeners' responses that were not stable. This would also permit a comparison of error patterns obtained from phonetic transcription and the forced-choice format of *SIP-CCLP Ver. 5*.

Item scaling. Item scaling refers to the number of response options available for each item (Kirshner & Guyatt, 1985). Items with few response options are recommended for discriminative measures to minimize variability among respondents in interpreting the possible responses. In *SIP-CCLP Ver. 5*, the number of response options for each item in the closed-set response task was increased to six (four minimally contrastive words, "other" and "can't identify") to address a number of limitations in the previous versions (e.g., multiple presentations of target stimulus words; see Chapter 2). Increasing the number of options was also expected to reduce listener's use of the "blank" response option to type in a response different from the choices provided, which may, in turn, reduce variability in the typed-in responses among listeners. Two of the three

listeners typed the same response in the "blank" for 43% of the errors identified for the 15 children with cleft palate in Gotzke (2005). In the current study, two of the three listeners typed the same response in the "blank" for only 18% of the errors in form 1 and 20% of the errors in form 2. The mean percentage of responses for which listeners chose the "blank" response option was 10.56% (SD = 9.74) for form 1 and 7.80% (SD = 7.48) for form 2. The percentage of errors for which no consensus was reached among listeners on the error response also decreased from 27.9% in Gotzke (2005) to 23.5% on form 1 and 18.8% on form 2. It appears that increasing the number of response options decreased listener's use of the "blank" response option, which may have also decreased variability among the three listeners' responses.

An advantage of the "blank response option" is that it may be used by listeners to identify error patterns that can not be captured using sounds in the American English phonological system (e.g., double articulation, nasal fricatives). Using the "blank" response option, listeners identified several instances in which a fricative was identified as a consonant cluster in form 1 (n = 7) and form 2 (n = 2), some of which may be substitution errors involving nasal fricatives. Examination of the relationship between the phonetic transcriptions of the children's productions of the *SIP-CCLP* stimulus words and listeners' identification of error patterns is needed to determine if children used non-English sound substitutions and if so, how listeners' responded to the children's use of these sounds.

Evaluation

Reliability. Methodologies from classical test, generalizability and item response theory were used to assess the reliability of *SIP-CCLP Ver. 5.* Chapter 3 examined reproducibility (i.e., the stability of an instrument over forms and occasions; and inter-rater agreement) and internal consistency (i.e., how consistently examinees performed across items) using classical test theory. The Standards for Educational and Psychological Testing, developed by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA & NCME, 1999), were followed in reporting the results of this evaluation. Generalizability theory was used to simultaneously evaluate forms, occasions and listeners as sources of error in *SIP-CCLP* scores (Appendix E), while item response theory was used to evaluate the equivalence of the two forms (Appendix F).

Two alternate forms were developed for *SIP-CCLP Ver. 5*. To be considered parallel forms in the classical test theory model, the means and variances of the scores from the two forms must not differ significantly for the two orders of administration (e.g., form 1 followed by form 2, form 2 followed by form 1) and for the combined data set. For *SIP-CCLP Ver. 5,* the means and variances were significantly lower for form 1 than form 2 for the children administered form 1 first, indicating that the two forms were not parallel. This result was confirmed using the limits of agreement method (Bland & Altman, 1986), as a bias was noted over forms. On average, intelligibility scores were 4.9% and phonetic accuracy scores were 3.7% lower on form 1 than form 2.

Unexpectedly, no main effect of forms was found using generalizability theory, suggesting that the mean score for each child was not different from form to form. However, when the variance components for the interactions of persons with form were summed, between 4.2% and 12.3% of variance in scores was explained, suggesting a possible difference between forms. The results of the analysis of form equivalence using item response theory confirmed that form 2 was easier than form 1. Overall, the results indicate that *SIP-CCLP* form 1 and 2 are not parallel. To address this issue, a classical test theory item analysis (Appendix D) was conducted to identify items to be exchanged between the two forms to improve form equivalence. To balance the number of "easy" items on the two forms, six exchanges were suggested (i.e., tea (F1) – two (F2); sap (F1) – sip (F2); sell (F1) – sail (F2); rail (F1) – row (F2); year (F1) – yell (F2); knee (F1) – no (F2)). To determine if these revisions create two parallel forms, both revised forms must be administered to a group of children in a single session, listener responses obtained and statistical testing for parallel forms completed.

Reproducibility of *SIP-CCLP* scores over time, forms, time and forms, and listeners was examined using classical test theory. Reliability coefficients were greater than 0.9 for all evaluations, indicating that *SIP-CCLP* scores are stable when differentiating between individuals over time, forms, and listeners. According to Kirshner and Guyatt (1985), high test-retest correlations are desirable for discriminative health status measures, as they indicate large and stable inter-examinee variation. For *SIP-CCLP Ver. 5*, intraclass correlation coefficients for test-retest reliability ranged from .93 to .97. Standard error(s) of

measurement (SEM) was also calculated to quantify the precision of scores over time, forms, and time and forms. SEM ranged from 2.95 (test-retest form 1 intelligibility) to 4.85 (test-retest form 2 phonetic accuracy). SEM was consistently higher for phonetic accuracy than for intelligibility scores, suggesting that children's ranks changed more with phonetic accuracy scores than with intelligibility scores. Possible reasons for the higher SEM associated with phonetic accuracy scores include differences among listeners in the amount of previous experience judging sounds as clear and distorted, lack of a clear definition for the two possible responses and ambiguous acoustic cues in the children's sound productions (Shriberg & Lof, 1991). While listeners were instructed to focus on the underlined sound(s) when choosing a rating, there may have also been differences in how listeners applied the rating, with some listeners rating only the target sound and others rating the whole word. In the examination of sources of variance affecting the reliability of phonetic transcription by Shriberg and Lof (1991), acoustic ambiguity was presented as the primary explanation for low inter-judge and intra-judge agreement on diacritics.

Inter-rater reliability was high (i.e., ICC > 0.9) for the groups of three listeners used to obtain *SIP-CCLP* scores. ICCs were also high for each possible combination of two listeners. Although the range of standard error of measurement was larger for the groups of two listeners (4.9 – 7.5) than for the groups of three listeners (5.2 – 6.8), the results suggested that *SIP-CCLP* scores obtained from a minimum of two listeners would be reliable. This result was

confirmed through the evaluation of the dependability of *Ver. 5* scores using generalizability theory.

Intra-rater reliability was examined for a subset of listeners who participated in two sessions one week apart. While intraclass correlation coefficients were high (> 0.9) and SEMs were relatively small (i.e., < 3.5), *SIP-CCLP* scores tended to increase from session one to session two, suggesting that a single exposure to a child's speech changes listeners' ability to identify target phonemes. Further research is needed to determine if increasing the amount of time between sessions minimizes this familiarization effect.

Internal consistency coefficients of 0.93 and 0.92 were obtained for form 1 and 2, respectively. This is within the range recommended by the Scientific Advisory Committee of the Medical Outcomes Trust (2002) for measures that will be used to assign scores to individuals.

In summary, the reliability of *SIP-CCLP Ver. 5* was found to be acceptable. Intraclass correlation coefficients for test-retest and inter-rater reliability and Cronbach's alpha for internal consistency exceeded the minimum standard of 0.75 needed to achieve a grade of A on Andresen's tool for assessing the quality of tools used to measure disability outcomes (2000). However, the two forms developed for *Version 5* are not parallel, as scores obtained using form 1 were lower than scores obtained using form 2. This finding was confirmed using limits of agreement (Bland & Altman, 1986), generalizability theory and item response theory. An item analysis was conducted to revise the two forms

(Appendix D). A future study will determine if this revision achieves form equivalence.

Validity. According to Kirshner and Guyatt (1985), test developers of discriminative health status measures are most concerned with establishing the measure's construct-related validity. To this end, the relationships between test scores and variables related to the construct that the test is purported to measure were examined. Validity of *SIP-CCLP Ver. 5* was assessed by examining relationships of its scores to 1) scores obtained from other measures of intelligibility and 2) measures of related constructs. While it is recognized that there is not a gold standard for measuring speech intelligibility, in the current study examination of the relationships between *SIP-CCLP Ver. 5* scores and scores obtained from other measures of intelligibility were referred to as criterion-related validity.

SIP-CCLP intelligibility scores were expected to be moderately positively correlated with intelligibility scores obtained from two "criterion" measures: a conversational speech and an imitative sentence sample (*TOCS+* Intelligibility Measures; Hodge, Daniels & Gotzke, 2009). The *TOCS+* sentence test is a computer-mediated measure of sentence-level speech intelligibility appropriate for young English-speaking children. Previous research with children with articulation/phonological disorders of unknown origin suggests that intelligibility scores obtained from an imitated word sample have similar relationships to intelligibility scores obtained from continuous speech and imitated sentences (r =. 79 and r = .81 respectively; Gordon-Brannan & Hodson, 2000). An imitative

sentence task has the added advantage of being a more efficient means of obtaining a connected speech sample from children and judgments from listeners compared to conversational or spontaneous speech. In this study, the correlations between *Ver. 5* scores (imitated words) and intelligibility scores obtained from a spontaneous speech sample (form 1: r = .61; form 2: r = .51) were slightly lower than those reported previously by Gordon-Brannan and Hodson (2000). However, the correlations between *Ver. 5* scores and *TOCS+* intelligibility scores were higher than expected (form 1: r = .90; form 2: r = .89). The mean difference between Form 1 and *TOCS+* intelligibility scores was -4.6% (SD = 8.7, Form 1 – *TOCS+*). The mean difference between Form 2 and *TOCS+* intelligibility scores was 0.3% (SD = 9.3, Form 2 – *TOCS+*). These results suggest that *SIP-CCLP* and *TOCS+* are measuring the same construct (i.e., imitative speech intelligibility). However, while *SIP-CCLP* scores and intelligibility scores obtained from a 100-word conversational speech sample are related, they do not appear to measure the same construct. Conversational speech intelligibility may be affected by factors (e.g., grammar, type-token ratio, linguistic planning load, speaking style) that are controlled in imitative samples. Further examination of the children's spontaneous speech samples using *Systematic Analysis of Language Transcripts* (SALT Software, LLC, 2012) and/or phonetic analysis may provide additional insight into what factors account for differences in intelligibility scores in the two conditions.

A secondary analysis of the correlation between intelligibility scores obtained from the imitative sentence task and conversational samples was conducted to better understand the relationship between different measures of speech intelligibility for young children with cleft palate. A significant positive correlation between intelligibility scores from these two samples was obtained ($r = .76$, $p < .01$), similar to the correlation reported by Gordon-Brannan and Hodson (2000) ($r = .85$). Similarities associated with connected speech samples (e.g., contextual cues, speaking rate, coarticulation) may account for the higher correlation between *TOCS+* and the spontaneous speech sample than between *SIP-CCLP* scores and the spontaneous speech sample.

SIP-CCLP intelligibility scores were expected to be moderately positively correlated with percentage of consonants correct and moderately negatively correlated with ratings of hypernasality and voice severity. Based on previous research that examined these relationships for speakers with cleft palate (e.g., Whitehill & Chun, 2002), percentage of consonants correct was expected to be the single variable that accounted for the greatest amount of variation in *SIP-CCLP* scores. Between 60 – 69% of the variance in *SIP-CCLP* intelligibility (form 1 only) and phonetic accuracy scores was explained by two predictors: percentage of consonants correct and hypernasality ratings. This result suggests that as articulation accuracy decreased and/or hypernasality increased, *SIP-CCLP* scores decreased. Percentage of consonants correct explained 51% of the variance in form 2 intelligibility scores and hypernasality ratings did not contribute additional unique variance. Severity may be the overarching variable underlying these

relationships. Unexpectedly, voice severity scores were not correlated significantly with *SIP-CCLP* scores. This may reflect a limitation of the single word phonetic contrast approach used in *Ver. 5* to measure intelligibility.

Hypernasality and voice severity ratings were obtained using direct magnitude estimation with a modulus. This method was an efficient, reliable and valid means of estimating hypernasality and voice severity in this study. Chapman, Hardin-Jones, Goldstein, Halter, Havlik and Schulte (2008) also used direct magnitude estimation with a modulus to rate articulation proficiency. In the current study, phonetic transcription was used to obtain a measure of articulation accuracy – percentage of consonants correct. Reliability of phonetic transcription was lower than reported in other studies (e.g., Zajac, Plante, Lloyd & Haley, 2011) likely due to the transcribers' inexperience in transcribing the speech of children with cleft palate. Direct magnitude estimation with a modulus may be a more reliable and efficient means of estimating articulation for young children with cleft palate. Research is needed to determine if measures of articulation obtained using different methods have similar relationships to *SIP-CCLP* scores.

Item reduction. When identifying the final set of items to be included in a discriminative health status measure, items that discriminate between examinees are retained and those that do not are deleted (Kirshner and Guyatt, 1985). A classical item analysis was conducted to identify which *SIP-CCLP Ver. 5* items should be deleted or revised (Appendix D). As a result of this analysis, fourteen items were flagged for deletion as they did not discriminate between the children.

Seven of these items were identified correctly for at least 17 of the 20 children (difficulty index $\geq$ 0.85), while the other seven items had discriminative indices less than 0.2. These deletions reduced the number of items from 63 to 49. However, it was recommended that at least some of these items be retained to ensure that the content domain was sampled adequately and to decrease the respondent burden of children, parents and listeners.

This item analysis also identified three items in need of revision. For these three items, the discriminative index was low only for one member of the stimulus word pair. Examination of the lexical characteristics considered in the construction of *Ver. 5* (i.e., word frequency, number of neighbours for children and listeners) suggested that, for two of the items, word frequency was the reason why the discriminative indices were different. In each case, the word with the higher word frequency was less discriminating. This result suggests that word frequency be considered when developing speech measures for children.

Guidelines for Administering, Scoring and Interpreting *SIP-CCLP Ver. 5*

Administering *SIP-CCLP* to obtain word recordings. The *SIP-CCLP Ver. 5* software allows users to selected whether the pictures, pre-recorded models, animations, and "beep" to cue responses are turned "on" or "off" during administration and whether instructions are played prior to administration. In this evaluation, the pictures, pre-recorded models, animations and "beep" were turned "on," instructions were played and the familiarization training was completed for all children. If the child required frequent verbal and visual reminders to wait for the "beep" before repeating the word, the "beep" was turned off after

familiarization training. Future users are encouraged to select these same options when administering *SIP-CCLP Ver. 5* to children. However, as the pre-recorded models were obtained from a Western Canadian speaker, it is recommended that users from areas with different English dialects obtain models of the *SIP-CCLP* stimulus words from a representative male speaker prior to administration. The design of the software allows replacement of the existing audio models of the stimulus words with new audio models of the same words with relative ease. Furthermore, users are encouraged to review the picture stimuli to ensure that the objects pictured are likely to be familiar to children in their geographic area (e.g., egg dying as pictured for the word "dye" may not be familiar). New pictures may need to be developed or children familiarized with the word and picture stimuli prior to administration.

Recording levels, recording environment, and recording hardware and peripherals (i.e., computer, microphone, preamplifier) are all factors that may affect the quality of recordings obtained using *Ver. 5*. In this study, recording levels were checked and adjusted prior to administration of the *SIP-CCLP* to each child and again during familiarization training. Adjustments to the recording level were primarily made using the external preamplifier. *SIP-CCLP* should be administered in a quiet environment.

Administering *SIP-CCLP* to obtain listener judgments. At the beginning of the listening task, listeners were given a short verbal summary describing the task they were about to complete. During this summary, listeners were not given any information about the age or gender of the child or the effects

of cleft palate on speech. After the practice items, the administrator asked listeners if they had any questions and if the volume of the playback was adequate. Listeners were debriefed at the end of the session. It is recommended that future users follow these same practices when administering *SIP-CCLP* to listeners.

In this study, listeners were university students ranging in age from 18 to 39 years with English as a first language and hearing within normal limits. They were not provided with training judging the speech of children with cleft palate. This population may not be representative of listeners in the children's environment (e.g., teachers, peers, family) or in the clinical environment (e.g., hospital). Test users are cautioned that scores obtained using *SIP-CCLP Ver. 5* may be higher if listeners familiar with the speech of children with cleft palate are used. Because listeners' familiarity with a child affects intelligibility scores, test users should always describe the characteristics of their listeners when reporting intelligibility scores obtained when using *SIP-CCLP* and use listeners with the same characteristics when comparing scores to other children or the same child over time.

Listener participation. The second objective of this project was to provide guidelines to test users about listener participation based on an examination of the effects of listener familiarity with speaker, test stimuli, and listening task on *SIP-CCLP* scores. Previous research has found that repeated exposure to the same speaker and same stimuli increased listeners' word identification scores for synthetic speech using a six-choice closed-set response

task (Greenspan, Nusbaum & Pisoni, 1988; Schwab, Nusbaum & Pisoni, 1985). However, the difference in means for the two listening occasions was less than 3% for both studies. In these studies, listeners heard recordings produced by "a speaker" with fixed severity. Each listener judged the recordings two times with inter-judging intervals set at either six or ten days.

SIP-CCLP scores increased with each exposure. The mean difference was less than 5% after three exposures to the same child's speech and more than 5% after four exposures to the same child's speech. As observed in Chapter 5, the increase in intelligibility scores was greatest for the child with the lowest PCC score after both three and four exposures but not for phonetic accuracy scores. Based on these findings, the results of the evaluation of inter and intra-rater reliability using classical test theory, and the evaluation of dependability using generalizability theory, guidelines for listener recruitment were developed. It is recommended that *SIP-CCLP* scores be based on the responses from a minimum of two listeners. If a listener judges *SIP-CCLP* recordings from the same child within a one-week period, test users can expect that scores will be on average 4% higher due to increased familiarity with the child's speech. Test users should not have the same listener judge the same child's recordings more than three times in a four-week period.

Scoring and interpretation of *SIP-CCLP* scores. Following administration of the *SIP-CCLP Ver. 5* response task to three listeners, the analysis component of the software was used to compile and analyze the listeners' responses. Prior to analyzing the responses, the software presents the entries in

the "blank" response option to the test administrator for verification or recoding as the target or foil word. Careful attention is needed to ensure that these items are coded correctly. After the check is completed, the software generates an analysis file that compares the listeners' responses on the *Ver. 5* items and presents the intelligibility and phonetic accuracy scores for each listener and the mean of the three listeners. The output also summarizes the number of phonetic contrast items correct and incorrect for each of the five error categories (i.e., manner preference error, place preference error, voicing error, sibilant error, and consonant cluster error). A sample of the analysis output is included in Appendix O.

SIP-CCLP Ver. 5 was developed to allow users to distinguish between children with respect to their speech intelligibility. SEMs were used to calculate the minimal detectable change (MDC), defined as the minimum difference between two scores from two subjects needed for their scores to be considered different (Weir, 2005). The largest estimate of MDC was for alternate forms over time which was 11.7% and 12.6% for intelligibility and phonetic accuracy scores, respectively. Therefore, a 12% difference in intelligibility scores (or a 13% difference in phonetic accuracy scores) from two individuals would be considered to be a difference and not a difference consistent with the measurement error of *SIP-CCLP Ver. 5*. It is important to note that MDC is a statistical construct that may not take into account what constitutes a clinically important difference in scores from the perspective of clinicians or clients.

Intelligibility scores have been used as a way to describe severity of an individual's speech impairment. For example, Gordon-Brannan and Hodson (2000) used intelligibility scores from a continuous speech sample to divide their sample of 48 children, aged 48 to 66 months, into four quartile groupings: adult-like speech (91 – 100% intelligible); mild speech involvement (83 – 90% intelligible); moderate involvement (68 – 81% intelligible); and severe (16 – 63% intelligible). Visual examination of the intelligibility scores from the children in this study (Figure 6-1) suggests three levels of severity: mild involvement (i.e., intelligibility score 78% - 90% on form 1 and 84% - 96% on form 2; 1 MDC), moderate involvement (i.e., intelligibility score 53% - 77% on form 1 and 59% – 83% on form 2; 2 MDC) and severe-to-profound involvement (i.e., intelligibility score $\leq$ 52% on form 1 and $\leq$ 58% on form 2; >3 MDC). Comparison of the children's severity levels on the two forms revealed that only one child changed in terms of severity from form 1 (mild) to form 2 (moderate). Three levels of severity are also suggested for phonetic accuracy scores: mild involvement (i.e., 72% - 85% on form 1 and 74% - 87% on form 2; 1 MDC), moderate involvement (i.e., 45% - 71% on form 1 and 47% – 73% on form 2; 2 MDC) and severe-to-profound involvement (i.e., $\leq$ 44% on form 1 and $\leq$ 46% on form 2). One child changed from a severity level of severe-to-profound on form 1 to moderate on form 2 using phonetic accuracy scores (Figure 6-2).

The *SIP-CCLP Ver. 5* analysis provides two summary scores: intelligibility and phonetic accuracy. An advantage of *Ver. 5* intelligibility scores is that they are calculated in a manner consistent with how intelligibility scores

are calculated in other measures of speech intelligibility (e.g., *Children's Speech Intelligibility Measure (CSIM)*; Wilcox & Morris, 1999) and are, therefore, familiar to speech-language pathologists, increasing the ease with which they may be interpreted and explained to families. An advantage of the phonetic accuracy score is that it provides information about how distortions may be contributing to a child's speech impairment by assigning more points to items that are identified correctly and assigned a rating of "clear" than to items that are identified correctly and assigned a rating of "distorted." As a result of this differential weighting, phonetic accuracy scores tend to be lower than intelligibility scores, decreasing the likelihood that children with mild speech impairments will ceiling using these scores. While phonetic accuracy and intelligibility scores are strongly correlated (F1: $r = .98$; F2: $r = .97$), they are not equivalent (F1: $t = 13.95$, $p < .01$; F2: $t = 12.86$, $p < .01$). For form 1, intelligibility scores ranged from 28.0 to 90.5% (mean = 63.8, SD = 18.5) and phonetic accuracy scores ranged from 20.1 to 85.5% (mean = 54.0, SD = 18.4). For form 2, intelligibility scores ranged from 32.3 to 95.8% (mean = 68.7, SD = 17.1) and phonetic accuracy scores ranged from 22.8 to 87.0% (mean = 57.7, SD = 17.1). In this study, the mean difference between these scores was greater for the 13 children with intelligibility scores between 50 and 84% (F1: mean = 11.2%, SD = 2.9; F2: mean = 12.6%, SD = 3.4) than for the children with intelligibility scores less than 50% (n = 4) or greater than 84% (n = 3). However, the ranking of the children was the same for all but two of the children on each form when their ranks on intelligibility and phonetic accuracy scores were compared. Although both the minimum and maximum

scores were lower for phonetic accuracy scores than intelligibility scores, the size of the range was similar for both scores on both forms suggesting that phonetic accuracy scores do not provide better discrimination of the children. Therefore, it is recommended that users report intelligibility scores instead of phonetic accuracy scores. However, users are cautioned that a "high" intelligibility score (e.g., > 90%) may not mean that a child does not have a speech impairment, as there may be remaining speech sound distortions. In these situations, the phonetic accuracy score provides additional useful information about the perceived accuracy of the child's speech. For example, CP11's intelligibility scores on form 1 and 2 were 90.5% and 95.8%, respectively, but her phonetic accuracy scores were 85.5% and 87%. Examination of her *Ver. 5* analysis output reveals that a minimum of two of the three listeners chose "distorted" for two items containing /r/ and identified two instances of gliding (MPE) for form 1. Similarly, on form 2, a minimum of two of the three listeners chose "distorted" for seven items containing /r/ and identified two instances of gliding (MPE). Based on these results, it is hypothesized that CP11 requires speech therapy on production of /r/ and, that following successful treatment, her phonetic accuracy scores (and possibly her intelligibility scores) will increase.

Limitations and Strengths

For the evaluations of reliability and validity of *SIP-CCLP Ver. 5*, recordings of form 1 and 2 were obtained from 20 children with cleft palate. While the recruitment goal for this study was met, Charter (1999) recommended that a minimum sample size of 400 subjects is required to have stable, precise

estimates of reliability and validity coefficients. A multi-center international study evaluating the reliability and validity of *SIP-CCLP* is needed to achieve this sample size. A study of this scope would also facilitate examination of whether there are differences in the occurrence of some of the *SIP-CCLP* error patterns which are related to variation in surgical timing and procedure or access to services in different geographic areas.

The goal of creating two parallel forms was not achieved. As a result, scores obtained using form 1 with one child and form 2 with another child cannot be compared directly. The item analysis (Appendix D) revealed that form 2 has more "easy" items than form 1 and identified items to be exchanged and revised to improve the similarity of the two forms with respect to difficulty. Administration of the revised forms will determine if these changes created two parallel forms. Test developers following the classical test theory model have developed a statistical means of establishing equivalent scores on two forms of a test that measure the same construct (i.e., equating) but do not meet the conditions for being parallel (Crocker & Algina, 1986). In order to equate scores from two forms, two assumptions must be met: 1) the two forms measure the same construct with equal reliability and 2) the percentile ranks corresponding to the scores are equal on the two forms. Data from a larger sample size are needed before these assumptions can be tested and equating applied. A major challenge faced in creating parallel forms for *SIP-CCLP Ver. 5* was identifying words for minimal pairs that were appropriate for young children and similar in phonetic structure, word frequency and phonological neighbourhood size.

Another limitation of this study is that the severity of the speech intelligibility impairment was not the same for the groups of children assigned to the two orders of form presentation (i.e., order 1: form 1– form 2 – form 2 – form 1; order 2: form 2 – form 1 – form 1 – form 2). As a result, order was identified as a significant main effect in the evaluation of the dependability of *Ver. 5* using Generalizability theory (Appendix E). Use of an intelligibility rating (e.g., *Intelligibility in Context Scale*; McLeod, Harrison & McCormack, 2012) obtained prior to the initial session to guide assignment of children to each order to control equivalence of participant severity for each order is recommended in future studies.

According to Kent et al. (1989), one aspect of the construct-related validity of measures of intelligibility developed using a phonetic contrast approach is that their results confirm or extend the results of articulation testing. Therefore, the error patterns identified using *Ver. 5* should be comparable to those obtained through phonetic transcription. However, phonetic transcription of the children's *SIP-CCLP* word productions was not completed as part of this study. It is recommended that the *SIP-CCLP* words be transcribed by an individual with experience transcribing the speech of children with cleft palate who is not familiar with the children or words. This would also provide insight into the stability of children's productions of the same words presented in each form and the relationship between listeners' use of "can't identify" and non-English sound substitutions.

One of the strengths of the *SIP-CCLP Ver. 5* is that its revision was guided by past experience with previous versions (e.g., Gotzke, 2005), guidelines for designing single-word lists to be used in assessment of cleft palate speech (EUROCRAN, 2009; Sell, Harding & Grunwell, 1999), and new research evaluating its content-related validity. Unlike the measure of intelligibility for children with cleft lip and palate developed by Zajac et al. (2011), which is a revised version of a pre-existing published measure (*CSIM*, Wilcox & Morris, 1999) with new software facilitating its administration to children and listeners, *Ver. 5* represents a complete revision of previous versions addressing everything from target error patterns and stimulus words to software. These revisions improved its sensitivity to the error patterns of children with cleft palate by ensuring that the error patterns targeted are relevant and representative for this population by examining results obtained with *Ver. 3* (Gotzke, 2005), reviewing research describing the speech errors of this population, and conducting an evaluation of content-related validity with experts in the area of cleft palate speech. In addition, guidelines for sampling the speech of children with cleft palate and age-appropriateness of the vocabulary were used to select stimulus words and thereby, improve sensitivity. Unlike *CSIM* (Wilcox & Morris, 1999) or Zajac et al.'s measure (2011), where words not found in frequency of usage counts were simply included, the *Ver. 5* software contains a familiarization activity to help introduce potentially unfamiliar vocabulary. Increasing the sensitivity of *Ver. 5* also increases the efficiency with which *SIP-CCLP* can be administered to children (average time to administer = 10 minutes) and listeners

(average time to administer = 10 minutes). Consequently, the user can obtain reliable and valid information about a child's imitative single word speech intelligibility and the error patterns that may be underlying the impairment within an hour.

The evaluation of *SIP-CCLP* is more rigorous and complete than evaluation conducted for other measures of speech intelligibility for children with cleft palate. Both *SIP-CCLP Ver. 5* and the measure developed by Zajac et al. (2011) were developed to index severity of speech intelligibility impairments. However, only *SIP-CCLP* was developed and evaluated using criteria outlined by Kirshner and Guyatt (1985) for measures with a discriminative purpose. It should be noted that the group of children used in the evaluation of the measure developed by Zajac et al. (2011) is not representative of the population of speakers with cleft palate, as only one of the 22 children in the study has a PCC score less than 70% and only children with cleft lip and palate older than 59 months were included. In the current study, children were more representative of the population of children with cleft palate, as children with different cleft types, with and without syndromes ranging in severity of speech impairment from mild to severe (as indicated by PCC on the spontaneous speech sample) and age from 37 to 84 months were included. For *Ver. 5*, reproducibility over time, forms and raters and internal consistency was examined following the standards developed by AERA, APA and NCME (1999). While reproducibility over forms and raters was examined by Zajac et al. (2011), results were reported for children with and without cleft lip and palate as a whole, not just for the population of interest (e.g.,

children with cleft palate) as per AERA, APA and NCME (1999) standards. Furthermore, standard error of measurement was not reported by Zajac et al. (2011); therefore, it is not possible to construct confidence intervals for children's scores. While content-related, construct-related and criterion-related validity were all evaluated for *Ver. 5,* only construct-related validity was examined by Zajac et al. (2011). One aspect of construct-related validity that was examined by Zajac et al. (2011) but not in the current study was a comparison of results for children with and without cleft palate. While this area needs to be addressed for *Ver. 5*, previous research conducted using *Ver. 3* (Gotzke, 2005) found that scores were significantly lower for children with cleft palate than for children without cleft palate. The construction and validation of *Ver. 5* outlined in this study may serve as a model to others who undertake development and evaluation of intelligibility measures.

This study provides new information about how number of exposures, speaker severity and exposure to different sets of words influence children's intelligibility scores obtained using a closed-set response task. The results of this evaluation have implications for how listeners are selected when conducting intelligibility assessment using *SIP-CCLP* or other measures of intelligibility that use a closed-set response task to obtain scores (e.g., *CSIM*, Wilcox & Morris, 1999). Neither Zajac et al. (2011) nor Wilcox and Morris (1999) provide guidelines for listener participation for their measures even though characteristics of the listener are recognized as a variable that may affect a speaker's intelligibility score (Walshe, Miller, Leahy & Miller, 2008). While this study

examined effects of repeated exposure when there was one week between administrations of the same form or a different form from the same speaker, further research is needed to determine if increasing the amount of time between administrations decreases the effect of repeated exposure on scores.

Future Research

In addition to the suggestions for future research identified in the previous section, the use of *SIP-CCLP Ver. 5* as an evaluative measure and the relationships between *Ver. 5* scores and acoustic and/or physiological measures could also be investigated. Clinically, intelligibility is used as a measure of speech disorder severity and intervention success (Whitehill, 2002). In this dissertation, the use of *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility with cleft palate was evaluated. The results of this research support using *SIP-CCLP* when the goal of measurement is to describe differences in severity of a speech disorder among a group of speakers with the same underlying condition. However, the use of *SIP-CCLP* as an evaluative outcome measure to quantify functional change following surgical, prosthetic, and/or speech intervention (e.g., intervention success) has not yet been investigated. If a measure is to be used for a different purpose than described, the test user must justify the new use, which may require re-evaluating item selection, item scaling, item reduction, reliability and validity of the measure for the new purpose (AERA, APA & NCME, 1999). In addition, for evaluative outcome measures, responsiveness, defined as the "power of the test to detect a clinically important difference" (Kirshner & Guyatt, 1985; p. 29), must also be evaluated. One

strategy that has been used to evaluate responsiveness is to examine whether scores improve following treatment known to have a desirable effect on the construct measured by a test. *SIP-CCLP* Ver. 5 recordings were collected for one child with cleft palate pre- and post-surgical intervention (i.e., insertion of a pharyngeal flap, palatal re-repair, and lip revision) during the study. Results from the two testing sessions are compared in Appendix N. For this child, both *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores increased by more than the minimal detectable change following surgical intervention, which suggests that *Ver. 5* has promise as an evaluative measure for speech intelligibility for young children with cleft palate. Future research will address the ability of *SIP-CCLP* to provide information about differences in intelligibility within a child over time. To facilitate the use of *SIP-CCLP* for multiple purposes, future software revisions may allow the test administrator to select whether scores are desired to provide a measure of severity (i.e., discriminative purpose) or outcome following intervention (i.e., evaluative purpose) prior to the analysis of listener results. This selection would direct the software to use only the items evaluated for the chosen purpose in the calculation of scores.

In this study, the relationships between *SIP-CCLP* scores and perceptual ratings of speech characteristics were examined as part of the evaluation of construct-related validity. Research addressing the relationship between *SIP-CCLP* scores and acoustic and/or physiological measures would provide additional insight into the speech characteristics underlying speech intelligibility in young children with cleft palate. Zajac et al. (2011) examined the relationship

between velopharyngeal closure, determined using pressure-flow testing, and speech intelligibility measured using an imitative single-word task. There was no significant difference in intelligibility scores for the children classified as having adequate closure and those classified as having inadequate closure. In the current study, nasalance scores on the *SNAP* sentences (Kummer, 2005) were collected for 18 of the 20 children as descriptive information but the relationships between these scores and *SIP-CCLP* scores were not examined. Additional acoustic measures of velopharyngeal closure that could be investigated include amplitude and spectral moments analysis of stop burst, and voice onset time.

Conclusions

The purpose of this dissertation was to develop and evaluate the reliability and validity of *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility for young children with cleft palate using a rigorous set of standards developed by AERA, APA and NCME (1999). The significance of the results described in the dissertation is summarized as follows:

1) *SIP-CCLP Ver. 5* is more sensitive, efficient and has greater utility than previous versions.

1.1 Stimulus words were selected on the basis of age-appropriateness and fit with guidelines for the phonetic content of single-word speech samples for individuals with cleft palate (e.g., EUROCRAN, 2009).

1.2 Word frequency and neighbourhood density were considered when balancing the two forms.

1.3 Content-related validity of the error patterns was evaluated using results from *Ver. 3*, published research and expert assessment.

1.4 The number of response options was increased to reduce the chance probability of listeners choosing the target, and to improve the sensitivity of *Ver. 5*.

1.5 The number of words elicited and items judged decreased, reducing the administration time for both children and listeners to about 10 minutes.

1.6 Two forms were developed.

2) Acceptable levels of test-retest, alternate forms, inter-rater and intra-rater reliability were obtained.

2.1 Form 1 and 2 did not satisfy the conditions for parallel forms using classical test theory. Revised forms were created as part of the item analysis and a future study is planned to evaluate their equivalence.

2.2 Appropriate statistics were used to assess the reproducibility of scores: intraclass correlation coefficients and standard error of measurement.

2.3 Reliability coefficients were greater than 0.9 for all evaluations.

2.4 Standard error of measurement was less than 5% for both intelligibility and phonetic accuracy scores for test-retest and alternate forms reliability.

2.5 The most conservative value for minimal detectable change was 12% for intelligibility scores and 13% for phonetic accuracy scores.

3) The validity of using *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility for children with cleft palate was supported.

3.1 Listeners identified errors in all five error categories in *Ver. 5* with the majority of errors identified from the manner preference error category.

3.2 All but two of the error patterns were identified for at least one child, confirming the content relevance and representativeness of the error patterns. One additional error pattern that was not tested was identified by listeners using the "blank" option and will be included in future versions.

3.2 Moderate-to-strong positive correlations were obtained between *SIP-CCLP* scores and intelligibility scores from two criterion measures: a 100-word spontaneous speech sample and the *TOCS+* sentence intelligibility test.

3.3 As predicted, *Ver. 5* scores were moderately correlated with percentage of consonants correct and hypernasality ratings.

4) *SIP-CCLP Ver. 5* is a better discriminative measure of speech intelligibility for young children than existing measures.

4.1 *SIP-CCLP Ver. 5* was developed and evaluated with particular attention to the purpose of the measure, as per Kirshner and Guyatt (1985).

4.1 Rigorous evaluation of validity and reliability (see 2 and 3) using appropriate methods and following AERA, APA and NCME (1999) standards for reporting was conducted.

4.3. Guidelines for administration to obtain word recordings and listener judgments and for scoring and interpretation are provided. Guidelines for listener participation are based on research conducted using the measure.

4.4 Intelligibility scores provide a summary measure of the severity of a child's speech intelligibility impairment, while phonetic accuracy scores provide additional information about distortions that may be contributing to perceived mild speech impairments.

In conclusion, the results support the use of *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility that is appropriate for young children with cleft palate, time-efficient, reliable, valid, and suitable for research and clinical applications.

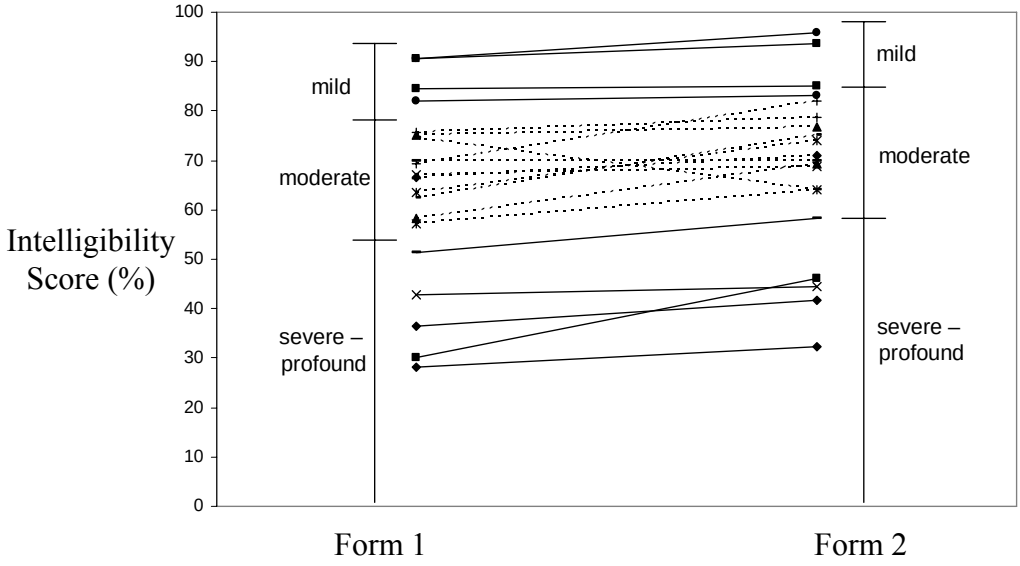

Figure 6-1. Severity groupings for form 1 and form 2 intelligibility scores.

Note. Children assigned a rating of mild or severe on form 1 are indicated by a black line. Children assigned a rating of moderate on form 1 are indicated by a dotted line.

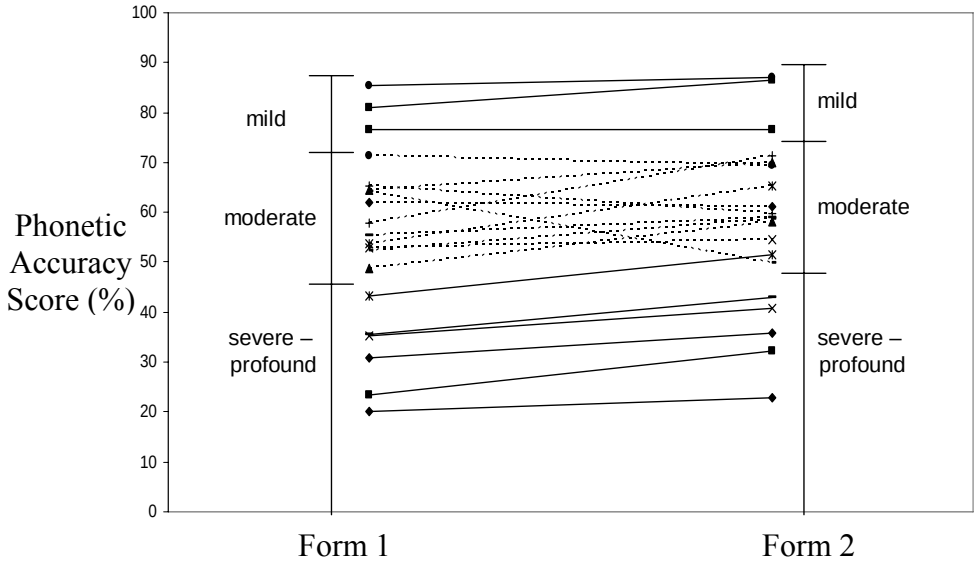

Figure 6-2. Severity groupings for form 1 and form 2 phonetic accuracy scores.

Note. Children assigned a rating of mild or severe on form 1 are indicated by a black line. Children assigned a rating of moderate on form 1 are indicated by a dotted line.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation, 81(12 Suppl)*, S15-S20.

Bland, J. M., & Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, i,* 207 -310.

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

Charter, R.A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21(4)*, 559-66.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

European Collaboration in Craniofacial Anomalies. (2009). *Single word lists*. Retrieved from http://www.eurocran.org/content.asp?contentID=1387

Gordon-Brannan, M., & Hodson, B. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology, 9*, 141-150.

Gotzke, C. L. (2005). *Speech intelligibility probe for children with cleft palate version 3: Assessment of reliability and validity*. (Unpublished master's thesis). University of Alberta, Edmonton, AB.

Greenspan, S. L., Nusbaum, H.C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(3)*, 421-433.

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

Hodge, M. M., Daniels, J., & Gotzke, C.L. (2009). *TOCS+ Intelligibility Measures.* Edmonton, Alberta: University of Alberta.

Hodge, M., & Gotzke, C. L. (2007). Preliminary results of an intelligibility measure for English-speaking children with cleft palate. *Cleft Palate-Craniofacial Journal, 44*(2), 163-174.

Kent R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*, 482-499.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disorders, 38*(1), 27-36.

Kolson, C. J. (1960). *The Vocabulary of Kindergarten Children*. (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Kummer, A. W. (2005). *The MacKay-Kummer SNAP Test-R: Simplified Nasometric Assessment Procedures*. KayPENTAX. Retrieved from http://www.kayelemetrics.com/index.php?option=com_product&view=product&Itemid=3&controller=product_innerpage&rec_id=50&no_id=2

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. *Cleft Palate-Craniofacial Journal, 46*(4), 347-362.

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.

SALT Software, LLC. (2012). *Systematic Analysis of Language Transcription* [computer software]. Middleton, WI: SALT Software, LLC.

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

Schwab, E.C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27(4)*, 395-408.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11*,193-205.

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

Stemach, G., & Williams, W. B. (1988). *Word express: The first 2, 500 words of spoken English illustrated*. Novato, CA: Academic Therapy Publications.

Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language & Communication Disorders, 43(6)*, 633-648.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the *SEM*. *Journal of Strength and Conditioning Research, 19(1)*, 231-240.

Whitehill, T. (2002). Assessing intelligibility in speakers with cleft palate: A critical review of the literature. *Cleft Palate-Craniofacial Journal, 39(1)*, 50-58.

Whitehill, T. L., & Chau, C. H. (2004). Single-word intelligibility in speakers with repaired cleft palate. *Clinical Linguistics and Phonetics, 18(4-5)*, 341-355.

Whitehill, T., & Chun, J. C. (2002). Intelligibility and acceptability of speakers with cleft palate. In: F. Windsor, M. L. Kelly, & N. Hewlett (Eds.), *Investigations in clinical phonetics and linguistics* (pp. 405-415). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilcox, K., & Morris. S. (1999). *Children's Speech Intelligibility Measure*. San Antonio, TX: The Psychological Corporation.

Zajac, D. J., Plante, C., Lloyd, A., & Haley, K. L. (2011). Reliability and validity of a computer-mediated, single-word intelligibility test: Preliminary findings for children with repaired cleft lip and palate. *Cleft Palate-Craniofacial Journal, 48(5)*, 538-549.

Appendix A

SIP-CCLP Ver. 3 Results, Supporting References and Expert Assessment of the Error Patterns Sampled in *SIP-CCLP Ver. 5*

	SIP-CCLP Ver. 5.0	*SIP-CCLP Ver. 3 Results*		References Describing Error Pattern	Expert Assessment of Error Patterns
		Number of Times Identified	Number of Children		
Manner Preference Errors	Glides for obstruents	2	2	Harding & Grunwell, 1996; Stokes & Whitehill, 1996	less than 10%
	Liquids for obstruents	4	3	Chapman, 1993; Morris & Ozanne, 2003	less than 10%
	Nasals for obstruents	17	6	Harding & Grunwell, 1996; Prins & Bloomer, 1965; Chapman, 1993; Bzoch, 1965	more than 10%
	Nasals for liquids	9	2	Lynch, Fox & Brookshire, 1983; Morris & Ozanne, 2003; Prins & Bloomer, 1965;	no agreement
	Stopping	20	7		more than 10%
	Deaffrication	8	3	Chapman, 1993; Morris & Ozanne, 2003	less than10%
	Gliding of liquids	16	5	Chapman, 1993; Morris & Ozanne, 2003	more than 10%
	Oral fricatives for liquids and glides	1	1	Chapman, 1993; Chapman & Hardin, 1992	less than 10%
	Oral stops for nasals	2	1	Harding & Grunwell, 1996; Peterson-Falzone, Trost-Cardamone, Karnell, & Hardin-Jones, 2006	less than 10%
	Affricates for oral stops	2	1	Chapman, 1993; Lynch, Fox & Brookshire, 1983	less than 10%
	Affricates for fricatives	3	1	Chapman, 1993; Lynch, Fox & Brookshire, 1983; Stokes & Whitehill, 1996	less than 10%

	SIP-CCLP Ver. 5	*SIP-CCLP Ver. 3* Results		References Describing Error Pattern	Expert Assessment of Error Patterns
		Number of Times Identified	Number of Children		
Place Preference Errors	Bilabial stops for alveolar stops	6	4	Chapman, 1993; Chapman & Hardin, 1992; Eurocleft Speech Group, 1993	less than 10%
	Velar stops for obstruents (i.e., oral stops, fricatives, affricates)	26	7	Harding & Grunwell, 1996; Morris & Ozanne, 2003; Ruiter, Korsten-Meijer & Goorhuis-Brower, 2009; Stokes & Whitehill, 1996	more than 10%
	Glottal stops for oral sounds a. stops b. fricatives c. affricates d. sonorants	 5 4 0 1	 3 3 0 1	Chapman, 1993; Harding & Grunwell, 1992; Lynch, Fox, & Brookshire, 1983; Morris & Ozanne, 2003	 a. more than 10% b. less than 10% c. more than 10% d. less than 10%
	Glottal fricatives for oral sounds a. fricatives b. affricates	 5 1	 1 1	Bzoch, 1965; Harding & Grunwell, 1996; Morris & Ozanne, 2003, Stokes & Whitehill, 1996	 a. more than 10% b. less than 10%
	Alveolar stops for velar stops	18	6	Chapman, 1993; Harding & Grunwell, 1996	no agreement
	Alveolar stops for bilabial stops	7	5	Harding & Grunwell, 1996; Lynch, Fox, & Brookshire, 1983	less than 10%
	Alveolar fricatives for labiodental and interdental fricatives	9	6	Harding & Grunwell, 1996; Lynch, Fox, & Brookshire, 1983	less than 10%

	SIP-CCLP Ver. 5	*SIP-CCLP Ver. 3 Results*		References Describing Error Pattern	Expert Assessment of Error Patterns
		Number of Times Identified	Number of Children		
Voicing Errors	Voiced sounds for voiceless sounds			Chapman, 1993	
	a. stops	1	1		a. less than 10%
	b. fricatives	9	6		b. less than 10%
	c. affricates	1	1		c. less than 10%
	Voiceless sounds for voiced sounds			Bzoch, 1965; Scherer, Williams, & Proctor-Williams, 2008	
	a. stops	28	12		a. less than 10%
	b. fricatives	15	10		b. less than 10%
	c. affricates	2	2		c. less than 10%
Sibilant Errors	Palatal fricative for alveolar fricatives	15	7	Albery & Grunwell, 1993; Eurocleft Speech Group, 1993; Morris & Ozanne, 2003	more than 10%
	Labiodental fricatives for alveolar fricatives	4	3	Chapman, 1993	less than 10%
	Addition of a nasal following a sibilant	4	1	Harding & Grunwell, 1996	less than 10%
	Fronting	14	6	Chapman, 1993; Harding & Grunwell, 1996; Morris & Ozanne, 2003	more than 10%
Cluster Errors	Deletion of an obstruent from an obstruent-obstruent cluster	2	2	Chapman, 1993	more than 10%
	Deletion of an obstruent from an obstruent-sonorant cluster	6	4	Morris & Ozanne, 2003; Ruiter, Korsten-Meijer & Goorhuis-Brower, 2009	less than 10%
	Backing and cluster reduction	0	0	Chapman, 1993	more than 10%

Appendix B

Forms for Content Review of *SIP-CCLP Ver. 5* Candidate Error Patterns

Description of the *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP)*

Intelligibility can be defined as the degree to which an individual's spoken message is recovered by a listener (Kent, Weismer, Kent & Rosenbek, 1989). When linguistic, morpho-syntactic, environmental and listener variables are controlled, intelligibility is considered to be an integrative measure of speech ability that reflects a talker's resonance, articulation, voice and prosody. This type of integrative measure of speech ability corresponds to the level of communicative activity (i.e., execution of a task) in the International Classification of Functioning, Disability and Health (ICF) framework (WHO, 2009).

The *Speech Intelligibility Probe for Children with Cleft Palate (SIP-CCLP)* is a computer-mediated measure of single-word intelligibility that uses a phonetic contrast approach to target the speech error patterns of children with cleft palate (Connolly, 2001; Feltz, McClure & O'Hare, 2002; Gotzke, 2005; Hodge & Gotzke, 2007). Children's word productions are elicited in response to a pre-recorded verbal model and pictorial cue and recorded directly to the computer as digital audio (.wav) files. In the most recent version of *SIP-CCLP* currently under development, listener judges complete a closed–set (i.e., multiple choice) word identification task in which they choose which of six choices best matches what was heard. These choices are four real words, a "blank" to be used to type in a response if what is heard does not match one of the provided choices and "can't identify" to be used if what is heard is not recognizable as an English sound. After choosing a response, listeners also rate the child's production of the target sound as "clear" or "distorted." Results obtained include an intelligibility score (i.e., percentage of words identified correctly by listeners), a phonetic accuracy score (i.e., a measure that reflects both identification and distortion judgments by listeners) and a list of the speech error patterns identified by listeners that may be contributing to the child's speech intelligibility deficit.

Items in the closed-set task were developed using a phonetic contrast approach to intelligibility assessment. In this approach, pairs of words that differ in one or two features (i.e., manner, place, voicing) serve as stimuli in a multiple-choice listening task. Each word pair targets error patterns of interest for children with cleft palate. Six different types of error patterns found in the speech of children with cleft palate are targeted in *SIP-CCLP*: manner preference errors (MP), place preference errors (PP), sibilant errors (SE), voicing errors (VE) and cluster errors (CE). These were identified from a comprehensive review of the literature on the speech error patterns of children with cleft palate and examination of error patterns identified by listeners using an earlier version of S*IP-CCLP* for a group of 15 children with cleft palate (Gotzke, 2005). Forty error

patterns that may be identified in the speech of children with cleft palate were identified. While some of these error patterns may be more unique to children with cleft palate (e.g., substitution of velar stops for alveolar stops), other patterns may be identified in the speech of children with typical speech, language and craniofacial development. Each of these error patterns are targeted in the most recent version of *SIP-CCLP*.

Content Review Form

Instructions: Considering the population of children with cleft palate who also have a speech disorder, please indicate if the error pattern described would be identified in fewer than 10% of these children.
Circle your response beside the error pattern and add comments to support your response as you wish.

Error Pattern	Occurs in fewer than 10%		Comments
Substitution of glides for obstruents e.g., fell → well	Yes	No	
Substitution of liquids for obstruents e.g., sail → rail	Yes	No	
Substitution of nasals for obstruents e.g., sail → nail	Yes	No	
Substitution of oral stops for nasals e.g., mat → bat	Yes	No	
Substitution of oral stops for liquids and glides e.g., whale → bale	Yes	No	
Substitution of oral fricatives for liquids and glides e.g., rail → veil	Yes	No	
Substitution of nasals for liquids e.g., lap → nap	Yes	No	
Substitution of affricates for stops e.g., two → chew	Yes	No	
Substitution of affricates for fricatives e.g., sue → chew	Yes	No	

Error Pattern	Occurs in fewer than 10%		Comments
Stopping e.g., zee → D	Yes	No	
Substitution of fricatives for affricates e.g., chew → sue	Yes	No	
Gliding e.g., row → whoa	Yes	No	
Substitution of velars for obstruents e.g., D → key	Yes	No	
Substitution of bilabials for alveolars e.g., tail → pail	Yes	No	
Substitution of glottal stops for oral stops e.g., toe → O	Yes	No	
Substitution of alveolar stops for bilabial stops e.g., pea → tea	Yes	No	
Substitution of alveolar fricatives for labiodental and interdental fricatives e.g., think → sink fail → sail	Yes	No	
Substitution of glottal stops for fricatives e.g., bash → baa	Yes	No	
Substitution of glottal stops for affricates e.g., G → E	Yes	No	
Substitution of glottal stops for sonorants e.g., Lee → E	Yes	No	

Error Pattern	Occurs in fewer than 10%		Comments
Substitution of glottal fricatives for oral stops e.g., pail → hail	Yes	No	
Substitution of glottal fricatives for oral fricatives e.g., ship → hip	Yes	No	
Substitution of glottal fricatives for affricates e.g., jail → hail	Yes	No	
Substitution of glottal fricatives for sonorants e.g., Lee → he	Yes	No	
Fronting e.g., gown → down	Yes	No	
Palatalization e.g., sip → ship	Yes	No	
Weakening e.g., sail → fail	Yes	No	
Addition of nasal following a fricative e.g., sail → snail	Yes	No	
Fronting e.g., ship → sip sink → think	Yes	No	
Substitution of voiceless stop for voiced in initial and medial position e.g., bale → pail	Yes	No	
Substitution of voiceless fricative for voiced in initial and medial position e.g., veil → fail	Yes	No	

Error Pattern	Occurs in fewer than 10%		Comments
Substitution of voiceless affricate for voiced in initial and medial position e.g., jeep → cheep	Yes	No	
Substitution of voiced stop for voiceless e.g., tear → deer	Yes	No	
Substitution of voiced fricative for voiceless e.g., fail → veil	Yes	No	
Substitution of voiced affricate for voiceless e.g., cheep → jeep	Yes	No	
Substitution of voiceless obstruents for voiced in final position e.g. robe → rope	Yes	No	
Deletion of an obstruent in an obstruent-obstruent cluster e.g., stew → two	Yes	No	
Backing (and cluster reduction) e.g., stew → coo	Yes	No	
Deletion of obstruent in an obstruent-sonorant cluster e.g., block → lock	Yes	No	
Deletion of a sonorant in an obstruent-sonorant cluster e.g., trip → tip	Yes	No	

In the space below, please provide descriptions and examples of error patterns that should be added to *SIP-CCLP* to provide adequate representation of the speech error patterns of children with cleft palate who have a speech disorder.

Rater Background and Experience

Please respond to the following questions.

1. What is your current or most recent academic and/or clinical position? Please describe below. For each, please indicate the length of time you have held that position in years.

Academic not applicable ☐

_____

_____

___0 – 5 years ___15 - 20 years ___more than 30 years
___5 – 10 years ___20 – 25 years
___10 – 15 years ___25 – 30 years

Clinical not applicable ☐

_____

_____

___0 – 5 years ___15 - 20 years ___more than 30 years
___5 – 10 years ___20 – 25 years
___10 – 15 years ___25 – 30 years

2. Please indicate how many years of experience you have had analyzing the speech of children with cleft palate. This experience may include, but is not limited to, phonetic transcription, analysis of speech error patterns (e.g., phonological assessment), and rating the speech of children with cleft palate.

___0 – 5 years ___15 - 20 years ___more than 30 years
___5 – 10 years ___20 – 25 years
___10 – 15 years ___25 – 30 years

3. Do you consider yourself to have expert knowledge about the speech characteristics of children with cleft palate?

☐ Yes ☐ No

Please comment below if you chose "No."

_____

Appendix C

SIP-CCLP Ver. 5 Stimulus Words

Sound Class	Target sound	Stimulus word	
		Form 1	Form 2
Stops	p	pea	pie
		pat	pad
		rope (final)	lap (final)
	b	bee	buy
		bat	beat
		robe* (final)	robe* (final)
	t	tea	two
		toe	tie
		tap	tip
		white (final)	light (final)
	d	dough	dye
		deer*	deer*
		dot	date
		bead (final)	seed (final)
	k	cow	K
		cap	cab
		leak (final)	lake (final)
	g	go	guy
		goat	gate
		log (final)	wig (final)

Note. *Same stimulus word is in both forms.

Sound Class	Target sound	Stimulus word	
		Form 1	Form 2
Fricatives	f	fail*	fail*
		fat	feet
		wife (final)	leaf (final)
	v	V*	V*
		veil*	veil*
	θ	thick	think
	ð	not targeted	not targeted
	s	see	Sue
		sap	sip
		sick	sink
		sell	sail
		pass (final)*	pass (final)*
	z	zee*	zee*
		zoo*	zoo*
		zip	zap
		As (final)	Ks (final)
	ʃ	she	shy
		shop	ship
		bash (final)	wish(final)
	ʒ	not targeted	not targeted
	h	not targeted	not targeted
Affricates	tʃ	chew*	chew*
		cheese*	cheese*
		peach (final)	beach (final)
	ʤ	G	J
		jeep*	jeep*
		badge (final)*	badge (final)*

Note. *Same stimulus word is in both forms.

Sound Class	Target sound	Stimulus word	
		Form 1	Form 2
Glides	w	wheel	well
	j	year	yell
Liquids	l	Lee lock	low lip
	r	rail write	row read
Nasals	m	mat	meat
	n	knee fan (final)	no ran (final)
	ŋ	long (final)	rang (final)
Consonant Clusters	st	stay stick	stew steak
	sl	slip	slap
	sn	snow	snail
	str	straight	strip
	sp	spell	spear
	sk	ski	sky
	tr	trail	trip
	dr	drip	dry

Appendix D

SIP-CCLP Ver. 5 Item Analysis

Introduction

Item analysis is used to identify the final set of items that will be included in a test. In conducting item analysis, the test developer calculates and evaluates item parameters that describe how a sample of examinees responds to each item on the test (Crocker & Algina, 1986). How item parameters are used to select the final items for a test depends on its purpose (Kirshner & Guyatt, 1985). If the purpose of a test is discriminative, items that discriminate between examinees with different degrees of ability on the construct would be retained, while others that do not discriminate would be deleted. If the purpose of the test is evaluative, items that are responsive to change in the underlying construct would be retained, while nonresponsive items would be deleted. Both classical test and item response theory may be used to develop item parameters.

In classical item analysis, item difficulty and item discrimination are often used. Item difficulty is determined by calculating the proportion of examinees who answer a dichotomously scored question correctly (Crocker & Algina, 1986). Item discrimination is used to identify items that are likely to be answered correctly by examinees with a high score on the measure and to be answered incorrectly by examinees with a low score on the measure. Item discrimination is usually considered to be more important than item difficulty when selecting items; however, this relationship may change depending on the purpose of the measure. Item parameters are group dependent (Gierl, 2008). As a result, each

time the sample of examinees changes, item parameters need to be recalculated. While no minimum sample size for calculating item parameters has been determined, Crocker and Algina (1986) suggest that item parameters developed from 200 examinees are relatively stable. Furthermore, they recommend that when developing item parameters, the sample size of examinees be five to ten times larger than the number of items in a test.

When conducting item analysis using item response theory (IRT), different item parameters are estimated or held constant depending on the model chosen (Hambleton, Swaminathan, & Rogers, 1991). In the one-parameter model (1PL), it is assumed that there is no guessing and all items have equal discrimination; difficulty is estimated. In the two-parameter model (2PL), it is assumed that there is no guessing and both discrimination and difficulty are estimated for each item. In the three-parameter model (3PL), all three parameters are estimated (i.e., guessing, discrimination, and difficulty). With IRT, the test developer is responsible for choosing the model and evaluating model-data fit. Item parameters are considered to be independent of the sample of examinees. To calculate stable item parameters using IRT, large sample sizes are recommended (i.e., 500 for 1PL model, 1000 for 2PL model and 1500 for 3PL model; Gierl, 2008).

The purpose of this study was to conduct a preliminary item analysis of *SIP-CCLP Ver. 5* using classical test theory. Item parameters were then used to identify items to be exchanged on the two forms to improve form equivalence and excluded in the calculation of *SIP-CCLP Ver. 5* scores. To identify the former,

guidelines for item reduction for discriminative health status measures described by Kirshner and Guyatt (1985) were followed. According to Kirshner and Guyatt (1985), each item in a discriminative measure should distinguish between examinees according to their functional status on the construct. Therefore, items with very high or very low difficulty, items with negative discrimination, and items which are affected by factors other than the construct being measured by the test should be deleted.

Method

Children and listeners. Listener responses for the children with cleft palate (n = 20) described in Chapter 3 were used.

Checking a listener's response file. Prior to analyzing the three listeners' responses, the analysis component of the *SIP-CCLP Ver. 5* software checked each listener's response file for entries in the "other/blank" response option. As described in Chapter 2, these entries were presented, item by item, above the target and foil words, for the test administrator to review and either verify or recode as the "target" or "foil." The test administrator examined each entry typed in by the listeners to see if it contained the sound in the contrastive position in the target or foil words. After checking, the software compiled and analyzed the three listener's responses. Phonetic contrast items in which a minimum of two of the three listeners chose the target were coded as "1." Phonetic contrast items in which no agreement was obtained among the three listeners, two of the three listeners chose "can't identify," two of the three

listeners chose the same foil or two of the three listeners typed a response in the "blank" button were coded as "0."

Data analysis. A difficulty index (p-value) was determined by dividing the number of children for whom listeners correctly identified the target by the total number of children (n = 20). Easier items (i.e., less difficult) have a higher proportion than harder items (i.e., more difficult). Item discrimination was examined by calculating a discriminative index. To calculate the discriminative index (*D* index), children were first divided into two equal groups according to the rank order of their intelligibility score on each form (see Appendix H). The proportion of children in the "lower" group for whom the item was identified correctly was subtracted from the proportion of children in the "upper" group for whom the item was identified correctly. Items with high discrimination indices better discriminate children with high intelligibility scores from children with low intelligibility scores.

Results

Item difficulty. Item difficulty ranged from 0.1 to 0.95 (median = 0.65, mean = 0.64, SD = 0.18) for form 1 and from 0.1 to 1.0 (mean/median = 0.7, SD = 0.20) for form 2. The distributions of the difficulty indices for form 1 and 2 are presented in Figure D-1. One item in each form had a difficulty index less than 0.2 (F1: "As"; F2: "Ks"). Three items in form 1 had a difficulty index greater than 0.9 (i.e., fan, pat, and rope). Eleven items in form 2 had a difficulty index greater than 0.9 (i.e., cab, lake, lap, light, meat, no, pad, pie, rang, well, yell).

Item discrimination. For form 1, intelligibility scores ranged from 28.0 to 66.7 (mean = 49.7, SD = 14.3) for the ten children in the "lower" group and from 67.2 to 90.5 (mean = 77.9, SD = 8.5) for the ten children in the "upper" group. For form 2, intelligibility scores ranged from 32.3 to 69.8 (mean = 55.9, SD = 13.6) for the ten children in the "lower" group and from 70.9 to 95.8 (mean = 81.5, SD = 8.2) for the ten children in the "upper" group. The discrimination index ranged from -0.2 to 0.7 (median = 0.3) for both forms (F1: mean = 0.3, SD = 0.21; F2: mean = 0.26, SD = 0.21). The distributions of the discrimination indices for form 1 and 2 are presented in Figure D-2. Three items in form 1 (i.e., bee, log, long) and four items in form 2 (i.e., lap, low, rang, yell) had negative discriminative indices. Nine items in form 1 had a positive discriminative index less than 0.2 (i.e., As, fan, leak, mat, pat, rail, rope, snow, toe). Nineteen items in form 2 had a positive discriminative index less than 0.2 (i.e., beat, cab, lake, light, lip, K, meat, no, pad, pie, ran, read, row, sail, seed, strip, think, veil, well).

Form equivalence. To improve the equivalence of the two forms with respect to item difficulty, p-values for stimulus words which targeted the same consonant on both forms were compared. Twelve stimulus word pairs (i.e., F1: year; F2: yell) where the p-value for the form 1 word was lower than the form 2 word and the difference in p-values was greater than 0.15 were identified. From this set of eleven words, six stimulus words in form 1 with lower p-values (i.e., tea (p-value = 0.45), sap (p-value = 0.50), sell (p-value = 0.5), rail (p-value = 0.35), year (p-value = 0.55), knee (p-value = 0.70)) were exchanged with six stimulus words in form 2 with higher p-values (i.e., two (p-value = 0.85), sip (p-

value = 0.75), sail (p-value = 0.75), row (p-value = 0.6), ye11 (p-value = 0.95), no (p-value = 1.0)).

For the revised forms, item difficulty ranged from 0.1 to 1.0 (median = 0.65, mean = 0.67, SD = 0.19) for form 1 and from 0.1 to 1.0 (median = 0.65, mean = 0.67, SD = 0.20) for form 2. The distributions of the difficulty indices for revised form 1 and 2 are presented in Figure D-3. Five items in form 1 had a difficulty index greater than 0.9. Nine items in revised form 2 had a difficulty index greater than 0.9. The mean intelligibility score was 66.11% (SD = 17.66; range: 30.16 – 91.00) for revised form 1 compared to 63.81% (SD = 18.48; range: 28.04 – 90.48) for the original form 1. The mean intelligibility score was 66.40% (SD = 17.77; range: 30.16 – 95.24) for revised form 2 compared to 68.70% (SD = 17.11; range: 32.28 – 95.77) for the original form 2. The discrimination index ranged from -0.2 to 0.7 (median = 0.3) for both revised forms (F1: mean = 0.3, SD = 0.21; F2: mean = 0.27, SD = 0.22).

Item reduction and revision. There were eight items on revised form 1 and ten items on revised form 2 with a difficulty index greater than 0.85, indicating that the stimulus word was identified correctly for a minimum of 18 children. Item parameters for these stimulus words and the equivalent item in the other form are listed in Table D-1 (form 1) and Table D-2 (form 2). For seven stimulus word pairs (form 1: pat, rope, fan, pea, white, toe, mat; form 2: pad, lap, ran, pie, light, meat), the difference in item difficulty between the items targeting the same sound in both forms equal to or less than 0.1. These seven stimulus word pairs target /p/ in initial and final position (3 items), /m/ in initial position,

/n/ in final position and /t/ in initial and final position (2 items). These items do not discriminate among children in either form and were flagged for deletion from the next version of *SIP-CCLP*. The difference in difficulty indices for the remaining two items in form 1 (i.e., no, yell) and four items in form 2 (i.e., lake, well, cab, rang) with high p-values was greater than 0.1. Examination of the discriminative indices for these items revealed that the *D* index was less than 0.2 in both forms for three of the stimulus word pairs These items target /k/ in final position (i.e., leak-lake) and /ŋ/ in final position (i.e., long-rang) and /j/ in initial position (i.e., yell-year). As these items do not discriminate among children in this preliminary analysis, they were flagged for deletion from the next version of *SIP-CCLP*. For the remaining items in form 1 (i.e., no) and form 2 (i.e., well, cab) with high p-values, item parameters for the stimulus words targeting the same sound in the opposite form are within acceptable limits, suggesting that the items in the two forms are not the same in some way. These three items were flagged for revision.

Item parameters and lexical characteristics for the three items flagged for revision are listed in Table D-3. For all three word pairs, word frequency per million words (Brysbaert & New, 2009) was higher for the stimulus word with very high difficulty. In the next version of *SIP-CCLP*, these stimulus words will be replaced with words with lower word frequency (i.e., "no" in revised form 1 will be replaced with "new," "well" in form 2 will be replaced with "whale" and "cab" in form 2 will be replaced with "cape"). The lexical characteristics of these three words are provided in Table D-3. Following deletion of these ten items,

intelligibility scores and discriminative indices were recalculated. One item in form 1 (i.e., row) and three items in form 2 (i.e., log, bee, lee) had negative discriminative indices. Item parameters for these stimulus words and the equivalent item in the other form are listed in Table D-4. As the item in both forms had a discriminative index less than 0.2, these four items were flagged for deletion from the next version of *SIP-CCLP (Ver. 6)*. Figure D-4 outlines the steps that were followed to reduce the number of items to 49.

Intelligibility scores and item parameters of the final forms were recalculated and internal consistency of each form was assessed using Cronbach's alpha. Intelligibility scores ranged from 15.3% to 93.1% (mean = 61.4, SD = 15.3) on final form 1 and from 17.7% to 95.0% (mean = 60.8, SD = 22.0) on final form 2. Item difficulty ranged from 0.1 to 0.85 (median = 0.63, mean = 0.62, SD = 0.16) for final form 1 (48 items[6]) and from 0.1 to 0.85 (median = 0.65, mean = 0.60, SD = 0.16) for final form 2 (47 items). The discrimination index ranged from 0 to 0.8 (mean = 0.37, SD = 0.15) for final form 1 and from 0 to 0.7 (mean = 0.34, SD = 0.18) for final form 2. The median discrimination index was 0.4 for both forms. Cronbach's alpha was 0.94 and 0.93 for form 1 and 2, respectively, for the 20 child participants.

[6] Although each of the final forms contains 49 items, scores were calculated using data for 48 items for final form 1 because one item was revised and 47 items for final form 2 because two items were revised.

Discussion

The purposes of this study were to evaluate the quality of the *SIP-CCLP Ver. 5* items by calculating indices of difficulty and discrimination, and then, to use this information to improve form equivalence and to reduce the number of items in the next version of *SIP-CCLP*. Form 1 and 2 were similar with respect to overall difficulty (mean = 0.64 (Form 1); 0.7 (Form 2)) and discrimination (mean = 0.3 (Form 1); 0.27 (Form 2)). Furthermore, both forms had items that sampled across the range of difficulty and discrimination.

Examination of the distribution of the difficulty indices revealed that form 2 had more easy items (i.e., 6 items with p-value = 1.0) and fewer items with p-values ranging between 0.4 and 0.55 than form 1 (i.e., F1: 21; F2: 12). To improve the equivalence of the two forms with respect to difficulty, six stimulus words to be exchanged between the two forms were identified.

To reduce the number of items in the next version of *SIP-CCLP*, guidelines for item reduction for discriminative health status measures outlined by Kirshner and Guyatt (1985) were followed. Kirshner and Guyatt (1985) recommend that items "to which most or all of the respondents give similar or identical answers" and "idiosyncratic items that in which patients who by other criteria have a low functional status perform well and vice versa" (p. 31) be excluded from discriminative measures. Ten items with high difficulty indices and four items with negative discriminative indices were flagged for deletion from the next version of *SIP-CCLP*. These items target stops in initial position (four items), stops in final position (four items) and sonorants (six items).

Kirshner and Guyatt (1985) also recommend that items that are affected by factors other than the construct being measured by the test be deleted. The likelihood that a word may be present in a young child's lexicon (i.e., age of acquisition) was one factor that was identified in development of *Ver. 5* that may affect children's accuracy of production. Seven stimulus words in *SIP-CCLP Ver. 5* were not listed in frequency usage counts by Stemach and Williams (1988) and Kolson (1960), but were primed prior to test administration in the "learn word" software feature described in Chapter 2. None of these seven words (i.e., bash, fail, lee, spear, sue, veil, zap) had low difficulty indices, suggesting that the "learn word" software feature was effective at minimizing the impact of age of acquisition on children's production accuracy. The number of other words which are similar to the target word in child's vocabulary (i.e., neighbourhood density) was another factor that was identified in development of *SIP-CCLP Ver. 5* as having a possible effect on children's accuracy of production. Production accuracy may be lower for words from low density neighbourhoods (Sosa & Stoel-Gammon, 2012). There were 16 words in form 1 and 15 words in form 2 with fewer than ten neighbours. None of these words had a difficulty index less than 0.3. Comparison of the difficulty index of the words with fewer than ten neighbours with the same number of words with more than 20 neighbours revealed that the average difficulty index for the words with fewer neighbours (F1: 0.59, SD = 0.18; F2: 0.6, SD = 0.13) was lower than the average difficulty index for the words with more than 20 neighbours (F1: 0.69, SD = 0.18; F2: 0.81, SD = 0.18). No stimulus words were flagged for deletion because of the effects

of age of acquisition or phonological neighbourhood. However, word frequency did appear to be a factor affecting the difficulty of two of the three items flagged for revision. In these two items, word frequency, as reported in the SUBTLEXus database (Brysbaert & New, 2009), of the item with the higher difficulty index was substantially higher than the item with acceptable difficulty in the opposite form (i.e., "no" (Form 1): p-value = 1.0, word frequency = 5971.55, "knee" (Form 2): p-value = 0.7, word frequency = 14.69). New stimulus words with lower word frequency were chosen to replace these stimulus items.

The results of this item analysis identified 14 items for deletion. As this analysis is based on results from a small number of children (n = 20), it is it is recommended that these items still be recorded in the next evaluation of *SIP-CCLP Ver. 5.* In a classical test theory model, item parameters are sample dependent; therefore, item parameters calculated with other samples of children may yield different results. Crocker and Algina (1986) suggest that relatively stable item parameters may be obtained using a minimum sample of 200 examinees. It is recommended that this item analysis be replicated with data from a larger sample of children before changes are made to *SIP-CCLP Ver. 5*. Furthermore, it is recommended that these items be maintained to adequately sample the speech sounds of English-speaking children and to decrease the respondent burden of children, parents, and listeners. Respondent burden is defined as "the time, effort and other demands place on those to whom the instrument is administered" (p. 202; Scientific Advisory Committee of the Medical Outcomes Trust, 2002). Including items that the majority of children are

expected to produce intelligibly will reduce the potential that children will feel discouraged about their ability to imitate the *SIP-CCLP* words and parents will be discouraged by their child's performance on the measure. Including these items also ensures that listeners are able to identify at least some of the words, thereby reducing listener stress. Characteristics of the 63 stimulus words in *SIP-CCLP Ver. 5* and *Ver. 6* are listed in Table D-5.

This item analysis was conducted following a classical test theory model. Item response theory offers an alternative means of evaluating items and constructing new forms that maximize the precision of ability estimates. However, large sample sizes (i.e., minimum 500 examinees) are recommended to calculate stable item parameters using item response theory (Gierl, 2008). The use of item response theory to evaluate items and maximize the precision of *SIP-CCLP* scores could be examined in future, large-scale studies.

Table D-1

SIP-CCLP Ver. 5 Revised Form 1 Items with Item Difficulty Greater than 0.9 and the Equivalent Items in Revised Form 2

Revised Form	Stimulus Word	P-value	*D* index	Revised Form	Stimulus Word	P-value	*D* index	Action
1	no	1.0	0.00	2	knee	0.7	0.60	Revise
1	pat	0.95	0.10	2	pad	0.95	0.10	Delete
1	rope	0.95	0.10	2	lap	0.95	-0.10	Delete
1	fan	0.95	0.10	2	ran	0.9	0.00	Delete
1	yell	0.95	0.10	2	year	0.55	-0.10	Delete
1	pea	0.9	0.20	2	pie	1.0	0.00	Delete
1	toe	0.9	0.00	2	tie	0.80	0.20	Delete
1	white	0.9	0.20	2	light	1.0	0.00	Delete

Table D-2

SIP-CCLP Ver. 5 Revised Form 2 Items with Item Difficulty Greater than 0.9 and the Equivalent Items in Revised Form 1

Revised Form	Stimulus Word	P-value	*D* index	Revised Form	Stimulus Word	P-value	*D* index	Action
2	lake	1.0	0.00	1	leak	0.8	0.00	Delete
2	meat	1.0	0.00	1	mat	0.9	0.00	Delete
2	well	1.0	0.00	1	wheel	0.85	0.30	Revise
2	cab	0.95	0.10	1	cap	0.8	0.40	Revise
2	rang	0.95	-0.10	1	long	0.8	-0.20	Delete

Note. pad, lap, ran, pie, and light in revised form 2 also had p-values greater than 0.9 and are reported with their form 1 counterparts in Table D-1.

Table D-3

SIP-CCLP Ver. 5 Stimulus Word Pairs to be Revised

Revised Form	Stimulus Word	P-value	*D* Index	Number of Neighbours[1]		Word Frequency[2] (per million words)
				CCC	IPhOD	
1	no	1.0	0.00	22	42	5971.55
2	knee	0.7	0.60	29	39	14.69
NEW 1	new			19	36	723.78
2	cab	0.95	0.10	12	30	35.8
1	cap	0.8	0.40	19	42	18.75
NEW 2	cape			16	29	8.24
2	well	1.0	0.00	15	40	2990.65
1	wheel	0.85	0.30	17	36	27.06
NEW 2	whale			n/a	40	11.25

Notes. Stimulus words with an unacceptable item parameter are listed first, followed by the stimulus word in the alternate form. [1]Number of neighbours was obtained from the Child Corpus Calculator (CCC) (Storkel & Hoover, 2010) and Irvine Phonotactic Online Dictionary (IPhOD) (Vaden, Hickok & Halpin, 2009). [2]Word frequency was obtained from the SUBTLEXus database (Brysbaert & New, 2009). n/a = not available.

Table D-4

SIP-CCLP Ver. 5 Items in Revised Form 1 and 2 with a Negative Item Discriminative Index and the Equivalent Items in the Opposite Revised Form

Revised Form	Stimulus Word	P-value	*D* Index	Revised Form	Stimulus Word	P-value	*D* Index
1	row	0.6	-0.20	2	rail	0.35	0.10
2	log	0.4	-0.20	1	wig	0.55	0.10
2	bee	0.75	-0.10	1	buy	0.85	0.10
2	lee	0.6	-0.20	1	low	0.8	0

Table D-5

Characteristics of SIP-CCLP Ver. 5 and Ver. 6 Stimulus Words

Characteristic		*SIP-CCLP Ver. 5*		*SIP-CCLP Ver. 6*	
		Form 1	Form 2	Form 1	Form 2
Number of Words Containing a High Vowel		27	29	23	26
Word Frequency[1] (per million words)	Mean	221.50	259.03	228.37	119.02
	SD	734.26	902.69	737.04	372.53
	Median	19.92	27.48	21.67	21.48
	Minimum	1.12	1.14	1.12	1.14
	Maximum	3793.04	5971.55	3793.04	2691.39
Number of Neighbours[2]	Mean	33.10	31.89	33.24	31.63
	SD	11.15	10.53	11.22	10.38
	Median	36	33	36	33
	Minimum	10	8	10	8
	Maximum	56	50	56	48

Notes. [1] Word frequency was obtained from the SUBTLEXus database (Brysbaert & New, 2009). [2]Number of neighbours was obtained from the Irvine Phonotactic Online Dictionary (IPhOD) (Vaden, Hickok & Halpin, 2009).

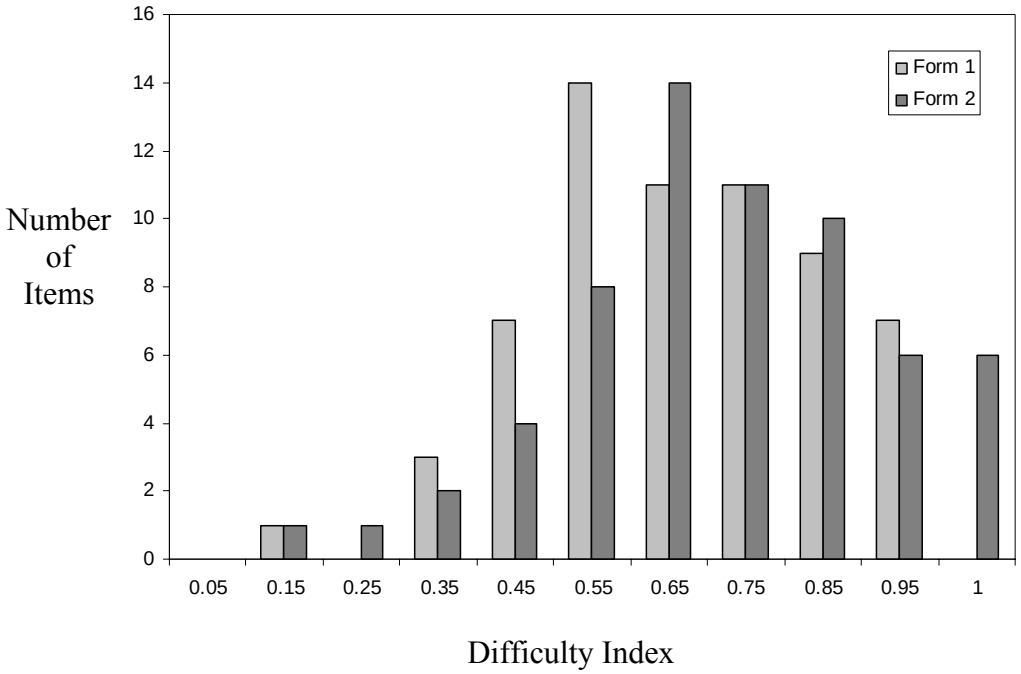

Figure D-1. Item difficulty distributions of *SIP-CCLP Ver. 5* form 1 and 2.

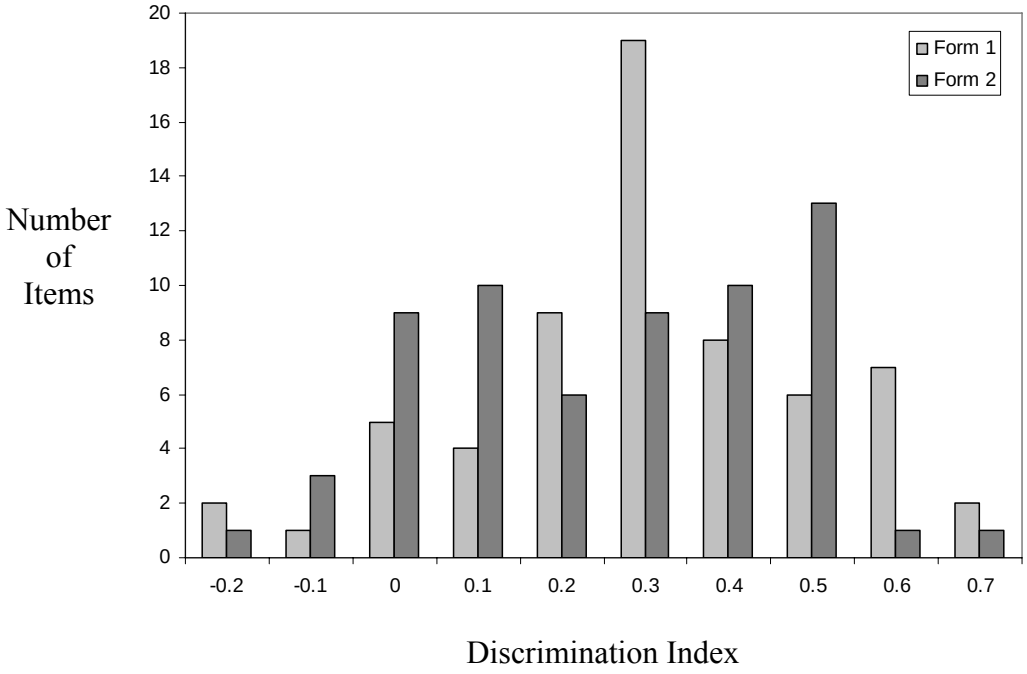

Figure D-2. Item discrimination distributions of *SIP-CCLP Ver. 5* form 1 and 2

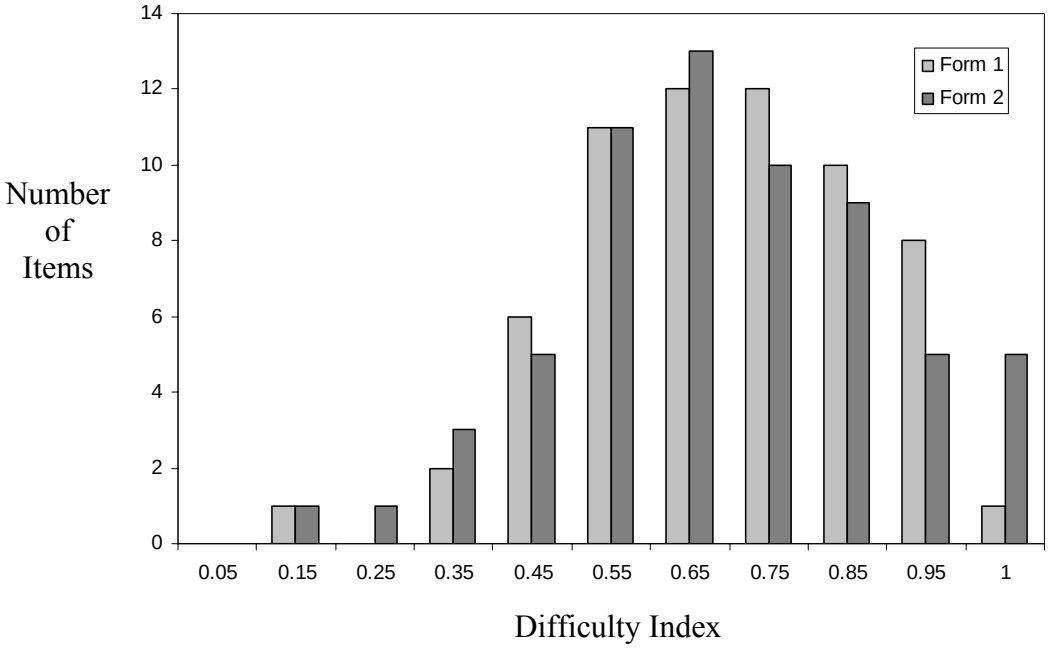

Figure D-3. Item difficulty distributions of revised form 1 and 2

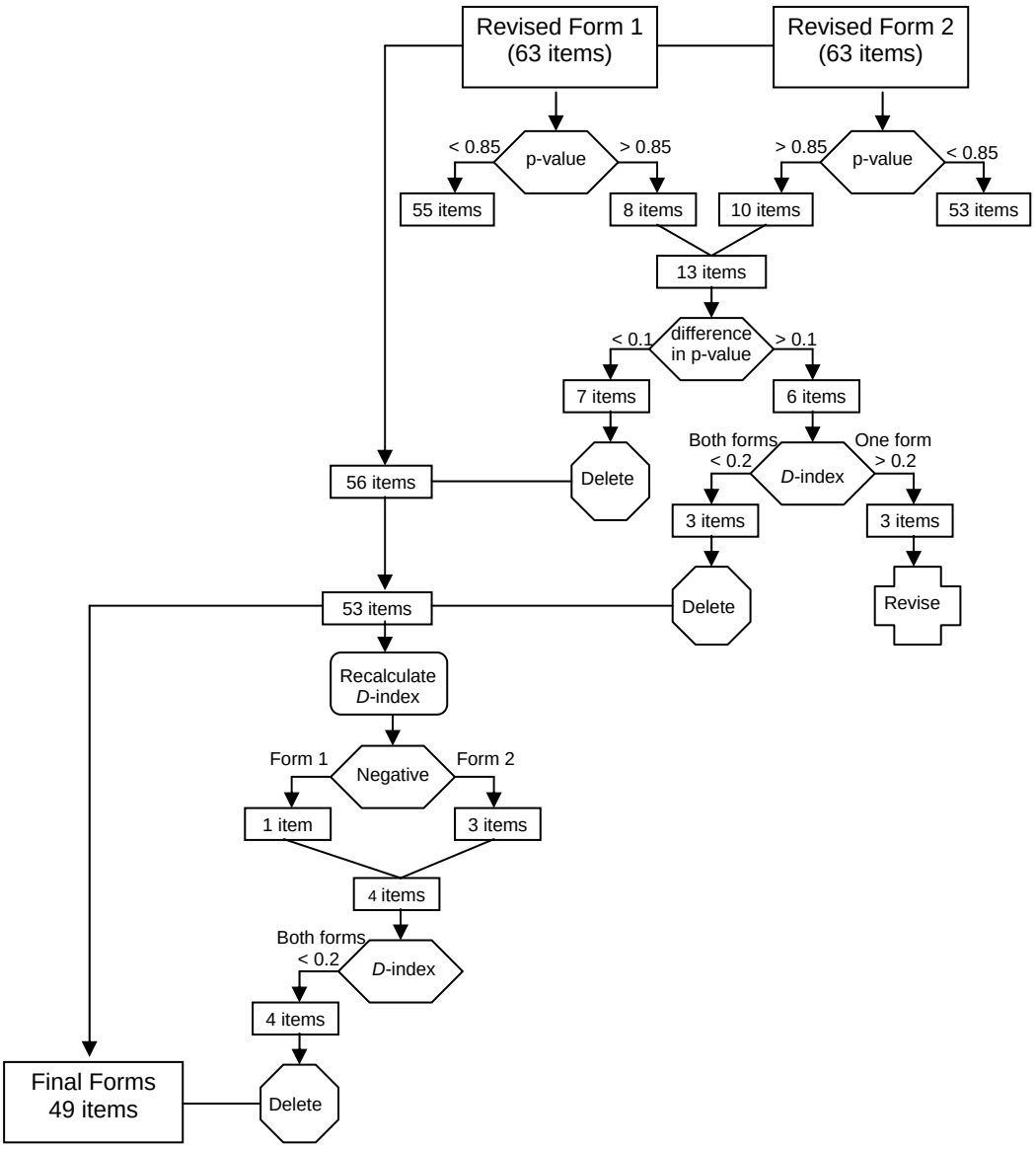

Figure D-4. Flow chart outlining the process of identifying potential items for deletion and revision for revised *SIP-CCLP Ver. 5* form 1 and 2

References

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41(4)*, 977-990.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Gierl, M. (2008). Unpublished class notes for Educational Psychology 508: Item Response Theory. University of Alberta, Edmonton, AB.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Disorders, 38*(1), 27-36.

Kolson, C. J. (1960). *The Vocabulary of Kindergarten Children*. (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality Life Research, 11,* 193-205.

Sosa, A.V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech, Language and Hearing Research, 55*, 596-608.

Stemach, G., & Williams, W. B. (1988). *Word express: The first 2, 500 words of spoken English illustrated*. Novato, CA: Academic Therapy Publications.

Storkel, H. L., & Hoover, J. R. (2010). An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods, 42(2),* 497-506.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). *Irvine Phonotactic Online Dictionary, Version 2.0*. [Data file]. Retrieved from http://www.iphod.com/

Appendix E

Evaluation of the Dependability of *SIP-CCLP Ver. 5* Scores Using Generalizability Theory

Introduction

In the classical true score model, each equation for estimating reliability of a measure identifies and quantifies a single source of measurement error (e.g., different occasions, test forms and combinations of items; Streiner & Norman, 2008). Generalizability (*G*) theory was developed as an extension to classical test theory to provide a way to simultaneously evaluate multiple sources of measurement error (Streiner & Norman, 2008). Error variances resulting from identified sources of measurement error and the interactions amongst them are used to provide a single overall estimate of dependability (reliability in the classical model) that describes the "accuracy of generalizing from a person's observed score to the average score that the individual would receive under all the possible conditions that the test user would be willing to accept" (i.e., the universe of generalization; Shavelson & Webb, 1991, p. 1).

In *G* theory, sources of measurement error are referred to as facets. Examples of possible facets include forms, raters and testing occasions. Facets are either crossed or nested in a *G* study. A facet is crossed if all objects of measurement (e.g., subjects) are observed with all conditions of the facet (Shavelson & Webb, 1991). A facet is nested if all objects of measurement are not observed with all conditions of the facet. For example, if each subject completes both forms of the test, "forms" is a crossed facet, but if each subject

completes only one form, "forms" is a nested facet. Whether facets are crossed or nested affects the types of questions that might be asked in a *G* study about interactions between facets (Streiner & Norman, 2008). Facets can also be fixed or random. If a facet is described as fixed, the test developer intends to generalize the results of the *G* study only to those conditions described in the study (Crocker & Algina, 1986). In achievement tests, subtest is often considered a fixed facet, as there are no untested conditions (e.g., subject areas) in the universe of generalization (Shavelson & Webb, 1991). If a facet is described as random, the test developer intends to generalize the results to all conditions, not just the conditions in the study. For example, if "items" is described as a random facet, the test user considers the set of X number of items on the test to be exchangeable with any other set of X number of test items from the universe of items (Shavelson & Webb, 1991).

Once the facets are described, the test developer conducts a *G* study. The purpose of this study is "to anticipate multiple uses of a measurement and to provide as much information as possible about possible sources of variation in the measurement" (Shavelson & Webb, 1991, p. 12). A *G* study is considered to be a planning study associated with development of the procedure for calculating a measure's dependability (Crocker & Algina, 1986; Shavelson & Webb, 1991). A decision (*D*) study is then conducted to design a procedure for using a measure that minimizes error (and maximizes dependability). Prior to conducting a *D* study, the test developer must define the interpretation of a measure as either relative or absolute (Shavelson & Webb, 1991). Relative decisions are

interpretations based on the ranking of subjects as in a norm-referenced framework, while absolute decisions are based on the absolute score as in a criterion-reference framework. The type of interpretation affects the definitions of error and how dependability coefficients are calculated. In a *D* study, once the variance components have been calculated, an optimal level of generalizability can be determined by increasing the number of levels of each facet(e.g., number of items) over which repeated measures are obtained and averaged to obtain a dependable score (Marcoulides, 1999).

Classical test theory was used to evaluate the test-retest and alternate forms reliability of *SIP-CCLP Ver. 5* (see Chapter 3). Pearson's correlation and intraclass correlation coefficients were greater than 0.9 and standard error of measurement was less than 5% of the total score (range = 2.95 – 4.56). Using limits of agreement (Bland & Altman, 1986), it was found that form 2 scores tended to be higher than form 1 scores. Gotzke concluded that the results of this evaluation provided support for *SIP-CCLP Ver. 5* as a reliable measure of speech intelligibility for young children with cleft palate, but suggested using *G* theory to evaluate form, occasion and raters as sources of measurement error. The purpose of this study was to use *G* theory to examine the dependability of *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores by estimating the error variances associated with four facets (i.e., order, occasion, form and rating[7]) and then, to conduct *D* studies to develop guidelines for the minimum number of ratings that

[7] In nested designs with raters nested within persons (i.e., *r:p x f)*, the same set of listeners judges a sub-sample of the examinees. This was not the case in this study, as groups of three independent listeners (i.e., raters) were assigned randomly to each child. To resolve this *G* study design problem, researchers have treated rating (r') as a random facet instead of rater (e.g., Huang, 2007). This strategy allows rating to be treated as a fully crossed facet.

must be averaged to ensure the dependability of *SIP-CCLP Ver. 5* scores for relative decisions.

Method

Data. *SIP-CCLP Ver. 5* form 1 and 2 recordings from 14 children with cleft palate were collected at session one and two as described in Chapter 3. Each child was randomly assigned to one of two orders of form administration (order 1: form 1 – form 2 – form 2 – form 1; order 2: form 2 – form 1 – form 1 – form 2). The number of days between sessions ranged from 5 to 14 (mean = 9.0 days) for the seven children assigned to order 1 and from 7 to 21 days (mean = 12.6 days) for the seven children assigned to order 2. Listener judgments were collected using the *SIP-CCLP Ver. 5* closed-set response task as described in Chapter 3. Each child's set of recordings for each form and occasion were judged by three independent listeners. Intelligibility and phonetic accuracy scores for each listener (i.e., ratings) were calculated by the *SIP-CCLP Ver. 5* software (see Chapter 3).

Data Analysis. *G* and *D*-study analyses were conducted using the *GENOVA for PC* program (Crick & Brennan, 2003). In *G*-study 1, separate person within order-by-time-by-form-by-rating (*p:o x t x f x r'*) analyses were conducted for *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores. Order (*o*), time (*t*) and form (*f*) were fixed facets with 2 levels; rating (*r'*) was a random facet with 3 levels. In *G* study 2, separate person-by-time-by-form-rating (*p x t x f x r'*) analyses were conducted for children in each order for both *Ver. 5* scores. To evaluate the effect of different numbers of ratings on the dependability

of *SIP-CCLP* scores, *D* studies of the same design as G study 2 (*p x t x f x r'*) were conducted. Time (*t*), form (*f*) and rating (*r'*) were crossed random facets. The *D* (dependability) coefficient and relative error, which is equivalent to standard error of measurement in classical test theory, were reported for each *D* study. Relative error was used to calculate the minimal detectable change (MDC).

Results

***G* study 1.** Mean *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores for the seven children in order 1 and the seven children in order 2 are reported in Table E-1 for session one and Table E-2 for session two. Examination of these results revealed that intelligibility and phonetic scores were consistently higher for the children in order 1, suggesting that the two groups were not similar in the severity of their speech intelligibility impairment.

The results of the *G* studies for *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores are presented in Table E-3. Persons within order (*p:o*) yielded the largest percentage of the total variance for both intelligibility (77.0%) and phonetic accuracy scores (74.69%), suggesting that person within order differed greatly in these scores. The residual (*ptfr':o*) yielded the second largest estimated variance component for both intelligibility (8.49%) and phonetic accuracy scores (7.10%). The residual contains the variability due to the interaction between time, rating, form and person within order and other unexplained systematic or unsystematic sources of error. Person within order-by-form-by-rating (*pfr':o*) yielded the third largest estimated variance component for

intelligibility scores (4.79%), suggesting that the three ratings assigned to person within order differed. Order (*o*) yielded the third largest variance component for phonetic accuracy scores (5.75%), suggesting that the children in order one had somewhat different phonetic accuracy scores than the children in order two. Because a main effect of order was found for phonetic accuracy scores, *G* study 2 was conducted.

***G* study 2**. The results of the *G* studies for *SIP-CCLP Ver. 5* intelligibility scores for each order are presented in Table E-4. Person (*p*) yielded the largest estimated variance component for order 1 (81.17%) and order 2 (69.57%), suggesting that, as expected, children differed in their intelligibility scores. The residual (*ptfr'*), which contains the variability due to the interaction between person, time, form and rating and other unexplained systematic and unsystematic sources of errors, yielded the second largest variance component for both orders (1: 7.75%; 2: 10.24%). The person-by-form-by-rating interaction (*pfr'*) yielded the third largest variance component for order 1 (4.48%), suggesting that the standing of children differed from form to form and rating to rating. For order 2, the person-by-time-by-rating interaction (*ptr'*) yielded the third largest variance component (6.98%), suggesting that standing of children differed from session to session and rating to rating.

The results of the *G* studies for *SIP-CCLP Ver. 5* phonetic accuracy scores for each order are presented in Table E-5. Person (*p*) yielded the largest estimated variance component for order 1 (84.78%) and order 2 (57.61%), suggesting that, as expected, children differed in their phonetic accuracy scores. The residual

(*ptfr'*) yielded the second largest variance component for both orders (1: 5.66%; 2: 13.90%). The person-by-time-by-rating interaction yielded the third largest variance component for order 1 (3.54%) and order 2 (12.08%), suggesting that standing of children differed from session to session and rating to rating, particularly for order 2.

***D* study**. The *D* coefficients for *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores are summarized in Table E-6. For *Ver. 5* intelligibility scores, the *D* coefficients for a design in which reported scores are based on three independent ratings for one form were .97 for order 1 and .93 for order 2. For the same design, the relative errors were 3.61 and 3.88 and the MDCs were 10.00% and 10.74% for order 1 and 2, respectively. Decreasing the number of independent ratings to two resulted in a *D* coefficients of .96 and .90, relative errors of 4.41 and 4.52 and MDCs of 12.23% and 12.53% for order 1 and 2, respectively. For phonetic accuracy scores, the *D* coefficients for a design in which reported scores are based on three independent ratings for one form were .97 for order 1 and .86 for order 2. For the same design, the relative error variances were 3.96 and 4.22 and the MDCs were 10.97% and 11.70% for order 1 and 2, respectively. Decreasing the number of independent ratings to two resulted in a *D* coefficients of .96 and .83, relative errors of 4.73 and 4.69 and MDCs of 13.12% and 13.01% for order 1 and 2, respectively.

Discussion

The purpose of this study was to examine the dependability of *SIP-CCLP Ver. 5* intelligibility and phonetic accuracy scores using *G* theory. As expected, person-within-order yielded that largest estimated variance component for both *SIP-CCLP* scores for *G* study 1. Because severity of the speech intelligibility impairment appeared to differ for the children assigned to each order and a main effect of order was found for phonetic accuracy scores, *G* study 2 was conducted to determine the variance components for time, form and rating for the two orders separately. The variance components for all three facets (i.e., time, form and rating) were small, with each accounting for less 1% of the total variance. Results of these *G* studies were similar for both orders and scores. However, the residual variance component (*ptfr'*) was consistently larger for order 2 and phonetic accuracy scores. *D* studies were then conducted to determine the minimum number of ratings that should be averaged when using *SIP-CCLP Ver. 5*. Results from the *D* studies suggest that scores based on a minimum of two ratings are dependable (i.e., *D* coefficient > 0.9, relative error < 5.0). However, relative error was slightly smaller and MDC was about 2% lower for scores based on three versus two raters.

Upon recruitment, children were randomly assigned to one of two orders with no a priori knowledge about the severity of their speech disorder or speech intelligibility impairment. The differences in intelligibility and phonetic accuracy scores for the two groups of children (i.e., order 1 and order 2) suggest that the two groups were not equivalent with respect to severity of speech intelligibility

impairment. Four of the seven children assigned to order 1 had *SIP-CCLP* intelligibility scores greater than 80%, while none of the seven children assigned to order 2 had intelligibility scores greater than 80%. This group difference necessitated that dependability of scores be examined separately for the two orders (*G* study 2). In future examinations of the reliability/dependability of *SIP-CCLP*, it is recommended that an intelligibility rating (e.g., *Intelligibility in Context Scale*; McLeod, Harrison & McCormack, 2012) be completed by parents prior to assignment. This rating could be used to ensure that the range of speech intelligibility impairment is similar in the two groups.

Although person yielded that largest estimated variance component for both *SIP-CCLP* scores for *G* study 2, the percentage of variance explained was relatively greater for the children in order 1 than for the children in order 2 for both intelligibility (order 1: 81.2%; order 2: 69.57%) and phonetic accuracy scores (order 1: 84.78%; order 2: 57.61%). This result is likely related to the restricted range in scores of the children assigned to order 2 compared to the children assigned to order 1 (Tables E-1 and E-2). Because of the restricted range, the variability (standard deviation) was also lower for order 2, resulting in the relatively lower estimated variance components for person for both *SIP-CCLP* scores.

In *G* study 2, the main effects of time, form and rating each accounted for less than 1% of the total variance for each order and *SIP-CCLP* score, suggesting that the score for each child was not systematically different from session one to session two, from form 1 to form 2 and from rating to rating. However, when the

variance components for the interactions of person with time (i.e., *pt, ptf, ptr'*), person with form (i.e., *pf, ptf, pfr'*) and person with rating (i.e., *pr', ptr', pfr'*) were each summed, the percentage of variance explained by each of the sums was relatively greater for the children in order 2 than for the children in order 1 for both intelligibility and phonetic accuracy scores. For example, the summed variance component for the interactions of person with time for intelligibility scores was 3.2% for order 1 and 11.4% for order 2. These results suggest that the standing of children differed somewhat from session to session, from form to form, and from rating to rating for the children in order 2 most importantly. Examination of the rank order of the children's mean intelligibility scores revealed that the only change in rank in order 1 was for the two children with the highest scores who exchanged ranks from administration of form 2 in session one to administration of form 2 in session two. However, rank changed at least once for all children in order 2. For mean phonetic accuracy scores, none of the seven children in order 1 changed ranks over sessions or forms, while five of the seven children in order 2 changed rank at least once. Each child's mean intelligibility and phonetic accuracy scores over each form and session is displayed graphically in Figures E-1 and E-3 for the children in order 1 in Figures E-2 and E-4 for the children in order 2. The restricted range of scores for the children assigned to order 2 is apparent and may account, in part, for the changing ranks.

With respect to the facet of time, it is important to note that the average number of days between sessions was longer for the children assigned to order 2 (12.6 days) than order 1 (9 days). The increased amount of time between sessions

may have resulted in more variability in scores for the children in order 2. As described in Chapter 3, timing of the second session was affected by availability of research space and families and travel time for the test administrator. In future evaluations of the dependability of *Ver. 5*, it is recommended that time between sessions be fixed to minimize the variance associated with time. Reducing the number of days between sessions would also minimize variance associated with this facet. Roebroeck, Harlaar and Lankhorst (1993) suggest that adding another session to allow children to become familiar with the examiner and the *SIP-CCLP* may also decrease measurement error associated with time.

There are a number of strategies that could be implemented to decrease the error variance associated with forms and ratings. Basing *SIP-CCLP* scores on results from more than one form would decrease the variability associated with forms. However, this strategy would increase respondent burden for both children (e.g., longer assessment session) and examiners (e.g., administering additional listening sessions). Improving form equivalence with respect to item difficulty and discrimination would also decrease error variance associated with this facet (see Appendix D). To decrease the variability associated with ratings, one strategy is to increase the number of ratings on which *SIP-CCLP* scores are based. Using one listener to judge all children would also decrease the measurement error associated with ratings. However, further research is needed to determine the effect of multiple exposures to the *SIP-CCLP* stimulus words on scores before this strategy could be implemented.

A relatively high percentage of variance remained in the residual (*ptfr'*) in G study 2 (i.e., range: 5.66% - 13.90%). The residual contains the variability due to the interaction between person, time, form and rating and other unexplained systematic and unsystematic sources of errors. Sources of measurement error that were not investigated in this study include the *SIP-CCLP* items and the familiarity of the examiner with the child subjects.

D studies were conducted to examine the effect of decreasing the number of ratings on the dependability of *SIP-CCLP*. For order 1, the *D* (dependability) coefficients were greater than 0.9 regardless of whether intelligibility and phonetic accuracy scores were based on single scores or the mean score over two or three ratings. For order 2, the *D* coefficients were greater than 0.9 when two or three ratings were used to calculate intelligibility scores but were never greater than 0.9 for phonetic accuracy scores. These results suggest that *SIP-CCLP* is dependable even when only two ratings are used to calculate intelligibility scores. Furthermore, the relative error variances are relatively small (i.e., < 5%). The MDC suggests that a difference greater than 12.5% in intelligibility scores is needed to be confident that the difference is not consistent with the measurement error of the test. Based on these results, it is recommended that *SIP-CCLP Ver. 5* scores be based on the mean score of a minimum of two ratings.

Results of this application of generalizability theory support the dependability of *SIP-CCLP Ver. 5* as a discriminative measure of speech intelligibility for young children with cleft palate. The results of this analysis should be considered preliminary as variance components in *G* study 2 were

calculated using data from only seven children. It is recommended that this study be replicated with data from a larger sample of children.

Table E-1

Mean SIP-CCLP Ver. 5 Intelligibility and Phonetic Accuracy Scores for the Seven Children in Each Order in Session One

Order 1	Form 1			Form 2		
	Mean	SD	Range	Mean	SD	Range
Intelligibility Score (%)	69.92	23.34	28.04 – 90.48	74.76	21.83	32.27 – 95.77
Phonetic Accuracy Score (%)	60.32	24.28	20.11 – 85.45	63.42	23.08	22.75 – 87.03

Order 2	Form 2			Form 1		
	Mean	SD	Range	Mean	SD	Range
Intelligibility Score (%)	62.43	13.98	41.80 – 78.83	60.70	15.27	36.51 - 75.66
Phonetic Accuracy Score (%)	51.13	10.17	35.72 – 59.79	50.38	13.34	30.95 – 65.35

Note. SD = standard deviation.

Table E-2

Mean SIP-CCLP Ver. 5 Intelligibility and Phonetic Accuracy Scores for the Seven Children in Each Order in Session Two

Order 1	Form 2			Form 1		
	Mean	SD	Range	Mean	SD	Range
Intelligibility Score (%)	74.75	18.60	43.92 – 94.71	71.05	22.67	29.63 – 91.01
Phonetic Accuracy Score (%)	64.51	21.09	33.60 – 89.68	64.25	24.30	22.75 – 87.83

Order 2	Form 1			Form 2		
	Mean	SD	Range	Mean	SD	Range
Intelligibility Score (%)	64.17	14.44	42.33 – 77.78	63.34	13.79	42.86 – 75.66
Phonetic Accuracy Score (%)	51.17	10.40	36.51 – 60.32	51.06	11.72	35.71 – 65.87

Note. SD = standard deviation.

Table E-3

Variance Components for Mixed Effects p:o x t x f x r' G-study Design (N_{order} = 2, Ntime = 2, N_{form} = 2, $N_{ratings}$ = 3)

		Intelligibility (%)		Phonetic Accuracy (%)	
Source of Variability	Degrees of Freedom	Variance Component	Percentage of Total Variance	Variance Component	Percentage of Total Variance
order (o)	1	0.80	0.18	25.59	5.75
person (p):o	12	339.50	77.00	332.43	74.69
time (t)	1	0.42	0.10	0.40	0.09
form (f)	1	0	0	0	0
rating (r')	2	0	0	0	0
ot	1	0	0	0.21	0.05
of	1	0	0	0	0
or'	2	0	0	0.27	0.06
pt:o	12	1.09	0.25	0	0
pf:o	12	0	0	8.13	1.83
pr':o	24	6.06	1.37	6.80	1.53
tf	1	2.45	0.56	0	0
tr'	2	0	0	0	0
fr'	2	0	0	0	0

Table E-3 continued

		Intelligibility (%)		Phonetic Accuracy (%)	
Source of Variability	Degrees of Freedom	Variance Component	Percentage of Total Variance	Variance Component	Percentage of Total Variance
otf	1	8.77	1.99	0.002	<0.01
otr'	2	0	0	0	0
ofr'	2	0	0	0	0
ptf:o	12	4.22	0.96	3.48	0.78
ptr':o	24	19.05	4.32	22.92	5.15
pfr':o	24	21.13	4.79	13.26	2.98
tfr'	2	0	0	0	0
otfr'	2	0	0	0	0
ptfr':o	24	37.42	8.49	31.60	7.10
Total	167	440.91	100	445.08	100

Table E-4

Variance Components for G-study 2 (Mixed Effects p x t x f x r' Design) for SIP-CCLP Ver. 5 Intelligibility Scores (N_{person}=7, N_{time}=2, N_{form}=2, N_{rating} =3)

		Order 1		Order 2	
Source of Variability	Degrees of Freedom	Variance Component	Percentage of Total Variance	Variance Component	Percentage of Total Variance
person (p)	6	486.18	81.17	192.82	69.57
time (t)	1	0	0	0.83	0.30
form (f)	1	0	0	0.09	0.03
rating (r')	2	0	0	0	0
pt	6	0	0	4.52	1.63
pf	6	0	0	0	0
pr'	12	4.19	0.70	7.93	2.86
tf	1	15.93	2.66	0	0
tr'	2	0	0	0	0
fr	2	0	0	0	0
ptf	6	0.63	0.11	7.82	2.82
ptr'	12	18.75	3.13	19.35	6.98
pfr'	12	26.86	4.48	15.40	5.55
tfr	2	0	0	0	0
ptfr'	12	46.44	7.75	28.39	10.24
Total	83	599.00	100	277.16	100

Table E-5

Variance Components for G-study 2 (Mixed Effects p x t x f x r' Design) for SIP-CCLP Ver. 5 Phonetic Accuracy Scores (N_{person}=7, N_{time}=2, N_{form}= 2, $N_{ratings}$=3)

		Order 1		Order 2	
Source of Variability	Degrees of Freedom	Variance Component	Percentage of Total Variance	Variance Component	Percentage of Total Variance
person (p)	6	557.09	84.78	107.76	57.61
time (t)	1	2.05	0.31	0	0
form (f)	1	0.01	<0.01	0	0
rating (r')	2	0	0	0	0
pt	6	0	0	3.02	1.62
pf	6	0.49	0.08	15.77	8.43
pr'	12	9.76	1.48	3.85	2.06
tf	1	0	0	0	0
tr'	2	0	0	0	0
fr	2	0	0	0.83	0.45
ptf	6	7.95	1.21	0	0
ptr'	12	23.24	3.54	22.59	12.08
pfr'	12	19.29	2.94	7.22	3.86
tfr	2	0	0	0	0
ptfr'	12	37.19	5.66	26.00	13.90
Total	83	657.08	100	187.06	100

Table E-6

D Coefficient, Relative Error and Minimal Detectable Change for Mixed Effects p x t x f x r D-study Design (N_{time} = 1,N_{form} = 1)

		Order 1			Order 2		
	$N_{ratings}$	*D* Coefficient	Relative Error	MDC (%)	*D* Coefficient	Relative Error	MDC
Intelligibility Score (%)	1	.93	6.23	17.26	.84	6.05	16.77
	2	.96	4.41	12.23	.90	4.52	12.53
	3	.97	3.61	10.00	.93	3.88	10.74
Phonetic Accuracy Score (%)	1	.93	6.52	18.08	.76	5.89	16.32
	2	.96	4.73	13.12	.83	4.69	13.01
	3	.97	3.96	10.97	.86	4.22	11.70

Note. MDC = minimal detectable change.

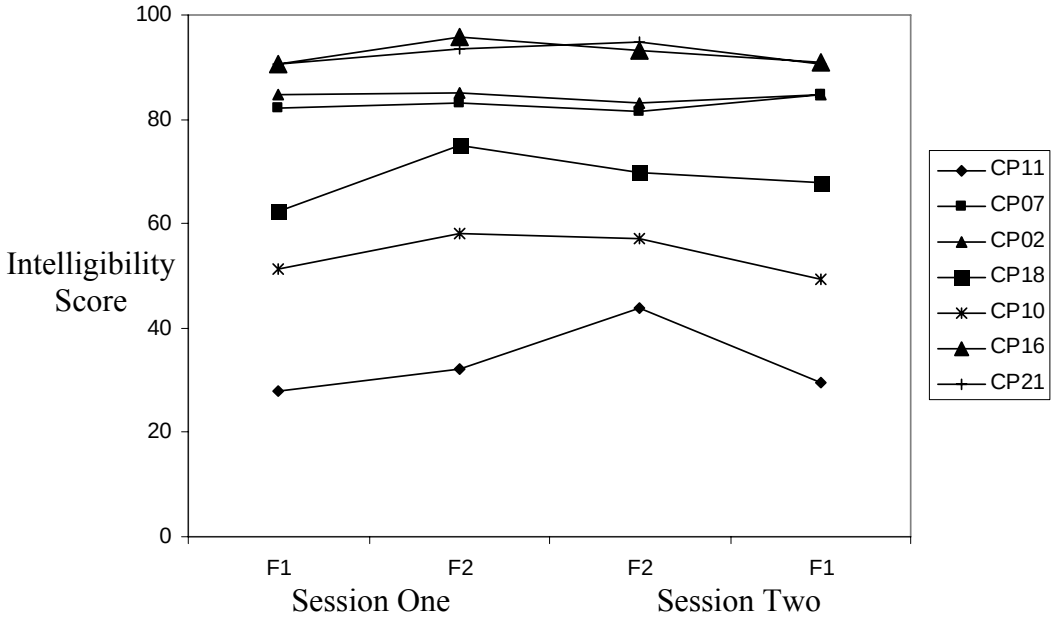

Figure E-1. Mean intelligibility scores for each session and form for the children in order 1 showing the stability/variability in the rankings.

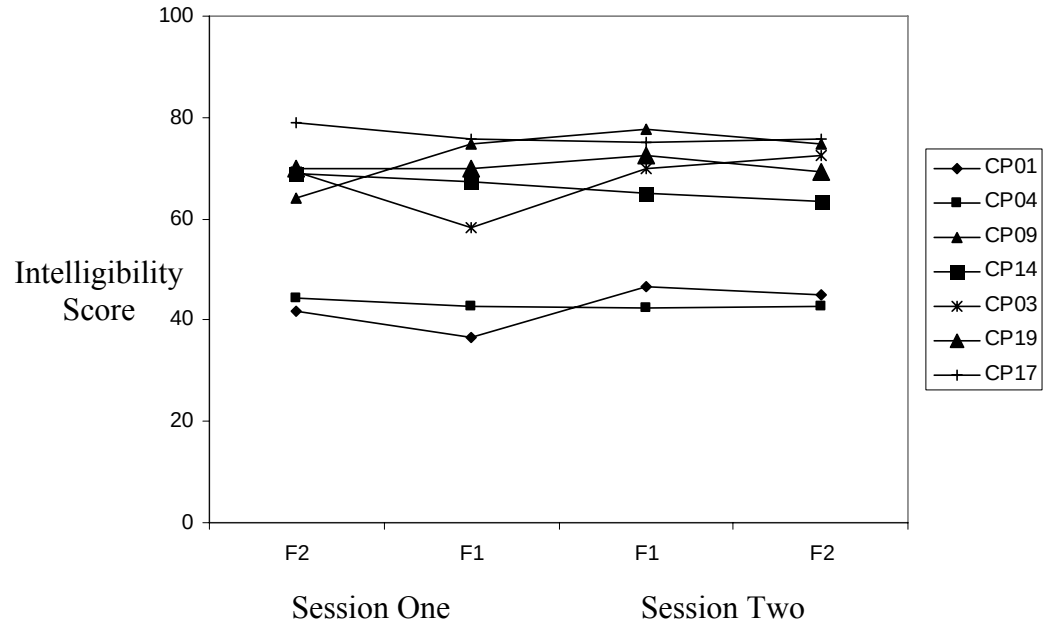

Figure E-2. Mean intelligibility scores for each session and form for the children in order 2 showing the stability/variability in the rankings.

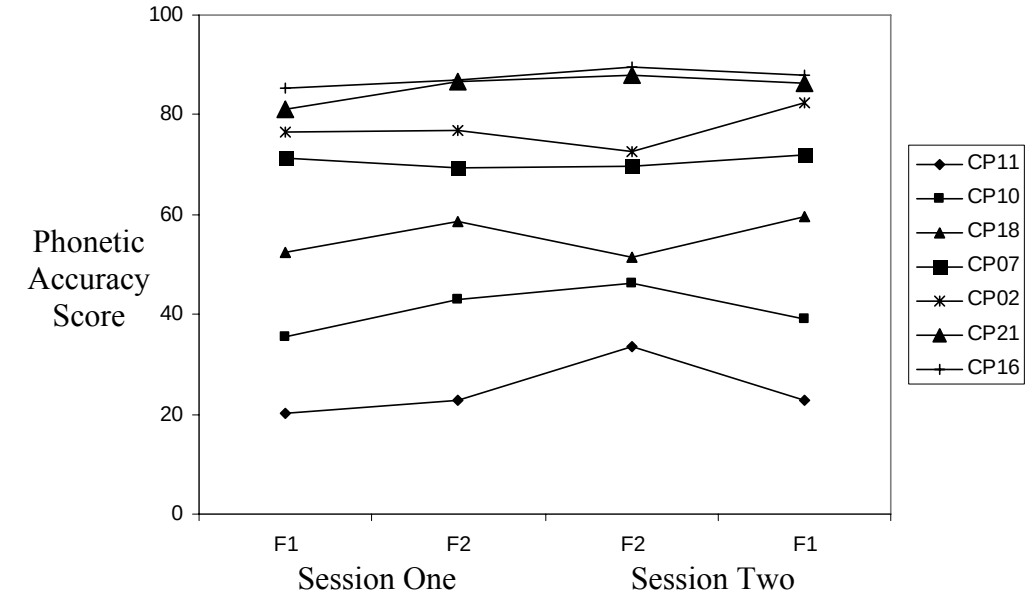

Figure E-3. Mean phonetic accuracy scores for each session and form for the children in order 1 showing the stability/variability in the rankings.

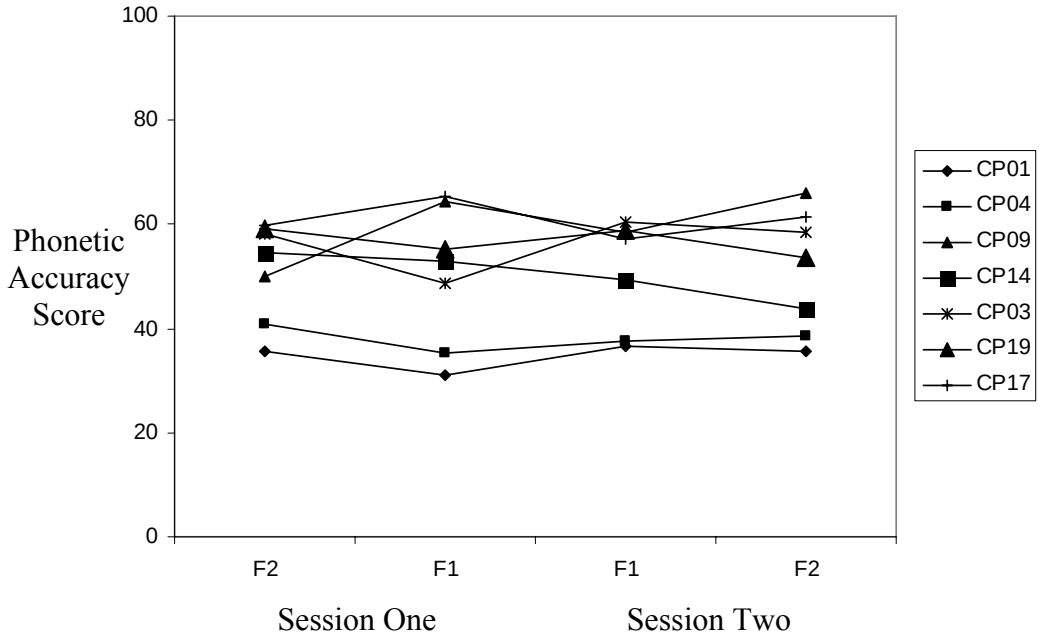

Figure E-4. Mean phonetic accuracy scores for each session and form for the children in order 2 showing the stability/variability in the rankings.

References

Bland, J. M., & Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, i,* 207 -310.

Crick, J. E., & Brennan, R. L. (2003). *GENOVA for PC* [computer software]. Downloaded from http://www.education.uiowa.edu/centers/casma/computer-programs.aspx#genova

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Huang, J. (2007). *Examining the Fairness of Rating ESL Students' Writing on Large-Scale Assessments*. (Unpublished doctoral dissertation). Queen's University, Kingston, ON.

Marcoulides, G.A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S.E. Embretson & S.L. Hershberger (Ed.), *The new rules of measurement: What every psychologist should know* (p. 129-152). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Roebroeck, M.E., Harlaar, J., & Lankhorst, G.J. (1993). The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical Therapy, 73(6)*, 386-395.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications, Inc.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: a practical guide to their development and use*. Oxford, UK: Oxford University Press.

Appendix F

Evaluation of the Reliability of *SIP-CCLP Ver. 5* Using Item Response Theory

Introduction

Item response theory (IRT) was developed as an alternative to classical test theory as a means to construct tests and interpret test scores. In IRT, it is assumed that "the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits or abilities" (Hambleton, Swaminathan, & Rogers, 1991, p. 7). Examinees with higher ability on the trait(s) underlying performance are expected to have a greater likelihood of obtaining the correct response on an item than examinees with lower ability.

With IRT, test developers must first choose which of three models to use. In these models, different item parameters are estimated or held constant (Hambleton et al., 1991). In the one-parameter model (1PL), it is assumed that there is no guessing and all items have equal discrimination; difficulty is estimated. In the two-parameter model (2PL), it is assumed that there is no guessing and both discrimination and difficulty are estimated for each item. In the three-parameter model (3PL), all three parameters are estimated (i.e., guessing, discrimination, and difficulty). Regardless of the model chosen, the test developer must evaluate model-data fit.

With classical test theory, the value of reliability estimates (e.g., correlation coefficients) depends upon the group of examinees from which they are obtained (Hambleton et al., 1991). As a result, reliability estimates calculated

using one group of examinees may be different than estimates obtained for another group. With item response theory, item characteristic curves and test information functions, which replace the concept of reliability (Gierl, 2008), are considered to be independent of the sample of examinees. Item characteristic curves (ICCs) describe the relationship between the examinees' performance on the item and the ability underlying performance on the item. Using ICCs, test developers create test information functions that display the amount of information a measure provides along the ability continuum. Large sample sizes are recommended to calculate stable item parameters using IRT, (i.e., 500 for 1PL model, 1000 for 2PL model and 1500 for 3PL model; Gierl, 2008).

In classical test theory, standard error of measurement (SEM) is calculated to evaluate the precision of test scores. A limitation of classical test theory is that the SEM is the same for examinees of all abilities (e.g., low and high scorers on a test). The standard error of the ability estimate in IRT is analogous to SEM but, unlike SEM, the value of the standard error of the ability estimate varies with ability level (Hambleton et al., 1991). Smaller values of the standard error of the ability estimate are desired across the ability scale.

A third limitation of classical test theory is that the means and standard deviations of scores obtained from two forms of a test must be equal to evaluate form equivalence (Hambleton et al., 1991). In an IRT framework, relative efficiency is used to compare the precision of two forms that measure the same ability but are not necessarily strictly parallel. To calculate relative efficiency, values describing test information along the ability continuum for one form are

divided by the test information for the second form. Values greater or less than one indicate that one form is functioning better than the other form. For example, if the relative efficiency of form A compared to form B is equal to 1.25, it indicates that form A provides more information at that ability level and is functioning as if it were 25% longer than form B (i.e., form A is functioning better than form B).

The reliability of *SIP-CCLP Ver. 5* has been examined using classical test and generalizability theory (see Chapter 3 and Appendix E). The purpose of this study was to use IRT to examine the relative efficiency of *SIP-CCLP Ver. 5* form 1 and 2. A 2PL model was selected for two reasons: 1) item discrimination indices were not equal for all items on each *SIP-CCLP* form (see Appendix D); and 2) it was assumed that there was no guessing, as scores were based on the combined responses of all three listeners. Because of the exploratory nature of this study and the small sample size on which it was based, model-data fit was not evaluated.

Method

Data. *SIP-CCLP Ver. 5* form 1 and 2 recordings from 20 children with cleft palate were collected and judged by listeners as described in Chapter 3. The *SIP-CCLP Ver. 5* analysis software was used to collate and analyze responses from listeners. Items in which a minimum of two of the three listeners chose the target were assigned a score of "1" by the software. All other items were assigned a score of "0."

Data Analysis. The *BILOG-MG for Windows (Version 3.0)* program (Zimowski, Muraki, Mislevy, & Bock, 2003) was used to calculate item parameters (i.e., a – discrimination, b – difficulty) for the two-parameter (2PL) model for each form. Discrimination (a) can range from 0 (i.e., non-discriminating) to 3 (i.e., very highly discriminating), while difficulty (b) can range from $-\infty$ (i.e., very easy) to $+\infty$ (i.e., very hard) but practically ranges from -3 to 3 (Gierl, 2008). Item parameters were used to calculate how much information each item provided at seven points along the ability continuum (i.e., -3, -2, -1, 0, 1, 2, 3) for each form. Test information at given ability levels for each form was calculated by summing item information at the aforementioned points along the ability continuum. The standard error of ability estimate was calculated using the formula SE = 1/(√test information). The minimal detectable change (MDC) was calculated using the formula MDC = 1.96 X SE X √2 (Weir, 2005). To determine relative efficiency of the two forms, the test information at given ability levels for form 1 was divided by the test information at given ability levels for form 2.

Results

The *BILOG-MG* program did not calculate item parameters for three items in form 1 (i.e., leak, log, long) and form 2 (i.e., lap, rang, read) with biserial correlations less than -0.15 and for six items in form 2 (i.e., lake, light, meat, no, pie, well) with a difficulty index equal to one (i.e., item was identified correctly for all 20 child participants). For the remaining 60 items in form 1, discrimination (a) ranged from 0.462 to 1.497 (mean = 0.833, SD = 0.222) and difficulty (b)

ranged from -2.766 to 2.303 (mean = -0.644, SD = 0.978). For the remaining 54 items in form 2, discrimination (a) ranged from 0.417 to 1.332 (mean = 0.878, SD = 0.207) and difficulty (b) ranged from -3.083 to 1.649 (mean = -0.713, SD = 0.918). The test information functions for each form are shown in Figure F-1. The standard error of the ability estimates and minimal detectable change across the ability continuum are reported in Table F-1. Standard error ranged from 0.21 to 0.83 for form 1 and from 0.21 to 0.96 for form 2. The minimal detectable change ranged from 0.57 to 2.31 for form 1 and from 0.58 to 2.66 for form 2. The standard error of the ability estimates and relative efficiency of form 1 with respect to form 2 are displayed graphically in Figures F-2 and F-3, respectively. Relative efficiency ranged from 0.94 to 1.3.

Discussion

The purpose of this study was to conduct an exploratory evaluation of the relative efficiency of *SIP-CCLP Ver. 5* form 1 and 2 using item response theory for a sample of 20 children with cleft palate. The test information functions revealed that form 1 provided maximum information near the middle of the ability scale (i.e, 0), while form 2 provided maximum information closer to -1 on the ability scale, indicating that form 2 is easier than form 1. Form 1 is more precise at both ends and the middle of the ability continuum compared to form 2. Standard error of the ability estimates were less than 1 across the ability continuum.

Item parameters for three items in each form were not calculated by the Bilog-MG software (Zimowski, Muraki, Mislevy & Bock, 2003) as biserial

correlation coefficients for these items were less than -0.15. Biserial correlation coefficients are a type of item discrimination index used in classical test theory, in which the construct underlying test performance is assumed to be normally distributed (Crocker & Algina, 1986). A negative biserial correlation coefficient indicates that listeners identified the item correctly for more children with low intelligibility scores than for children with high intelligibility scores. In the classical item analysis outlined in Appendix D, five of these six items were flagged for deletion from the next version of *SIP-CCLP*. Item parameters were also not calculated for six items in form 2 with a difficulty index equal to one. Four of these items were flagged for deletion and two of these items were flagged for revision (i.e., "no" and "well") in Appendix D.

The shape of the *SIP-CCLP Ver. 5* test information functions indicates that both forms provide less information (i.e., less precise) for children at high and low ability levels. As *SIP-CCLP* will most likely be used to assess the speech intelligibility of children with low to moderate ability levels, the limited information provided at high ability levels may not be of concern. According to Hambleton et al. (1991), a "fairly flat" test information function is desired for a test that is designed to sample the range of abilities on the test construct (p. 101). One of the advantages of IRT is that it allows the test developer to construct new versions of a test that match a target test information function. To create a version of *SIP-CCLP* that more closely approximates a flat target test information function, the test developer would select items with high discrimination (a) and difficulty in the target range (e.g., -3 to 2). After each item is added to the test,

the test developer recalculates the test information function, stopping once the test information function approximates the target test information function. It is important to note that selecting items based only on item parameters does not necessarily result in a test that has content-related validity.

The test information functions peaked at different points for both forms with form 1 peaking closer to the midpoint of the ability scale than form 2, indicating that form 2 is easier than form 1. This result was also obtained using the limits of agreement method (Bland & Altman, 1986) in the evaluation of the reliability of *SIP-CCLP Ver. 5* in Chapter 3. In the classical item analysis outlined in Appendix D, items that could be exchanged between the two forms to improve equivalence were identified. The item parameters estimated using IRT could also be used for this purpose. The relative efficiency of form 1 to form 2 indicated that form 1 is more precise at both ends and the middle of the ability scale. Form 1 is functioning as if it were 20% longer than form 2 at the low end of the ability scale (i.e., -3); as if it were 33% longer at the high end of the ability scale (i.e., 3) and as if it were 12% longer at the midpoint of the ability scale (i.e., 0). Deleting items with difficulty at these points on the ability scale from form 1 would improve the equivalence of the two forms.

The standard error of the ability estimates were low for both forms (i.e., < 1.0) across the range of scaled ability scores. According to Hambleton et al. (1991), smaller standard errors are associated with tests that are longer, have highly discriminating test items in which the correct response can not be guessed, and have items with a range of difficulties that match the ability of the examinee.

The small standard errors obtained for SIP-CCLP is likely a factor of the number of items. Like standard errors of measurement in the classical test theory model, standard error of the ability estimates can be used to calculate confidence intervals around a child's ability score. In this study, standard errors were used to calculate the minimal detectable change (MDC) that is needed to determine if scores from two examinees are different. MDCs were less than one point on the ability scale for most points on the ability continuum.

One limitation of this study is that the assumptions of model-data fit were not evaluated. When using IRT, Gierl (2008) recommends that test developers evaluate three areas of model-data fit: model assumptions, model features and model predictions. With respect to model assumptions, the test developer evaluates dimensionality (i.e., is there a single factor underlying ability on the measure?), equality of the item discrimination indices, role of guessing in test performance and speededness (i.e., is the test timed?). In this study, item discrimination indices were determined to be not equal (D indices ranged from -0.2 to 0.7 for both forms; see Appendix D) suggesting that the 1PL model was not appropriate for this data. Guessing was not formally evaluated but was suggested not to be a factor, as scores were based on the combined responses of three listeners. Speededness was not a factor for either children or listeners. Dimensionality, as well as model features (i.e., item and ability invariance) and model predictions, must be examined in future examinations of *SIP-CCLP* using IRT.

This study is exploratory as it is based on data from only 20 children. However, the results indicate that form 1 provides more information at the high and low ends and the middle of the ability continuum than form 2 and that form 1 is more difficult than form 2. Therefore item parameter estimates could be used to improve form equivalence by eliminating some items with difficulty at approximately -3, 3 and 0 on the ability scale. The feasibility of conducting an IRT analysis on the revised version of *SIP-CCLP* after form equivalence has been improved will be determined by the opportunities to collect data from very large samples of children. A 2PL model is recommended, which, according to Gierl (2008) would require 1000 subjects.

Table F-1.

Standard Error of the Ability Estimates and Minimal Detectable Change for SIP-CCLP Ver. 5 Form 1 and 2 over the Ability Scale

	Form 1		Form 2	
Ability Scale	Standard Error	MDC	Standard Error	MDC
-3	0.40	1.12	0.44	1.22
-2	0.27	0.77	0.28	0.78
-1	0.22	0.60	0.21	0.58
0	0.21	0.57	0.23	0.61
1	0.29	0.82	0.30	0.84
2	0.50	1.37	0.48	1.33
3	0.83	2.31	0.96	2.66

Note. MDC = minimal detectable change. The standard error of the ability estimate and MDC are reported in the units of the ability scale underlying performance on *SIP-CCLP Ver. 5*.

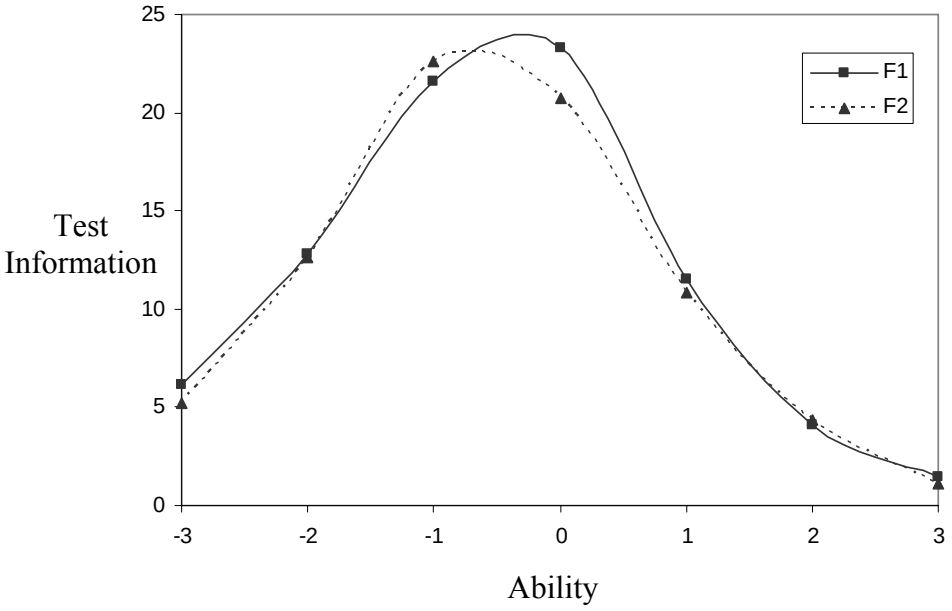

Figure F-1. Test information functions for *SIP-CCLP Ver. 5* form 1 and 2.

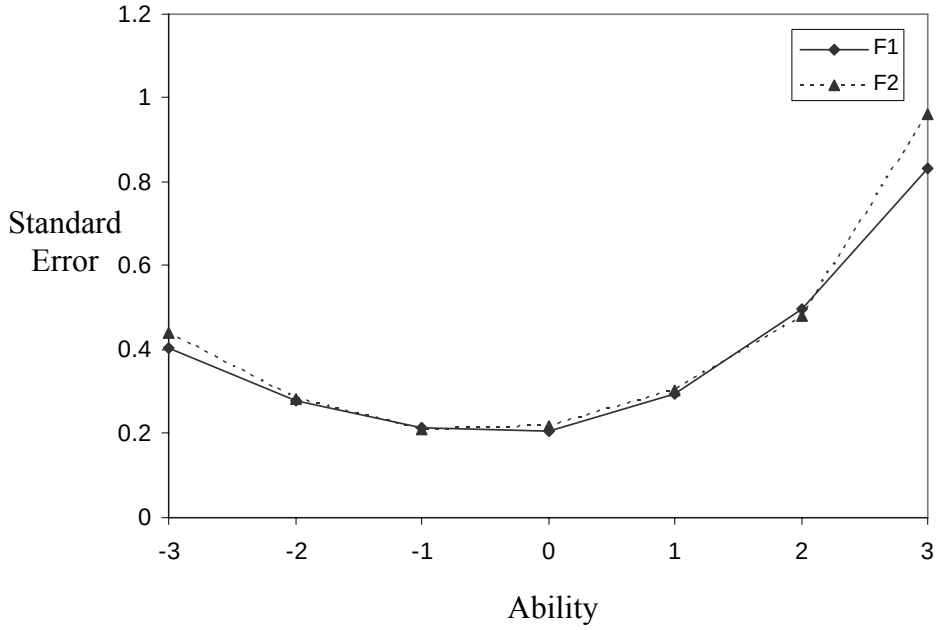

Figure F-2. Standard error of the ability estimate for *SIP-CCLP Ver. 5* form 1 and 2.

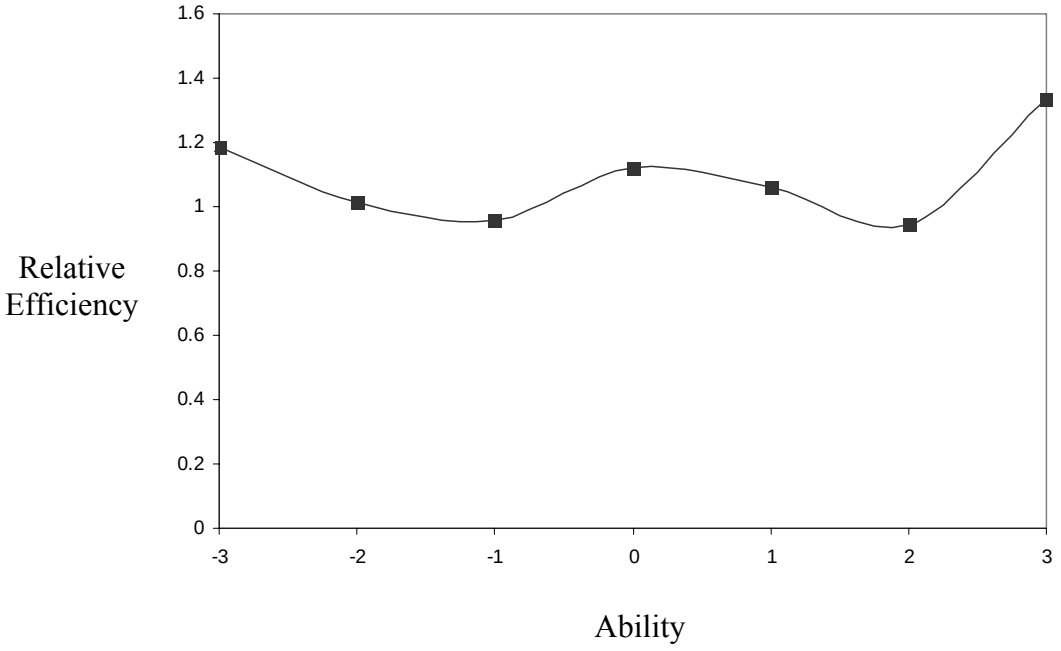

Figure F-3. Relative efficiency of Form 2 to Form 1.

References

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, i,* 207 - 310.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Gierl, M. (2008). Unpublished class notes for Education Psychological 508: Item Response Theory. University of Alberta, Edmonton, AB.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the *SEM*. *Journal of Strength and Conditioning Research, 19(1)*, 231-240.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG for Windows (Version 3.0)* [computer software]. Scientific Software International, Inc.

Appendix G

Excerpt from the *Zoo Passage* (Fletcher, 1978) Elicited from Child Participants

Look at this book with us.

It's a story about a zoo.

That is where bears go.

Today it's very cold out of doors

But we see a cloud overhead

That's a bright fluffy shape.

Appendix H

Children's Intelligibility Scores (*SIP-CCLP Ver. 5*, Spontaneous Speech Sample, *TOCS+*), *SIP-CCLP Ver. 5* Phonetic Accuracy Scores, Hypernasality Ratings, Voice Severity Ratings and Percentage of Consonants Correct Scores

	SIP-CCLP Ver. 5 Intelligibility Score		*SIP-CCLP Ver. 5* Phonetic Accuracy Score		Spontaneous Speech Intelligibility Score		*TOCS+* Sentences Intelligibility Score	Hypernasality Rating	Voice Severity Rating	PCC
	Form 1	Form 2	Form 1	Form 2	Time 1	Time 2				
CP01	36.51	41.80	30.95	35.71	52.81	64.31	45.42	200	30	45.83
CP02	84.66	85.19	76.46	76.72	80.91	70.09	90.00	40	80	80.33
CP03	58.20	69.30	48.68	58.20	46.79	52.83	50.65	150	50	47.59
CP04	42.86	44.44	35.19	40.74	67.3	76.57	48.40	100	25	39.04
CP05	63.49	74.07	53.70	65.34	33.33		54.32	225	200	40.31
CP07	82.01	83.07	71.43	69.31	76.05	72.17	78.67	75	50	60.47
CP08	69.31	82.01	57.94	71.43	69.87		84.58	180	10	64.40
CP09	74.60	64.02	64.29	50.00	81.93	73.40	67.52	200	30	51.69
CP10	51.32	58.20	35.45	42.86	55.45	60.06	60.76	280	30	36.00
CP11	28.04	32.28	20.11	22.75	61.95	44.23	30.74	300	50	38.38
CP12	30.16	46.03	23.54	32.28	30.42		29.11	260	55	27.72
CP13	75.13	76.72	64.55	69.84	71.47		80.70	175	140	50.77
CP14	67.20	68.78	52.91	54.50	40.33	43.14	59.11	90	15	39.53
CP15	57.14	64.02	43.12	51.59	53.72		67.86	200	45	48.15
CP16	90.48	95.77	85.45	87.04	81.05	78.90	91.36	40	20	82.04
CP17	75.66	78.84	65.34	59.79	83.01	71.65	90.42	120	25	45.00
CP18	62.43	75.13	52.38	58.73	85.71	72.70	86.75	200	25	63.92
CP19	69.84	69.84	55.29	58.99	57.84	52.56	77.35	180	45	45.11
CP20	66.67	70.90	61.90	61.11	63.14		73.66	70	18	41.79
CP21	90.48	93.65	80.95	86.51	86.27	88.14	97.53	90	15	59.99

Appendix I

Graphs of the Relationships between *SIP-CCLP Ver. 5* Scores and Session One Spontaneous Sample Intelligibility Scores

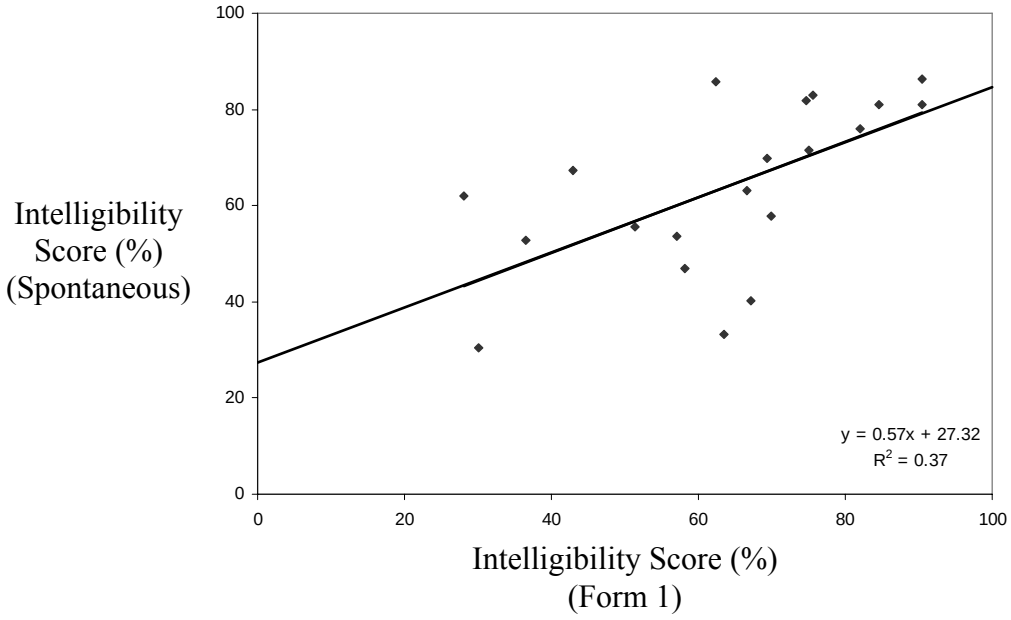

Figure I-1. Relationship between form 1 and session one spontaneous sample intelligibility scores.

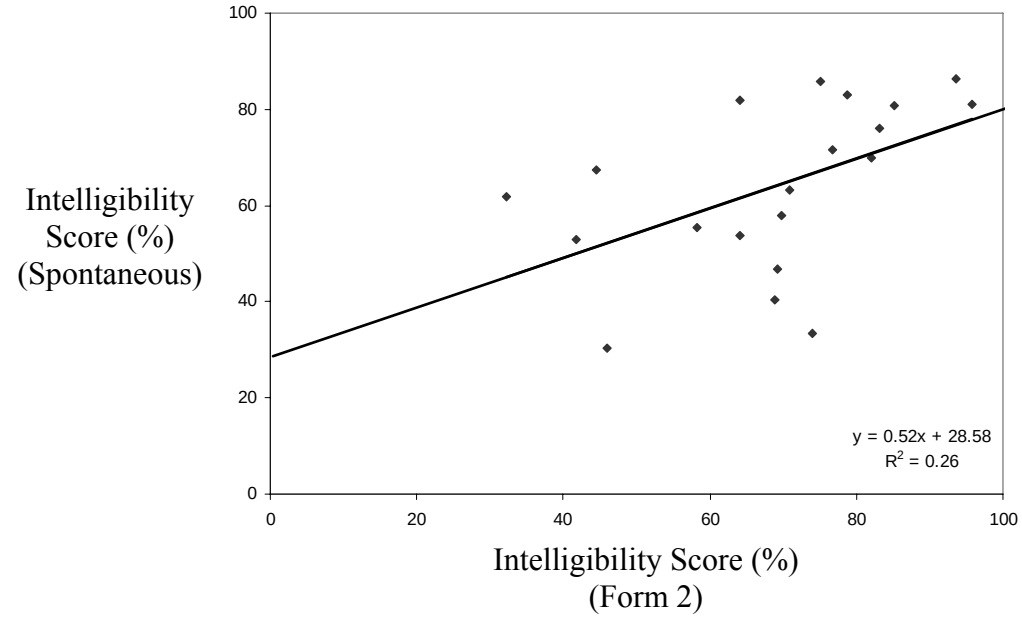

*Figure I-2***.** Relationship between form 2 and session one spontaneous sample intelligibility scores.

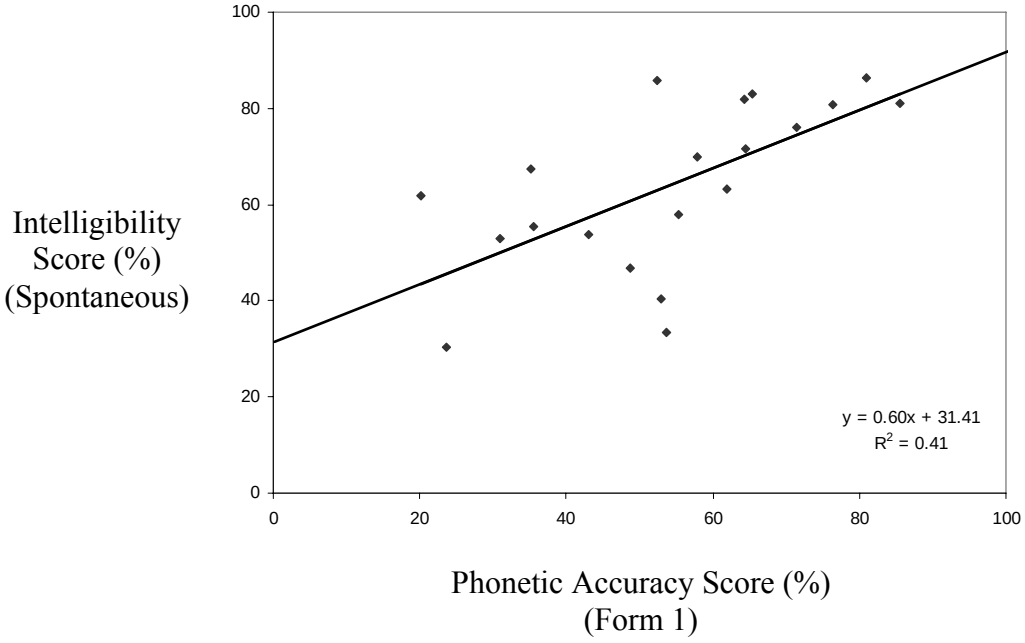

*Figure I-3***.** Relationship between form 1 phonetic accuracy scores and session one spontaneous sample intelligibility scores.

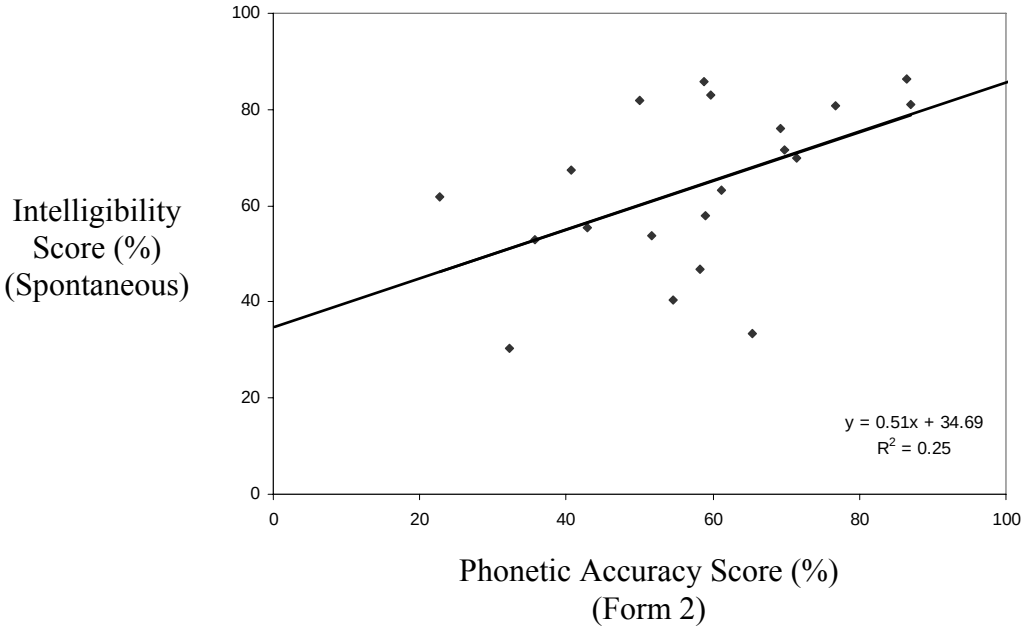

Figure I-4. Relationship between form 2 phonetic accuracy scores and session one spontaneous sample intelligibility scores.

Appendix J

Graphs of the Relationships between *SIP-CCLP Ver. 5* Scores and *TOCS+* Intelligibility Scores

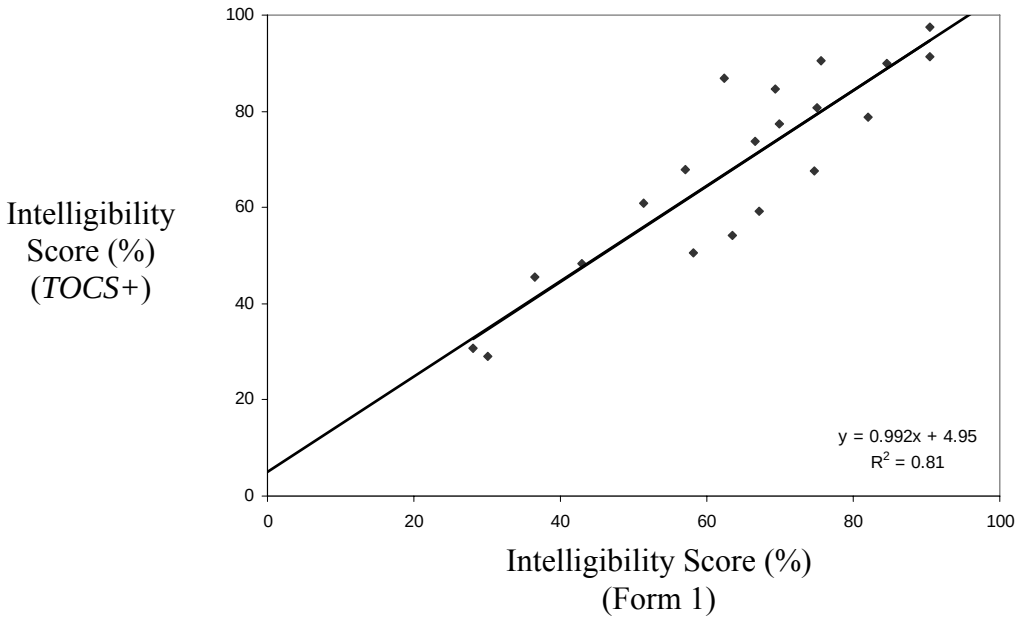

Figure J-1. Relationship between form 1 and *TOCS+* intelligibility scores.

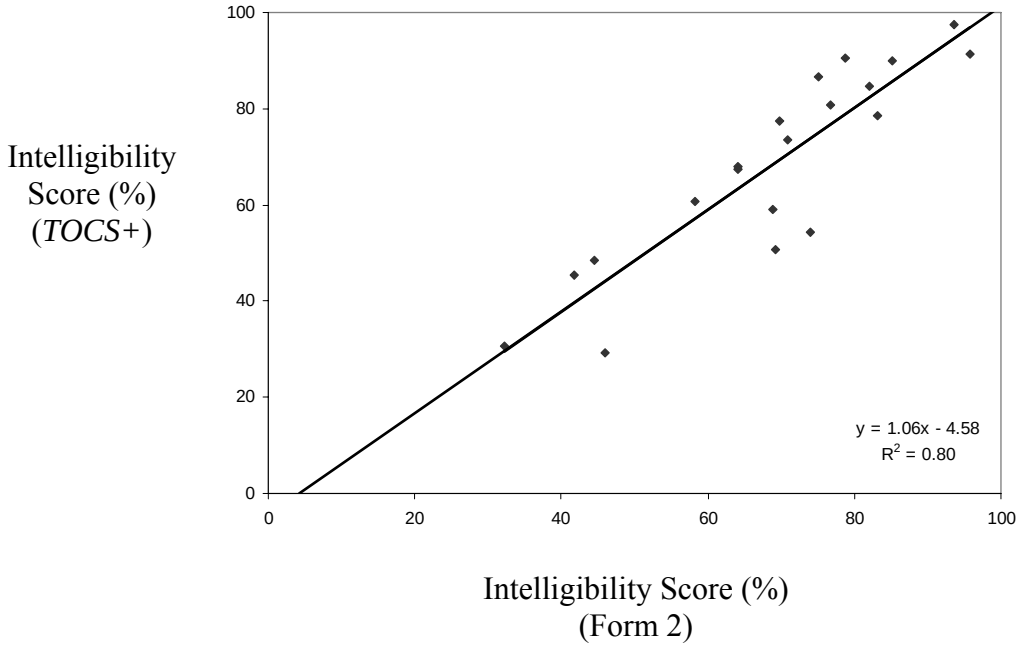

*Figure J-2***.** Relationship between form 2 and *TOCS+* intelligibility scores.

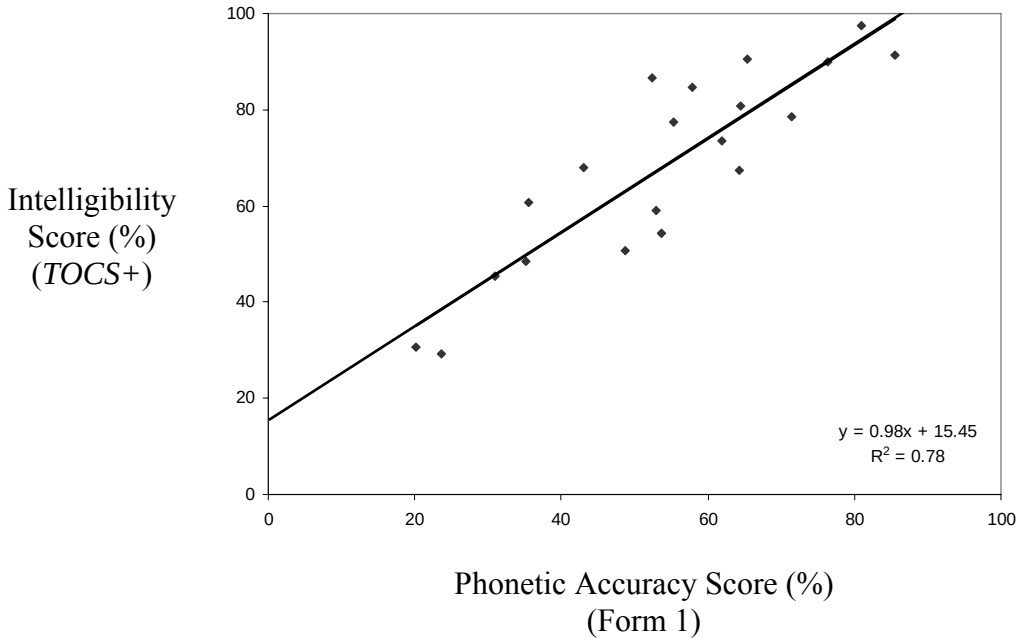

*Figure J-3***.** Relationship between form 1 phonetic accuracy scores and *TOCS+* intelligibility scores.

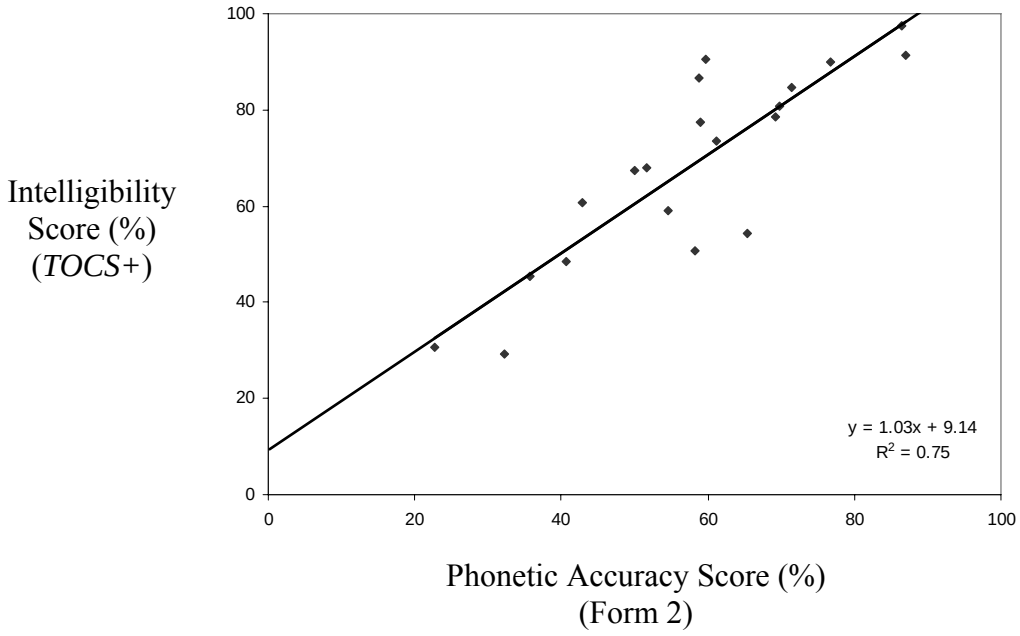

Figure J-4. Relationship between form 2 phonetic accuracy scores and *TOCS+* intelligibility scores.

Appendix K

Graphs of the Relationships between *SIP-CCLP Ver. 5* Scores and Session Two Spontaneous Sample Intelligibility Scores

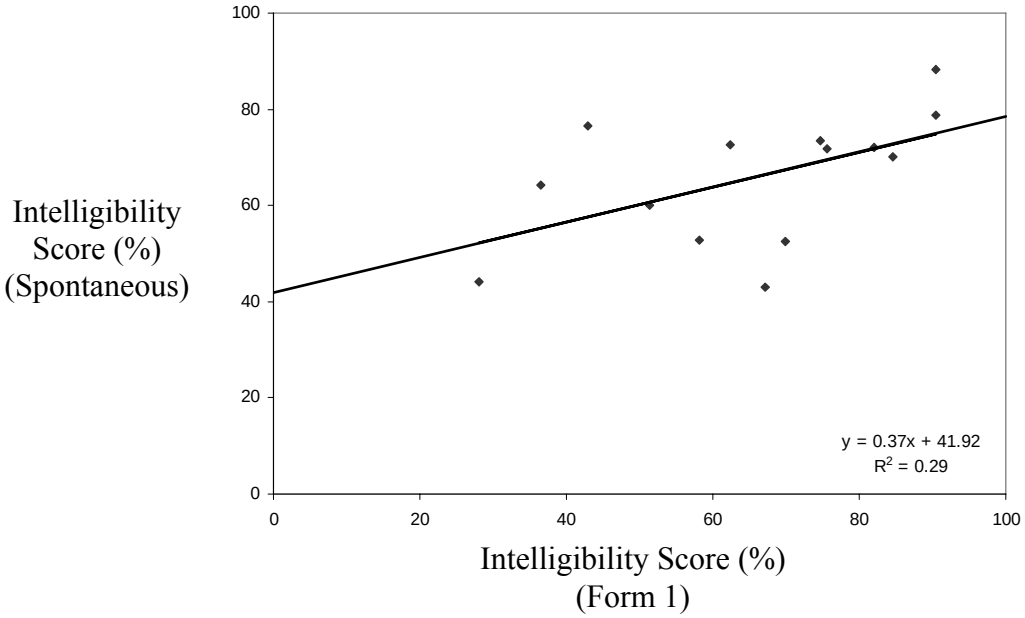

Figure K-1. Relationship between session one form 1 intelligibility scores and session two spontaneous sample intelligibility scores.

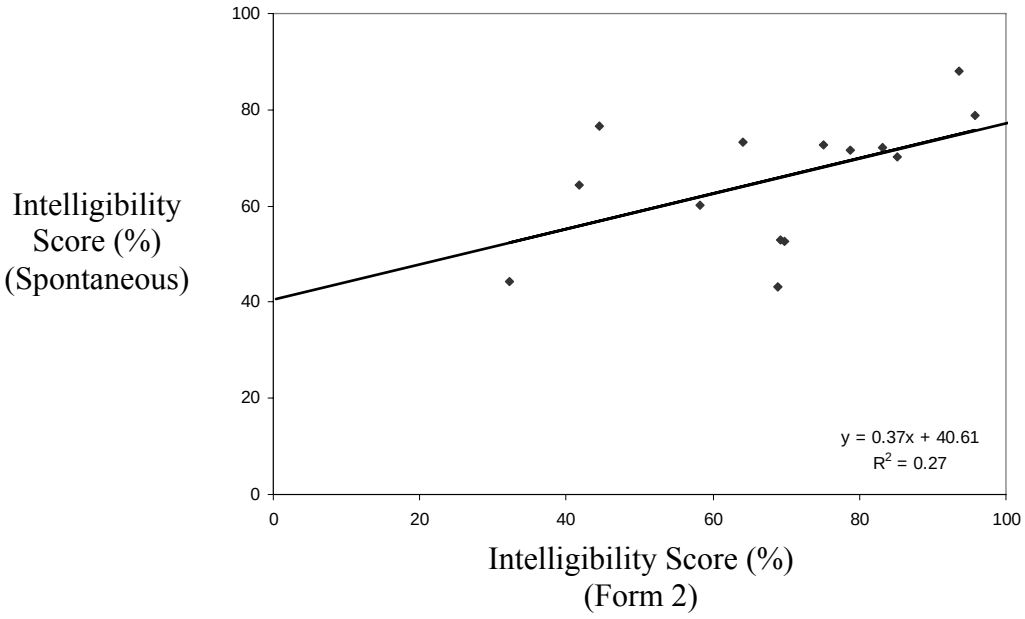

Figure K-2. Relationship between session one form 2 intelligibility scores and session two spontaneous sample intelligibility scores.

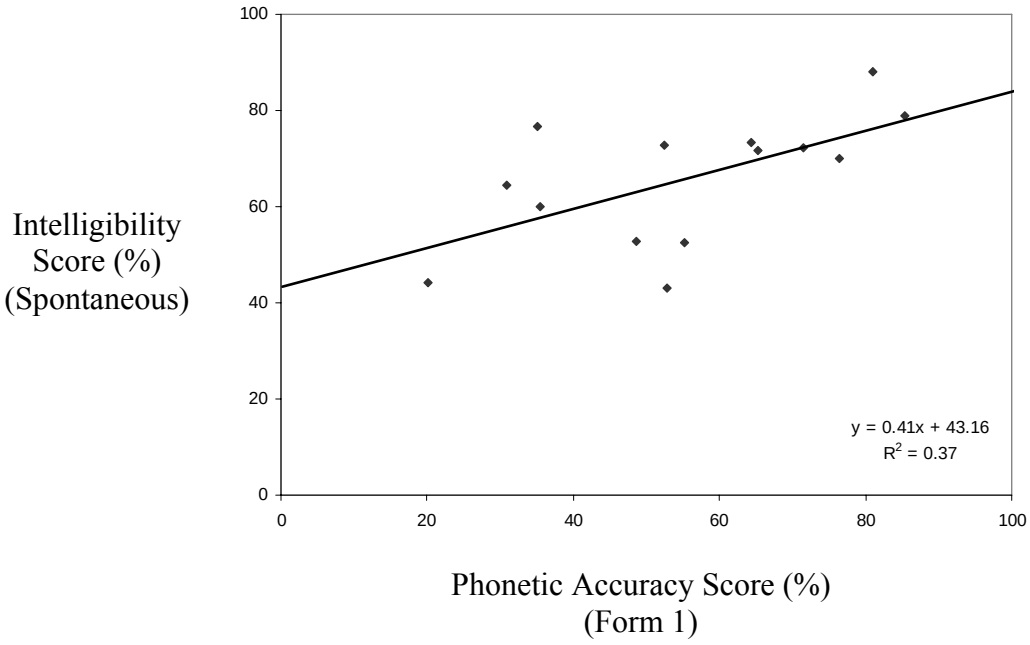

Figure K-3. Relationship between session one form 1 phonetic accuracy scores and session two spontaneous sample intelligibility scores.

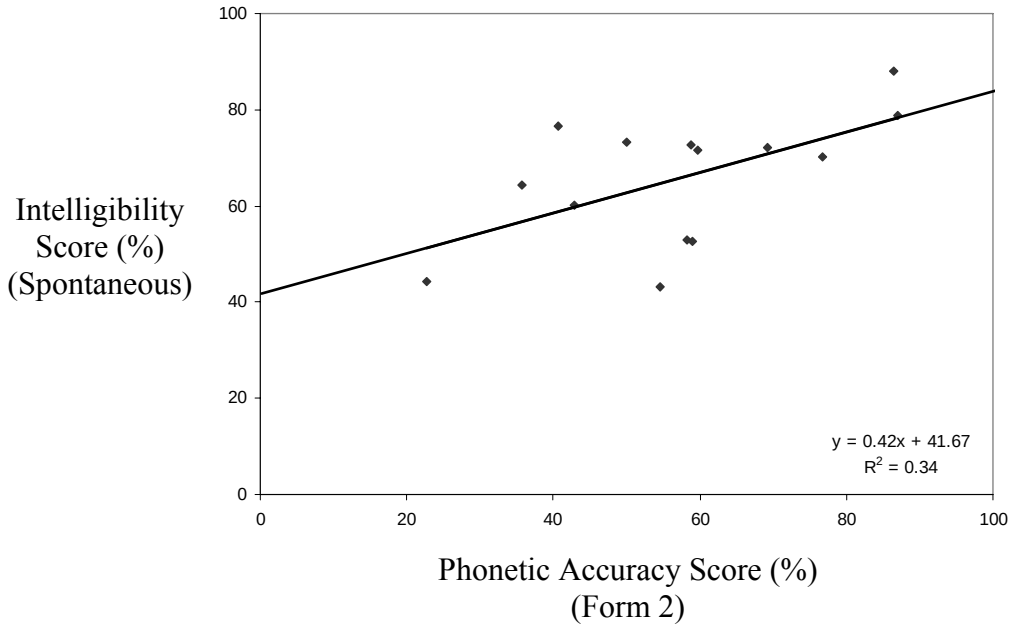

Figure K-4. Relationship between session one form 2 phonetic accuracy scores and session two spontaneous sample intelligibility scores.

Appendix L

Graphs of the Relationships between Hypernasality Ratings and *SIP-CCLP Ver. 5* Scores

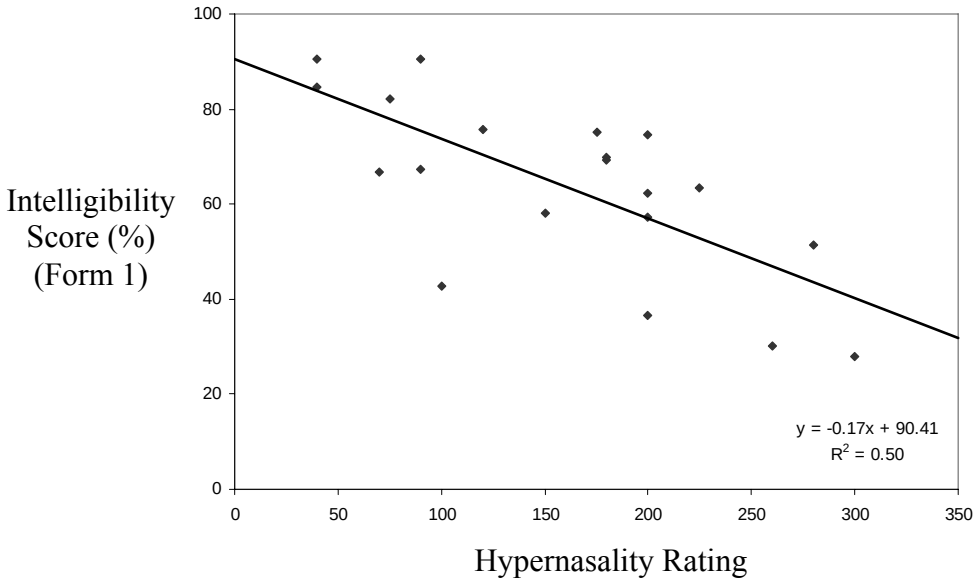

Figure L-1. Relationship between hypernasality ratings and form 1 intelligibility scores.

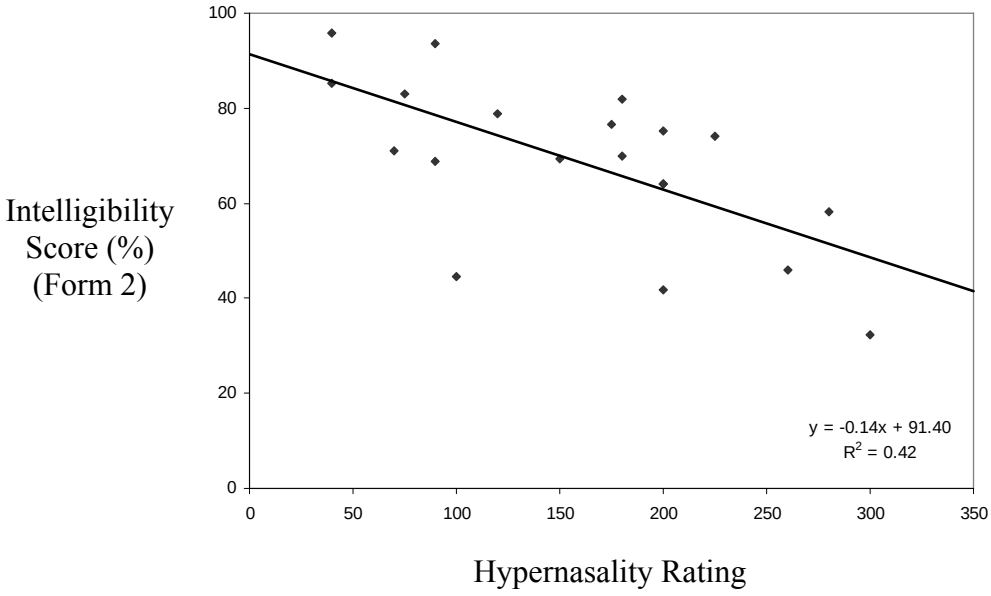

Figure L-2. Relationship between hypernasality ratings and form 2 intelligibility scores.

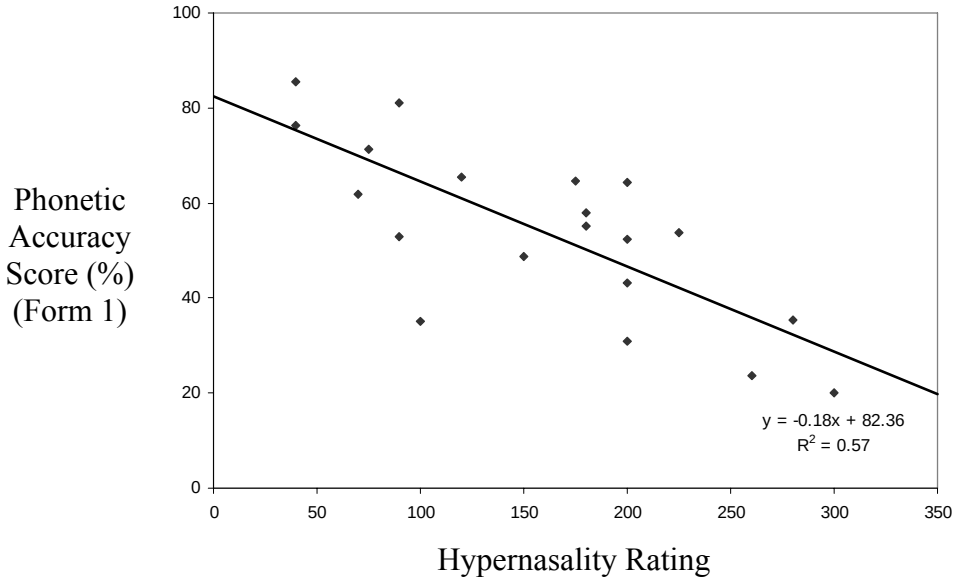

Figure L-3. Relationship between hypernasality ratings and form 1 phonetic accuracy scores.

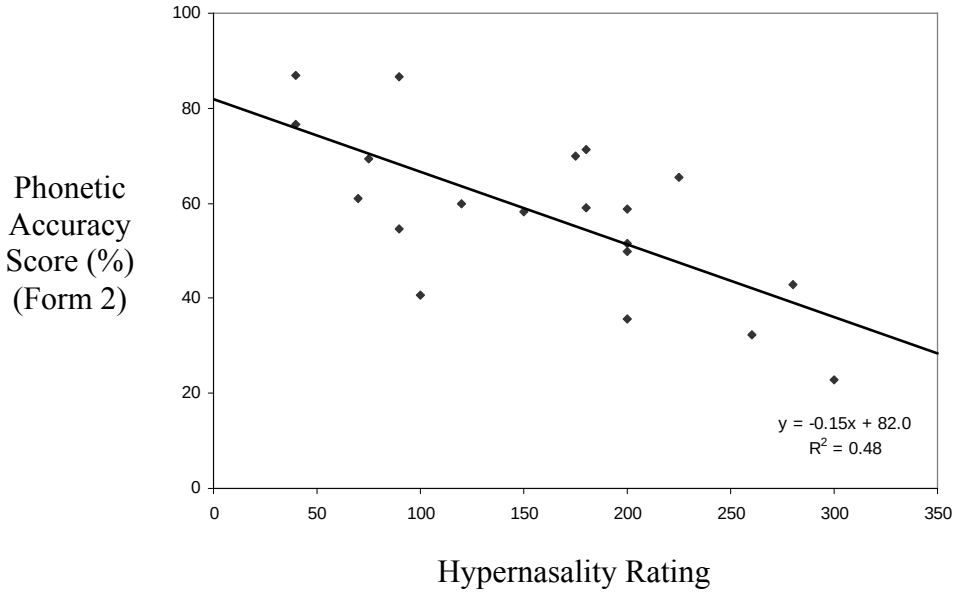

*Figure L-4***.** Relationship between hypernasality ratings and form 2 phonetic accuracy scores.

Appendix M

Graphs of the Relationships between Percentage of Consonants Correct and *SIP-CCLP Ver. 5* Scores

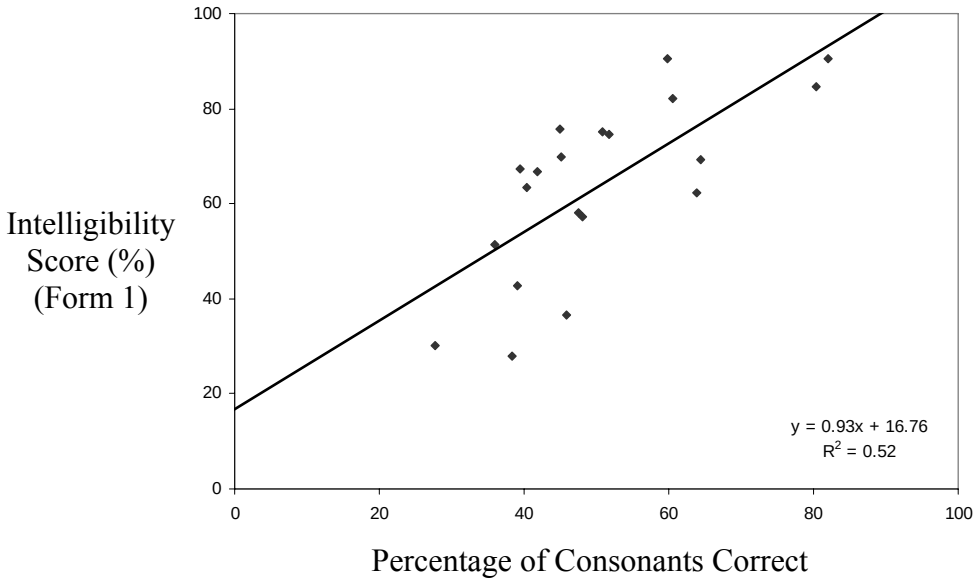

Figure M-1. Relationship between percentage of consonants correct and form 1 intelligibility scores.

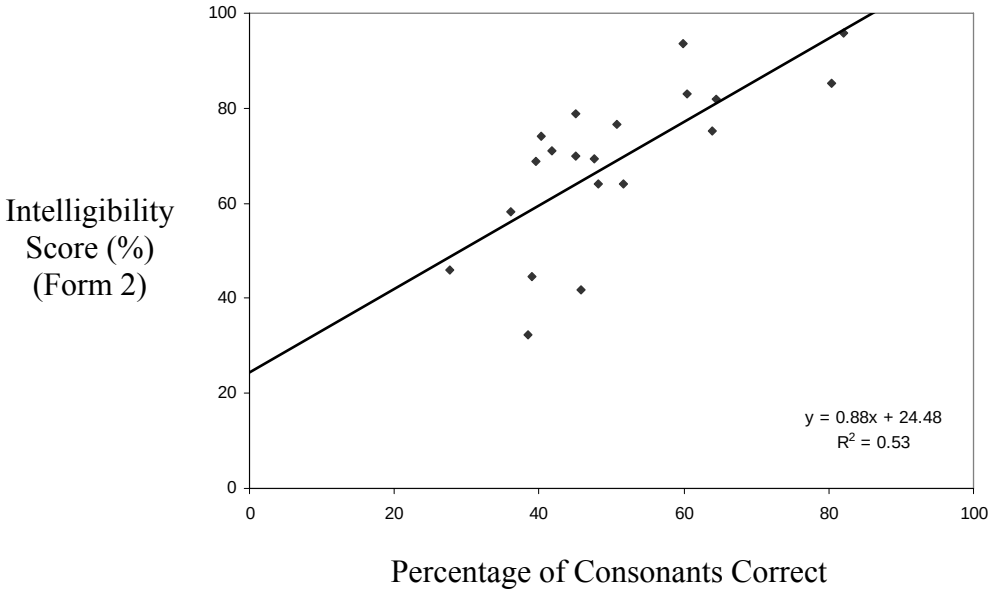

Figure M-2. Relationship between percentage of consonants correct and form 2 intelligibility scores.

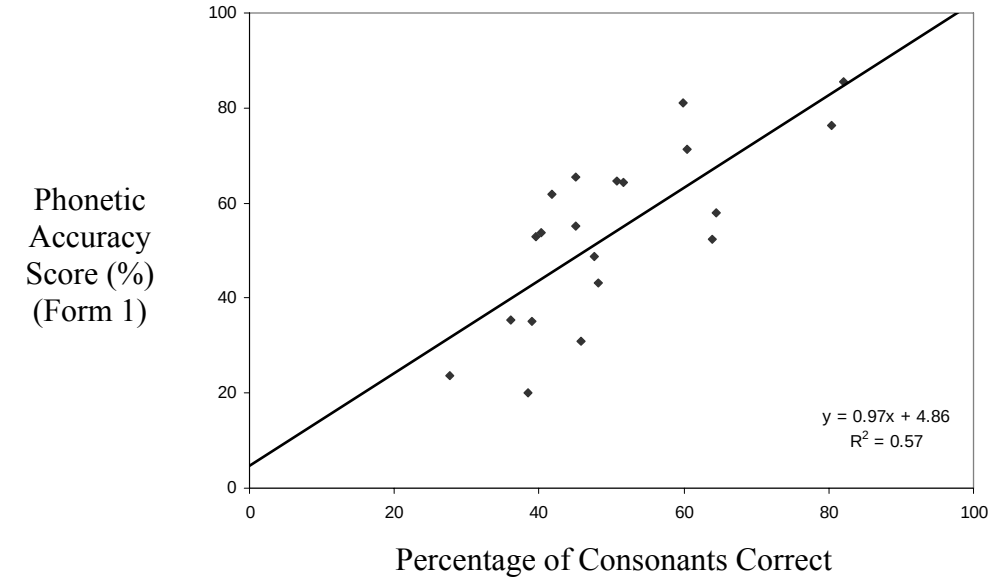

Figure M-3. Relationship between percentage of consonants correct and form 1 phonetic accuracy scores.

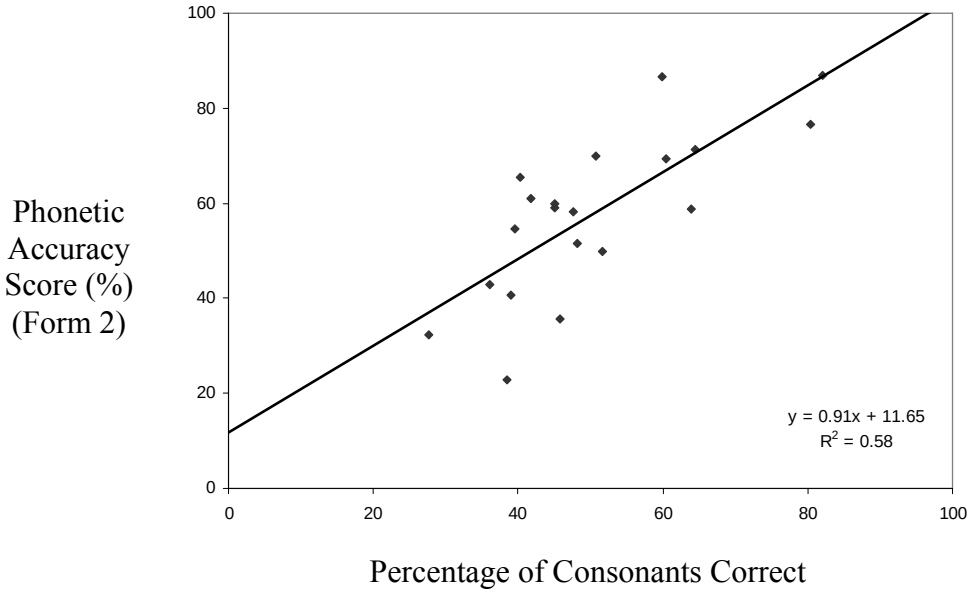

Figure M-4. Relationship between percentage of consonants correct and form 2 phonetic accuracy scores.

Appendix N

Results for CP11 in 2011 and 2012

Child Participant

At the time of initial data collection, CP11 was a 60-month old male with repaired bilateral cleft lip and palate adopted from China at 36 months of age. CP11 received surgery to insert a pharyngeal flap and palatal re-repair (i.e., closure of inner layers of palate) in July 2011, speech therapy throughout the school year (as part of Alberta Health and Wellness' Program Unit Funding), and lip revision surgery in April 2012. CP11's parents requested a second assessment to determine how his speech intelligibility had changed since his first assessment. In May 2012 (age: 73 months), recordings of the *SIP-CCLP Ver. 5* Form 1 and Form 2 stimuli words, *TOCS+* sentences (maximum sentence length: 7), *Zoo Passage* phrases, and a spontaneous speech sample were obtained as described in Chapters 3 and 4. As in 2011, form 1 was administered before form 2. CP11 had age-appropriate receptive language based on results from the *Fluharty -2* (Fluharty, 2001) on the day of testing. Nasalance scores on the picture-cued sub-test of the SNAP (Kummer, 2005) were also obtained on the day of testing and are reported in Table N-1.

Obtaining Listener Judgments

Twelve listeners were recruited to judge his recordings from the pool of students at the University of Alberta. All listeners had Canadian English as their first language and normal hearing as determined by a hearing screening. Each set of recordings was judged by two students in a graduate speech-language

pathology program and one student in a different course of study. None of the listeners had judged CP11's 2011 recordings. Listening sessions were conducted as described in Chapters 3 and 4. Dependent variables (i.e., *SIP-CCLP* phonetic accuracy score, intelligibility scores (*SIP-CCLP*, *TOCS+*, spontaneous sample), *SIP-CCLP* error patterns) were calculated as described in Chapters 3 and 4.

Results

CP11's speech intelligibility results in 2011 and 2012 are presented in Table N-2. His intelligibility scores were similar on the spontaneous speech samples over the two sessions. However, his speech intelligibility scores increased by at least 10% on the imitative sentences (*TOCS+*) and words (*SIP-CCLP*) tasks. Unexpectedly, CP11's intelligibility scores on form 1 were 18% higher than his intelligibility scores on form 2 in 2012. It is hypothesized that a lack of engagement during administration of form 2 may account, in part, for this difference. During administration of form 2, CP11 required extra encouragement to continue with the task and was increasingly restless.

The *SIP-CLLP* error types identified for CP11 are presented in Table N-3. Overall, fewer errors were identified in 2012 compared to 2011. Listeners achieved consensus on more errors in 2012 compared to 2011. A breakdown of the errors into the five error categories for form 1 and 2 are shown in Table N-4 and N-5, respectively. Fewer errors were identified in all categories on both forms in 2012 compared to 2011, except for sibilant errors on form 1. In 2011, listeners identified more than three errors in the following four error subtypes on form 1: nasals for obstruents (MPE), stopping (MPE), glottal stops for oral sounds

(PPE) and glottal fricatives for oral sounds (PPE). In 2012, listeners identified more than three errors for glottal stops for oral sounds (PPE) only. Similarly on form 2, listeners identified more than three errors for three error subtypes in 2011 (i.e., nasals for obstruents (MPE), glottal stops for oral sounds (PPE), and glottal fricatives for oral sounds (PPE)) and for two error subtypes in 2012 (i.e., velar sounds for obstruents (PPE) and glottal stops for oral sounds (PPE)).

On form 1, percent stops, fricatives and affricates correct were 40%, 6% and 0% in 2011 and 65%, 67% and 33% in 2012, respectively. On form 2, percent stops, fricatives and affricates correct were 30%, 33% and 0% in 2011 and 30%, 61% and 0% in 2012, respectively. On form 1, percent liquids, glides and nasals correct were 25%, 50% and 75% in 2011 and 75%, 100% and 100% in 2012, respectively. On form 2, percent liquids, glides and nasals correct were equal to or greater than 75% in 2011 and 2012. Across both forms, percent fricatives correct increased the most, which may have accounted for the increased percentage of consonant cluster targets identified correctly by listeners (i.e., 2011: 0% on both forms, 2012: F1 – 67%, F2 – 44%). Percent target sounds correct by manner are shown in Figure N-1 for the two forms.

Discussion

The purpose of this assessment was to evaluate how CP11's speech had changed since his first assessment in 2011. His speech intelligibility scores increased on the imitative sentences (*TOCS+*) and words (*SIP-CCLP*) tasks but did not increase on the spontaneous speech sample. On the *SIP-CCLP*, fewer errors were identified in 2012 compared to 2011. The glottal stop for oral sounds

(PPE) error pattern was the most frequently identified pattern in 2012 on both forms. Percent target sounds correct increased in all manner categories (i.e., stops, fricatives, affricates, liquids, glides, nasals, consonant clusters) on form 1 and three of the seven categories for form 2 (i.e., fricatives, liquids and consonant clusters).

As part of the evaluation of alternate test reliability (over forms and time) of *SIP-CCLP Ver. 5,* the minimal detectable change was calculated. The minimal detectable change was 11.69% for intelligibility scores and 12.63% for phonetic accuracy scores (see Chapter 3). Therefore, a change of at least 11.69% is needed to be confident, at the 95% level, that the change in intelligibility score reflects a real change in speech intelligibility measured at the single word level and not a change consistent with the measurement error of the test. Similarly, a change of at least 12.63% is needed to be confident that the change in phonetic accuracy score reflects a real change in phonetic accuracy as measured using *SIP-CCLP Ver. 5*. Examination of CP11's results reveals that the differences in intelligibility and phonetic accuracy scores from 2011 to 2012 were greater than the minimal detectable change for both forms. Therefore, we can be confident, at the 95% level, that these differences in scores reflect real changes in CP11's speech intelligibility and phonetic accuracy scores as measured using *SIP-CCLP Ver. 5*. Information on minimal detectable change is not available for the *TOCS+* sentence intelligibility test. However, Hodge and Gotzke (2010) estimated the measurement error for this measure using the limits of agreement procedure (Bland & Altman, 1986) for 18 children with dysarthria and cerebral palsy. These

estimates at the 95% level indicated that increases in *TOCS+* sentence intelligibility scores greater than 19.7% fall outside the range likely to be accounted for by measurement error. CP11's intelligibility scores increased by 28.84%, which is outside the 95% confidence interval for the limits of agreement. This increase likely represents a real change in CP11's intelligibility at the imitative sentence level. CP11's results on the *SIP-CCLP Ver. 5* and *TOCS+* sentence intelligibility test suggest that he is using new strategies to improve the intelligibility and clarity of his speech when imitating words or sentences such as more precise articulation of his consonant sounds.

Unlike his *SIP-CCLP Ver. 5* and *TOCS+* sentence intelligibility scores, CP11's intelligibility scores on the 100-word spontaneous speech sample were lower in 2012 than 2011 by 7.78%. The minimal detectable change for the 100-word spontaneous speech sample is 17.54% (see Chapter 4). As the difference in CP11's scores is less than the minimal detectable change, it is consistent with the measurement error of the test and does not reflect a change in performance. This result suggests that, although CP11 has learned strategies to improve his intelligibility, he has not yet generalized these to his spontaneous speech. Comparison of CP11's speaking rate on the 100-word spontaneous speech sample and the *TOCS+* sentence intelligibility test revealed that CP11 used a faster speaking rate on the spontaneous sample than on the imitated sentences (i.e., 116.5 words per minute and 91.7 words per minute, respectively). In connected speech, the *TOCS+* "models may help to slow his speech rate, facilitate accurate

articulation and thereby, increase intelligibility" (Hodge, 2009, case example 1, discussion, para. 1).

Overall, fewer errors were identified for CP11 in 2012 compared to 2011. On form 1, listeners did not achieve consensus on more items in 2011, suggesting that CP11 is using fewer non-English sound substitutions in 2012. On form 2, listeners did not achieve consensus on a similar number of items in 2011 and 2012. CP11's lack of engagement during administration of form 2 may have resulted in decreased attention and effort with respect to articulation accuracy, manifesting as less identifiable productions. Listeners identified more than three errors for fewer error sub-types in 2012 than in 2011. In 2012, two patterns accounted for the majority of errors identified by listeners on both forms (i.e., nasals for obstruents (MPE) and glottal stops for obstruents (PPE)). Speech therapy focusing on these patterns is recommended to help CP11 realize further speech intelligibility gains.

Comparison of CP11's 2011 and 2012 results revealed that percent fricatives, liquids and consonant clusters correct increased on both forms. Percent stops and affricates correct also increased on form 1 but did not increase on form 2. CP11's lack of engagement during administration of form 2 may account, in part, for the difference in results for the two forms. It may also be the reason why a greater percentage of fricatives than stops were identified correctly on form 2. Decreased attention and effort during administration of form 2 may have affected the degree to which he closed his velopharyngeal port during production of these sounds and may have affected his ability to create oral air pressure. As stops

require greater oral air pressure than fricatives to produce, incomplete closure would have had a greater impact on CP11's production of stops than fricatives. In describing a procedure for selecting targets for speech therapy intervention, Harding and Grunwell (1998) suggest that "in the presence of nasal escape, weak fricative production appears to be more readily achieved than weak plosive production" (p. 348). It is also possible that weak fricatives may be easier to identify than weak stops due to their longer duration.

CP11's results on the *SIP-CCLP Ver. 5* and *TOCS+* indicate that surgical and speech therapy intervention over the past year resulted in increased intelligibility at the imitative level. The lack of change in intelligibility scores obtained from the spontaneous speech sample indicate that further speech therapy intervention is needed to help CP11 generalize his clear speech strategies (e.g., decreased speaking rate and increased articulatory accuracy) to his conversational speech.

Table N-1

CP11's 2011 and 2012 Nasalance Scores on the SNAP sentences

	2011 (age: 60 months)	2012 (age: 73 months)
Bilabials	38	38
Alveolars	40	40
Velars	32	39
Sibilants	48	52
Nasals	59	67

Table N-2

CP11's 2011 and 2012 Intelligibility Scores (SIP-CCLP Ver. 5, Spontaneous Speech Sample, TOCS+) and SIP-CCLP Ver. 5 Phonetic Accuracy Score

		2011 (age: 60 months)	2012 (age: 73 months)
SIP-CCLP Ver. 5 Intelligibility Score	Form 1	28.04	65.08
	Form 2	32.28	47.09
SIP-CCLP Ver. 5 Phonetic Accuracy Score	Form 1	20.11	49.21
	Form 2	22.75	37.57
Spontaneous Speech Intelligibility Score		61.37	53.59
Spontaneous Sample Speaking Rate (wpm[1])		148.40	116.48
TOCS+ Sentence Intelligibility Score[2]		30.74	59.58
TOCS+ Sentence Speaking Rate (wpm)		84.98	91.70

Note. [1]wpm = words per minute. [2]The maximum sentence length of the *TOCS+* sentence test was six words in 2011 and seven words in 2012.

Table N-3

Types of Errors Identified for CP11 in 2011 and 2012

Error Type	2011		2012	
	Form 1	Form 2	Form 1	Form 2
Foil	23	27	14	16
Listener-generated	9	4	1	3
Can't identify	1	2	0	4
Unclassified	1	0	0	0
No consensus	15	8	6	9
TOTAL	48	41	21	32

Table N-4

SIP-CCLP Ver. 5 Form 1 Error Patterns Identified for CP11 in 2011 and 2012

Error Category	2011		2012	
	Number	Error Subtype	Number	Error Subtype
Manner Preference Errors	2	Liquids for obstruents	1	Liquids for obstruents
	5	Nasals for obstruents	3	Nasals for obstruents
	1	Nasals for liquids		
	5	Stopping		
	2	Gliding	1	Gliding
Place Preference Errors	2	Velar sounds for obstruents	1	Velar sounds for obstruents
	7	Glottal stops for oral sounds	4	Glottal stops for oral sounds
	4	Glottal fricative for oral sounds		
Voicing Errors				
Sibilant Errors			1	Palatal fricatives for alveolar fricatives
			1	Labiodental fricatives for alveolar fricatives
	1	Fronting	1	Fronting
Cluster Errors			1	Deletion of an obstruent from an obstruent-obstruent cluster
	3	Deletion of an obstruent from an obstruent-sonorant cluster	1	Deletion of an obstruent from an obstruent-sonorant cluster

Table N-5

SIP-CCLP Ver. 5 Form 2 Error Patterns Identified for CP11 in 2011 and 2012

Error Category	2011		2012	
	Number	Error Subtype	Number	Error Subtype
Manner Preference Errors			1	Glides for obstruents
			1	Liquids for obstruents
	8	Nasals for obstruents	3	Nasals for obstruents
	1	Gliding		
Place Preference Errors			4	Velar sounds for obstruents
	7	Glottal stops for oral sounds	5	Glottal stops for oral sounds
	6	Glottal fricatives for oral sounds	1	Glottal fricatives for oral sounds
Voicing Errors				
Sibilant Errors	1	Palatal fricatives for alveolar fricatives	1	Palatal fricatives for alveolar fricatives
	2	Labiodental fricatives for alveolar fricatives	1	Labiodental fricatives for alveolar fricatives
	1	Fronting		
Cluster Errors	2	Deletion of an obstruent from an obstruent-obstruent cluster	2	Deletion of an obstruent from an obstruent-obstruent cluster
	2	Deletion of an obstruent from an obstruent-sonorant cluster		
	1	Backing and cluster reduction		

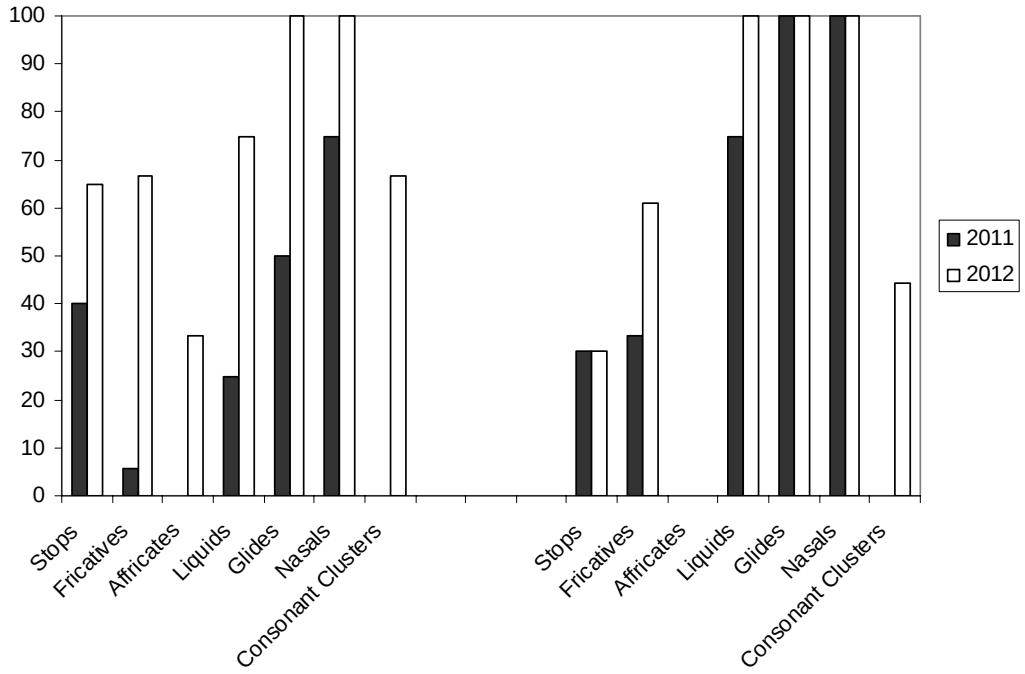

Figure N-1. Percentage of target sounds correct organized by manner for CP11 in 2011 and 2012

Appendix O

Sample Analysis Output

SIP-CCLP ver.5.0 - Closed Set Analysis

Files: CP11_18Apr11_F1_C2.xls
CP11_18Apr11_F1_C4.xls
CP11_18Apr11_F1_C5.xls

Part One: Comparison of Three Listener's Responses for the 63 Contrast Items

Code	Target Word	Foil Word 1	Foil Word 2	Foil Word 3	Mean RT	L1	L2	L3		L1 Rating	L2 Rating	L3 Rating
Fricatives	As	age	ate	ace	15.42733	ate	*ay*	*A*	0	4	5	5
Affricates	G	lee	D	he	7.609667	lee	*ye*	*E*	0	4	6	6
Fricatives	V	zee	bee	E	14.08833	bee	V	*me*	0	4	2	6
Affricates	badge	bad	back	batch	12.10467	*CI*	*anne*	*an*	0	7	6	6
Fricatives	bash	batch	back	bath	11.88567	bash	bath	back	0	2	4	4
Stops	bat	mat	at	pat	1.667	at	at	at	-1	4	3	3
Stops	bead	bean	bee	beat	10.79667	bean	bean	bee	-1	4	4	4
Stops	bee	we	E	D	2.989667	bee	bee	bee	1	2	2	1
Stops	cap	nap	tap	gap	5.974	*hap*	tap	nap	0	5	4	4
Affricates	cheese	tease	keys	Gs	13.06733	keys	*he*	*CI*	0	4	6	7
Affricates	chew	you	shoe	who	5.734333	shoe	*two*	*two*	0	4	5	6
Stops	cow	now	ow	how	3.635667	how	how	cow	-1	4	4	2
Stops	deer	near	year	ear	1.958333	ear	ear	ear	-1	4	3	3
Stops	dot	knot	caught	bought	15.14567	dot	*ought*	caught	0	2	5	4
Stops	dough	go	bow	toe	9.609333	*CI*	*ough*	*CI*	0	7	5	7
Clusters	drip	whip	trip	dip	5.521	trip	whip	*lip*	0	4	4	6
Fricatives	fail	whale	pail	veil	4.213667	pail	pail	fail	-1	4	4	1
Nasals	fan	fad	fat	fang	2.791667	fan	fan	fan	1	1	1	1
Fricatives	fat	mat	sat	hat	1.422	hat	hat	hat	-1	3	3	3
Stops	go	O	hoe	dough	4.979	O	O	O	-1	3	3	3
Stops	goat	note	oat	coat	2.541667	oat	oat	oat	-1	4	3	4

Affricates	jeep	leap	sheep	cheep	14.401	jeep	*-eep*	*CI*	0	2	5	7
Nasals	knee	D	tea	E	2.88	knee	knee	knee	1	2	1	1
Stops	leak	lean	lee	league	3.932667	leak	leak	leak	1	2	2	1
Liquids	lee	see	we	E	4.245	lee	lee	lee	1	1	1	2
Liquids	lock	sock	walk	knock	2.370333	knock	knock	lock	-1	3	3	1
Stops	log	lawn	long	lock	5.859667	long	long	long	-1	4	4	4
Nasals	long	log	lock	lawn	1.875333	long	long	long	1	1	2	1
Nasals	mat	bat	at	pat	6.093667	*nat*	mat	*rack*	0	6	2	6
Fricatives	pass	pal	pat	path	11.781	pass	*paih*	*CI*	0	2	6	7
Stops	pat	mat	at	bat	1.880333	pat	pat	pat	1	1	1	1
Stops	pea	we	E	tea	1.999667	pea	pea	pea	1	2	2	1
Affricates	peach	pete	pea	peek	7.406	peach	pete	*pink*	0	2	4	6
Liquids	rail	sail	whale	hail	9.515667	whale	whale	rail	-1	4	4	1
Stops	robe	roam	row	rope	8.161667	roam	roam	roam	-1	4	4	3
Stops	rope	roam	row	robe	3.083	rope	rope	rope	1	1	2	1
Fricatives	sap	yap	chap	snap	9.974	*tap*	*cap*	*cap*	0	6	6	6
Fricatives	see	knee	she	zee	8.250333	*he*	*hey*	*CI*	0	6	6	7
Fricatives	sell	well	tell	fell	8.578	sell	tell	tell	-1	2	4	4
Fricatives	she	D	E	see	7.494667	see	see	see	-1	4	4	4
Fricatives	shop	chop	top	hop	4.328333	top	top	shop	-1	4	4	2
Fricatives	sick	lick	wick	thick	8.166667	*hick*	*hick*	thick	0	5	5	4
Clusters	ski	tea	he	see	13.25533	see	he	*E*	0	4	4	6
Clusters	slip	snip	flip	lip	6.760333	slip	lip	lip	-1	2	4	4
Clusters	snow	hoe	show	no	6.151333	no	no	no	-1	3	3	3
Clusters	spell	smell	fell	sell	14.08333	*hell*	*ell*	*CI*	0	6	5	7
Clusters	stay	neigh	say	k	9.265333	*ay*	*ay*	*A*	0	5	5	6
Clusters	stick	sick	tick	kick	11.03667	*ick*	*ick*	sick	0	5	5	4
Clusters	straight	wait	skate	rate	7.234333	wait	wait	*make*	-1	4	4	6
Stops	tap	yap	nap	chap	5.828	tap	tap	*CI*	1	2	2	7
Stops	tea	knee	key	pea	13.422	key	key	key	-1	4	4	4
Fricatives	thick	lick	sick	tick	4.453	sick	*pick*	thick	0	4	5	1

Stops	toe	O	go	dough	3.400667	toe	toe	go	1	2	2	4
Clusters	trail	whale	rail	tail	18.01067	*snail*	tail	rail	0	6	4	4
Fricatives	veil	whale	mail	fail	1.125	mail	mail	mail	-1	3	4	3
Glides	wheel	feel	seal	eel	4.187333	*meal*	wheel	*meal*	0	6	2	6
Stops	white	whine	why	wide	3.343667	white	white	*like*	1	2	2	6
Fricatives	wife	while	wipe	why	8.281333	wipe	wipe	why	-1	4	4	4
Liquids	write	sight	fight	white	6.036667	white	white	*like*	-1	4	3	6
Glides	year	fear	sear	ear	3.223667	year	year	*near*	1	2	1	6
Fricatives	zee	D	V	see	4.37	*lee*	V	*lee*	0	6	3	6
Fricatives	zip	nip	lip	yip	1.927	lip	lip	lip	-1	4	4	3
Fricatives	zoo	who	you	coo	4.953	zoo	zoo	you	1	2	1	4

Part Two: Summary of Results

A. Word Identification/Distortion Scores

a) Total '1s' (Correct and Clear) = 12.169%

1 - L1 = 5/63

2 - L2 = 6/63

3 - L3 = 12/63

b) Total '2s' (Correct and Distorted) = 15.873%

1 - L1 = 16/63

2 - L2 = 11/63

3 - L3 = 3/63

c) Intelligibility Score = 28.042%

1 - L1 = 33.333%

2 - L2 = 26.984%

3 - L3 = 23.81%

d) Phonetic Accuracy Score = 20.106%
1 - L1 = 20.635%
2 - L2 = 18.254%
3 - L3 = 21.429%

B. Contrast Targets Correct Summary (1 or 2)
a) Total Contrast Targets Correct = 14/63
1 - Total Stops Correct = 8/20
2 - Total Fricatives Correct = 1/18
3 - Total Affricates Correct = 0/6
4 - Total Liquids Correct = 1/4
5 - Total Glides Correct = 1/2
6 - Total Nasals Correct = 3/4
7 - Total Consonant Clusters Correct = 0/9

C. Contrast Error Profiling (3 or 4)
a) Total Contrast Errors (foil) = 23
1 - Total Manner Preference Errors = 12

Stops	bead	bean
Fricatives	fail	pail
Stops	log	long
Liquids	rail	whale
Stops	robe	roam
Fricatives	sell	tell
Fricatives	shop	top
Consonant Clusters	straight	wait
Fricatives	veil	mail
Fricatives	wife	wipe
Liquids	write	white
Fricatives	zip	lip

2 - Total Place Preference Errors = 8

Stops	bat	at
Stops	cow	how
Stops	deer	ear
Fricatives	fat	hat
Stops	go	O
Stops	goat	oat
Liquids	lock	knock
Stops	tea	key

3 - Total Sibilant Errors = 1

Fricatives	she	see

4 - Total Voicing Errors = 0

5 - Total Consonant Cluster Errors = 2

Clusters	slip	lip
Clusters	snow	no

b) Total Contrast Errors (other) = 26

	Target	Listener 1	Listener 2	Listener 3
Fricatives	As	ate	*ay*	*A*
Affricates	G	lee	*ye*	*E*
Fricatives	V	bee	V	*me*
Affricates	badge	*CI*	*anne*	*an*
Fricatives	bash	bash	bath	back
Stops	cap	*hap*	tap	nap
Affricates	cheese	keys	*he*	*CI*
Affricates	chew	shoe	*two*	*two*
Stops	dot	dot	*ought*	caught
Stops	dough	*CI*	*ough*	*CI*
Clusters	drip	trip	whip	*lip*
Affricates	jeep	jeep	*-eep*	*CI*

Nasals	mat	*nat*	mat	*rack*
Fricatives	pass	pass	*paih*	*CI*
Affricates	peach	peach	pete	*pink*
Fricatives	sap	*tap*	*cap*	*cap*
Fricatives	see	*he*	*hey*	*CI*
Fricatives	sick	*hick*	*hick*	thick
Clusters	ski	see	he	*E*
Clusters	spell	*hell*	*ell*	*CI*
Clusters	stay	*ay*	*ay*	*A*
Clusters	stick	*ick*	*ick*	sick
Fricatives	thick	sick	*pick*	thick
Clusters	trail	*snail*	tail	rail
Glides	wheel	*meal*	wheel	*meal*
Fricatives	zee	*lee*	V	*lee*