Published in final edited form as:
J Speech Lang Hear Res. 2011 February ; 54(1): 1–18. doi:10.1044/1092-4388(2010/10-0030).

Developing a Weighted Measure of Speech Sound Accuracy

Jonathan L. Preston[1], **Heather L. Ramsdell**[2], **D. Kimbrough Oller**[2], **Mary Louise Edwards**[3], and **Stephen J. Tobin**[4]

[1]Haskins Laboratories, New Haven, CT

[2]The University of Memphis, Memphis, TN

[3]Syracuse University, Syracuse, NY

[4]University of Connecticut, Storrs, CT

Abstract

Purpose—The purpose is to develop a system for numerically quantifying a speaker's phonetic accuracy through transcription-based measures. With a focus on normal and disordered speech in children, we describe a system for differentially weighting speech sound errors based on various levels of phonetic accuracy with a Weighted Speech Sound Accuracy (WSSA) score. We then evaluate the reliability and validity of this measure.

Method—Phonetic transcriptions are analyzed from several samples of child speech, including preschoolers and young adolescents with and without speech sound disorders and typically developing toddlers. The new measure of phonetic accuracy is compared to existing measures, is used to discriminate typical and disordered speech production, and is evaluated to determine whether it is sensitive to changes in phonetic accuracy over time.

Results—Initial psychometric data indicate that WSSA scores correlate with other measures of phonetic accuracy as well as listeners' judgments of severity of a child's speech disorder. The measure separates children with and without speech sound disorders. WSSA scores also capture growth in phonetic accuracy in toddler's speech over time.

Conclusion—Results provide preliminary support for the WSSA as a valid and reliable measure of phonetic accuracy in children's speech.

One of the continuing needs in the fields of developmental phonology and speech-language pathology is for accurate, sensitive, and viable measures of speech production for research and clinical practice (Flipsen, Hammer, & Yost, 2005). Phonetic transcription is the basis for most analyses of normal and impaired speech production in children. However, associating a meaningful numeric value to phonetic transcriptions of children's speech is a tremendous challenge. This is not a small matter, given that such numeric values are then used in statistical analyses, to track changes in speech over time, or to compare to normative data. It is therefore critical to use measures that are reliable and meaningful and that have known psychometric properties. Many existing measures lack sufficient investigation of reliability and validity, or are not sensitive to the nature of the types of errors that children produce. To meaningfully quantify speech production, it is essential that we measure it in such a way that clinically relevant aspects are being captured (i.e., the degree of phonetic accuracy in a child's productions). Consequently, the current study presents a description of, and preliminary psychometric data on, a weighted measure of speech sound accuracy.

Correspondence concerning this article should be addressed to Jonathan L. Preston, Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511. preston@haskins.yale.edu.

In both research and clinical practice, measurement issues are not trivial. A recent study of "independent measures" used to describe the productive phonology of toddlers without reference to the corresponding adult forms (e.g., phonetic inventory, word shape) found many of the commonly used measures to be unreliable between different testing occasions (Morris, 2009). This calls into question some of the common practices of clinicians and researchers, who may be using measures with psychometric properties that have not been sufficiently investigated. In research, inferential statistics rely on measured variables to make inferences about particular relationships. Traditional linear models rely on ratios between variance attributable to the factors of interest and variance attributable to error. Although random error can never be avoided, we can conduct more efficient experimental or observational studies if we do not increase the error term by using poor measures. That is, measures that are sensitive and reliable are needed to reach appropriate statistical conclusions. Thus, both clinical and research methods could benefit from psychometric investigations of measures before they become widely implemented. Here, we focus on demonstrating the feasibility of establishing validity and reliability of a relational measure of productive phonology (i.e., one that compares children's productions with the corresponding adult forms).

Flipsen et al. (2005) reviewed several relational measures thought to index the severity of speech sound disorders (SSDs). They suggested that speech-language pathologists' judgments of severity might be a useful standard for comparison. Many of the measures derived from phonetic transcriptions of the speech of 17 children did not correlate strongly with these severity ratings. In addition, not all of those measures take into account the degree of phonetic accuracy, and many of the measures that were reviewed lack published psychometric data. Clearly, further research is needed to develop valid and reliable measures of phonetic accuracy.

One measurement scheme that has obtained widespread use is Percent Consonants Correct (PCC) (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997; Shriberg & Kwiatkowski, 1982). PCC is a reasonable starting point for discussing numeric quantification of children's speech given its simplicity of calculation, ease of interpretation, and psychometric properties that are well reported. It has been used as a measure of severity of involvement of speech impairment in many studies. Variants on this theme have been described, such as PCC-Revised (PCC-R, which does not consider common or uncommon clinical distortions as errors), PCC-Adjusted (PCC-A, which does not consider common clinical distortions as errors), and Percent Phonemes Correct (which considers vowels in the calculation as well as consonants). As reported by Shriberg et al. (1997), each of the variants on PCC might be appropriate for different clinical or research purposes. However, all of them are based on binary (right/wrong) judgments of speech sounds, and none of them assigns different weights to different types of errors.

Phonetic accuracy is important to consider when attempting to quantify child speech. In particular, when a child's productions are inaccurate, it may be helpful to consider the types of sound errors made. For example, some errors (e.g., deletions) may have a greater impact on intelligibility than others (e.g., distortions). However, many analysis schemes and standardized tests of articulation take into account only the number of errors produced in a sample while ignoring the nature of those errors. Atypical sound errors (i.e., those that do not commonly occur in typically developing children) may be especially important to consider, as they may suggest that children have arrived at unusual solutions to satisfy the constraints of their phonological system (Leonard, 1985). Atypical errors may also be useful in classifying children with SSDs into various subgroups (Dodd, 2005); they are likely to have a significant impact on intelligibility (Dodd & Iacano, 1989), and they have been found to relate to early phonological processing skills (Preston & Edwards, 2010; Rvachew, Chiang, & Evans, 2007). Thus, measures that weight errors according to the degree of difference between the produced sound and the target sound based on phonologically grounded evidence would be of value clinically and in research.

Consider the adult target “Sue”, /su/. Multiple mispronunciations of this word are possible. Phoneme substitutions might include [tu], [du], or [gu]. Phoneme omissions would include [u]. By simply tallying errors, these differences cannot be captured. For example, [tu] involves a change of one feature (manner of articulation); [du] involves a change in both manner of articulation and voicing; [gu] involves an error in manner and voicing, along with an unusual pattern of substituting a velar for an alveolar place of articulation. PCC would consider all of these errors to be the same. However, it is plausible (and arguably preferable) to rank these errors (from most to least accurate: [tu], [du], [gu], [u]) and to count them differently. Below, we describe how we assign numeric values to each of these error types.

Additionally, it is rarely the case that vowels are included when examining speech errors. However, vowels have been found to be in error in many children with SSDs (Pollock, 1991), including those with childhood apraxia of speech (Crary, 1984). Vowel errors should therefore be considered when examining a child’s speech production accuracy. A weighted measure of speech sound accuracy should also take vowel production into account, differentially weighting vowel errors that are closer to and farther from the target.

One issue that frequently arises when using transcription-based measures is between-transcriber reliability. Point-by-point reliability of transcription-based measures is often in the 90-95% range for broad transcription of consonants and around 80% for narrow transcription (Shriberg, et al., 1997; Shriberg & Lof, 1991). As a general rule, it is harder to achieve reasonable agreement for disordered speech. However, as described by Oller and Ramsdell (2006), it is often the case that two transcribers perceive the child’s production similarly, but may differ in their assignment of one or two features. Minor differences in two listeners’ perception of a phone (e.g., [do] vs. [to]) will inherently result in reduced reliability. However, if the measurement system considers all of the features of the production that were perceived the same (i.e., that both transcribers perceive an alveolar stop followed by [o]), the reliability assessment can be more precise.

In validating a measurement tool that quantifies phonetic accuracy in children’s speech, it is important to show, from converging lines of evidence, that the measure captures what it is intended to capture. Thus, in addition to carefully constructing the measure based on theoretically relevant principles, it is important to show that the measure can differentiate between children with and without SSDs and thus capture the relative degree of impairment, as evidenced by correlations with standardized tests, listener judgments, and other types of transcription-based analyses. Sensitivity of a measure, which is the capacity to reflect systematic variation, change, or differences between different events, should also be considered (Kazdin, 2003). Further, sensitive measures of phonological development should be able to capture small improvements in speech production as children become more phonetically accurate over time.

Description of the Weighted Measure of Speech Sound Accuracy (WSSA)

The WSSA is in some ways similar to the commonly used clinical and research metric, PCC (Shriberg, et al., 1997; Shriberg & Kwiatkowski, 1982), in that sounds are vertically aligned to compare an adult “target” form with the child’s produced form of a word. However, the new measure is intended to be more fine-grained in that it weights error types differently (e.g., phoneme omissions and unusual changes in manner, place, and voicing are weighted heavily, but errors involving a common substitution are given smaller weights), rather than using a binary (right-wrong) score for each sound. The WSSA is implemented in the LIPP

(Logical International Phonetics Program) software program (Oller & Delgado, 1999), and relies on definitions of errors specified in a LIPP alphabet and analysis program described by Oller and Ramsdell (2006). These definitions are based on the match between a listener's phonetic transcriptions and an adult form. The definitions used in WSSA are an adaptation of those described by Oller and Ramsdell (2006) designed to compare two transcribers' renditions of a talker's speech. Any two transcriptions of the same vocalization can, however, be compared with this sort of analysis, as used here. Computations are based on comparison between transcriptions of a child's speech and the "target" adult form of the word (adapting the adult form for dialectally acceptable variations).

Because there are many possible errors between different consonants and different vowels, a complete review of all permutations is not possible; however, a review of the major concepts is presented. (The LIPP analysis program and accompanying alphabet can be obtained from the first author.) For a given segment, word, utterance, or sample, a score is computed that reflects how well, on average, the child's production matches the target form in the number of segments (termed global structural agreement) and in features of the segments that are represented (termed featural agreement). In general, deletions of phonemes are weighted most heavily because they represent structural changes. With respect to features, major changes (e.g., producing a glottal sound for an orally articulated sound) are weighted heavily and minor feature changes (e.g., producing a homorganic stop in place of a fricative) are weighted less. Scores can range from 0.0 (no match in features and number of segments) to 1.0 (complete match of features and segments) in the segment, word, utterance, or sample. The following is a description of how the WSSA weights errors for sound substitutions, followed by the penalties for additions and deletions of sounds. Finally, we outline how the WSSA is calculated, and provide computational examples in Appendix A.

Consonant Substitutions

Consonant substitutions are weighted based on features of manner, place and voicing. If the child production and the target form of a consonant are exact matches, all features of a consonant are correctly represented (i.e., the segment consonant is produced correctly) and the feature agreement is 1.0. If the child produces a substitution, "credit" is given for place, manner, and voicing features that are correct (each is worth 0.333 if produced correctly). For example, minor errors in place of articulation result in a small penalty (score reduction), whereas major errors result in a much bigger penalty. Based on developmental phonological principles, the direction of change also plays a role in weighting; for example, a backing error (e.g., coronal → dorsal) results in a greater penalty than a fronting error (dorsal → coronal), because backing errors are rarely seen in typically developing children (Edwards & Shriberg, 1983). Table 1 lists feature weights for consonant substitution errors. Following Oller & Ramsdell (2006), errors are ranked as teeny, small, big, or huge. Each place, manner or voicing feature is credited with a maximum score of 0.333, and a rank-ordering of errors within each feature results in equal-stepped reductions. That is, consonant manner or place each could be completely accurate (with a score of 0.333), or reduced by 25% for a teeny error [0.333 − (0.25 *0.333)= 0.250], by 50% for a small error [0.333− (0.50* 0.333) = .167], by 75% for a big error [0.333 −(0.75*0.333) = 0.0833], or by 100% for a huge error [0.333 − (1.0 *0.333) = 0]. Similarly, voicing (total value of 0.333) could be completely accurate or could be reduced by 33% for a teeny error, 67% for a small error, or 100% for a huge error.

Vowel Substitutions

Vowels are defined in the LIPP alphabets as having features coded for tongue advancement (front, back, or central), tongue height (high, high lax, mid, mid lax, and low), rounding (round or not round) and nasalization (nasal or not nasal). Vowel substitution errors are captured in the WSSA, with errors in height (0.4 point credit if correct), tongue advancement (0.4 point credit), rounding (0.1 point credit), and nasalization (0.1 point credit) calculated by the analysis program. Again, a correct production of a target vowel results in a score of 1.0 for that vowel; minor deviations from a target (e.g., /i/ → [ɪ]) are weighted less than serious deviations (e.g., /i/ → /a/). Table 2 provides examples of the error weights for vowel substitutions.

Consonant-Vowel or Vowel-Consonant Alternations

There are also calculations devised to capture errors in which vowels are substituted for syllabic or non-syllabic consonants, and vice-versa. These errors are generally weighted heavily, as they involve crossing a major class of sounds (consonants and vowels). The most extreme errors result in a score of 0 for that sound. For example, these errors include producing an obstruent in place of a vowel or vice versa (e.g., t → æ), or a producing a low vowel for a high semivowel (e.g., j→ a). Less severe errors include substituting a vowel for a syllabic liquid (e.g., "*bicycle*", ba□s□k□ → ba□s□ko). Table 3 lists penalties for consonant-vowel alternations. Note that some of these errors are not observed in the speech of children included in the present study (e.g., vowels being produced as obstruents), but to be comprehensive the WSSA has a provision for scoring such productions.

Computing the WSSA Score

Appendix A provides a computational example for the WSSA. Following Oller and Ramsdell (2006), *slots* are positions for (vertical) alignments of phones between the target (adult) form and the child production. For the slots in which there is a target phoneme and a produced phoneme (paired slots), the WSSA computes the *featural agreement* score (indicating how closely the produced sounds match the target sound). Each *paired* slot begins with a value of 1.0 and is reduced to penalize for any consonant or vowel substitutions based on the consonant-consonant, vowel-vowel, or consonant-vowel alternations described above. The average phonetic accuracy of the sounds that are produced is then computed to derive the featural agreement score. Thus, if all of the sounds produced by the child exhibit complete phonetic accuracy when compared to the target (e.g., [su] for /su/), the average feature agreement would be 1.0. If many of the features were in error (e.g., [□u] or [□o] for /su/), a lower featural agreement score is achieved. Note that the mean featural agreement for the child production [u] for /su/ is 1.0, because the phoneme that is produced is phonetically accurate.

Because the featural agreement score captures only correct productions of sounds or sound substitutions, an additional computation is required to capture phoneme deletions and adjunctions (additions). This is done through the *global structural agreement* score, which penalizes for sound deletions and adjunctions (i.e., unfilled slots in either the target form or the child's production) by penalizing totally and assigning a 0 agreement to the slot. The current system weights glottals, glides, and other weak segments half as much (0.5) as slots for strong segments (e.g., orally articulated consonants and vowels, weighted 1.0). This global structural agreement score is then multiplied by the featural agreement score to derive the WSSA (see Appendix A).

Goals of the Study

Having defined scoring within the WSSA, we now describe an initial investigation of its validity and reliability using data from both typically developing children and children with SSDs. A reasonable measure of phonetic accuracy would correlate well with existing measures of phonetic accuracy, be able to distinguish between children with and without SSDs at various ages, be able to account for phonetic growth over time, be reasonably stable across different speech samples, and show strong correlations between transcribers. Throughout the study, we will examine how the WSSA performs in relation to other measures. We use PCC as a standard for comparison and demonstration purposes because PCC is a well-respected and widely-used measure of accuracy that has relatively well-specified reliability and validity (Shriberg, et al., 1997; Shriberg & Kwiatkowski, 1982).

Method

Transcriptions were entered into the LIPP software program, with a broad adult form of the target word(s) on the "target" row and the child's production on the "transcription" row. Adult forms were adjusted as necessary to reflect dialectally acceptable variations of words. For example, reduced vowels in unstressed syllables (e.g., the final syllable of "elephant") are sometimes realized as a schwa [ə] and sometimes realized as a high lax front vowel [ɪ]. The vowel in the target row would be adjusted to be the same as the child's production, so that the WSSA algorithm did not compute a penalty for the vowel in such a case. Other dialectally acceptable variations, such as affrication of /tr/, and /dr/ clusters were allowed (e.g., "tree" [tɹi] as [ʃɹi]).

In aligning the adult target and the child's production, procedures described by Oller and Ramsdell (2006) were followed. The first step was to align the nuclei. Rarely do children add nuclei, but if this occurred, the added nucleus was aligned with other consonants or included in empty (orphan) slot. In instances where consonant clusters were reduced to a single consonant that was not one of the constituents of the sequence, (e.g., /sp/ → [m]), the consonant produced was aligned with the target consonant with which it shared the most features (in this case [m] was aligned with the target /p/).

Speech samples were obtained from children of a variety of ages from several prior studies, including samples of typically speaking (TS) children and those with SSDs. We provide a brief description of each sample, and refer readers to prior work for further details of the participants, recording parameters, etc. Data are included from cross-sectional and longitudinal samples, including: a sample of preschoolers with SSD (Group 1), adolescents with and without SSD (Group 2), preschoolers with and without SSD (Group 3) and typically developing toddlers with longitudinal speech samples (Group 4).

Group 1: Preschoolers with SSDs

Forty-four preschoolers (ages 4;0 to 5;9) from Upstate/Central New York were referred by speech-language pathologists as having a SSD of unknown origin. Participants achieved a standard score on the Goldman-Fristoe Test of Articulation-2 (GFTA-2, Goldman & Fristoe, 2000) below 90, were monolingual speakers of General American English and did not have significant developmental, cognitive or receptive language delays (see Preston & Edwards, 2010, for further description). Most were seen at home, although a small number were seen in a room reserved for child research at Syracuse University. These participants were digitally audio recorded naming 125 pictures chosen to elicit many consonant clusters and multisyllabic words. Responses from all 125 words were later phonetically transcribed into the LIPP software program by the first author (a speech-language pathologist with expertise in SSDs and six years of graduate training in childhood phonological disorders). The WSSA analysis routine was then run on the transcription data to generate a score from 0 to 1. PCC was derived from the same speech sample. Also, children's speech sound errors on the 125-item picture-naming task were coded as to the number of nondevelopmental/atypical phonological processes exhibited in the sample using a measure termed "atypical errors per consonant" (Preston, 2008; Preston & Edwards, 2010). Examples of nondevelopmental sound errors include backing of alveolars to velars, cluster creation, labialization of back sounds, and liquids replacing glides (see also Preston & Edwards, 2010).

Additionally, to demonstrate reliability across listeners, the fourth author (who has over 30 years of experience in phonetic transcription of children's speech) independently transcribed 25 words from each of 41 participants. These were entered into LIPP as well, and a WSSA score was derived. Thus, WSSA scores from 25 words from each of two transcribers were compared.

To test the validity of the WSSA with respect to clinical judgment, the 125-word speech sample audio files from 20 preschoolers were edited and used to obtain judgments of severity. Each of the 125 words was extracted from the digital sound file and paired with a spoken number from 1 to 125. These number-word pairs were then concatenated into a single audio file of approximately five minutes for each child, as follows:

Audio Indicator: *Number one*	Child's production: *parachute*
Audio Indicator: *Number two*	Child's production: *baby carriage*
…	
Audio Indicator: *Number 125*	Child's production: *teacher*

To determine the perceived severity of the child's SSD, the audio samples of the 20 children were rated by 12 ASHA (American Speech-Language-Hearing Association) certified speech-language pathologists (SLPs), all of whom had a Certificate of Clinical Competency (CCC) and clinical experience working with preschoolers. Each SLP, informed only of the child's age and gender, listened to speech samples of five children. They were provided with the orthographic form of each target word that the child was attempting to produce and were instructed only to listen, not to transcribe the child's speech. The SLPs used the following nine-category rating scale to indicate the perceived severity, circling one category per child: Advanced speech (above average, no SSD), Normal (no SSD), Mild SSD, Mild-moderate SSD, Moderate SSD, Moderate-severe SSD, Severe SSD, Very Severe SSD, Profound SSD. Each child's speech sample was rated by three different SLPs, and no two children were ever rated by the same combination of SLPs. The median rating of the three SLPs was used for subsequent analysis of perceived "severity" of the child's SSD.

Group 2: Young Adolescents with and without SSDs Naming Pictures

A group of adolescents (ages 10 to 15) from Upstate/Central New York were recorded naming 64 pictures (see Appendix A of Preston & Edwards, 2007). Fourteen of these adolescents were recruited because they had difficulty producing rhotics (/ɹ, ɝ, ɚ/), and the 19 remaining participants were typically speaking adolescents (TS) with no history of speech-language difficulty. All of these participants were recorded and transcribed together by the first and fourth authors to achieve a consensus (Shriberg, Kwiatkowski, & Hoffmann, 1984) (see Preston & Edwards, 2007, for recording and transcription information). Another five children between 9 and 13 years (one with a lateral lisp and four with errors primarily involving liquids) were recruited for a pilot study of SSDs at Haskins Laboratories and were recorded naming the same 64 pictures. Thus, the sample included 19 young adolescents with known SSDs (11 male) and 19 TS young adolescents (8 male) with no history of speech problems, all between 9 and 15 years of age.

Group 3: Preschoolers with and without SSDs Naming Pictures

As part of a larger study conducted at Syracuse University, 18 male children between the ages of 3;10 and 5;4 participated in an extensive picture-naming task (see Conture, Louko, & Edwards, 1993; Louko, Edwards, & Conture, 1990; Wolk, Edwards, & Conture, 1993).

Ten of the children exhibited SSDs (confirmed by clinicians, and had GFTA percentile <12), and eight had typical phonological development (GFTA percentiles >42). No other communication or developmental problems were noted. Each child was audio and video-recorded while naming 120 colored pictures illustrating familiar objects and actions. The words were selected to elicit all consonant sounds of English at least twice in each word position and in a variety of consonant clusters, as appropriate; many multisyllabic words were also included. The 120 words were made up of two randomized 60-word lists, A and B, each of which provided a representative sample. For approximately half of the children, the pictures were presented in the A-B order, and for the other half, the B-A order was used. All 120-word speech samples were transcribed on-line by the fourth author. The audio-video recordings were later reviewed to refine the transcriptions.

Group 4: Longitudinal Study of Spontaneous Speech of Typically Developing Toddlers

To evaluate the ability of the WSSA to capture longitudinal change in phonological production, transcriptions from a publicly available online dataset from the PhonBank portion of the CHILDES project (MacWhinney, 2000) were used (http://childes.psy.cmu.edu/browser/index.php?url=PhonBank/English-Davis/). This included samples, provided by Dr. Barbara Davis, of typically developing English-speaking children from Texas who were seen approximately twice per month over the course of several months (see Davis & MacNeilage [1995] and Davis, MacNeilage, & Matyear [2002] for further information on data collection and transcription). We selected seven children who had data spanning at least 10 weeks (Aaron, Anthony, Ben, Cameron, Charlotte, Hannah, Rachel). A child's data from a session was included and entered into LIPP for the present study based on the following criteria: the child had at least five samples at different times, and each sample had at least 12 different utterances transcribed in which the "gloss" or adult target was known. Hence, unintelligible words were not used. If a child produced a word or phrase multiple times, we included it only twice. In order to capture production variability, we used the first two renditions that were different (e.g., [bau] and [ba] for /bal/). Up to one hundred utterances per session were entered into LIPP (with a minimum of 12 different utterances). Appendix B lists the PhonBank samples used here.

Finally, data were included from a typically developing English speaking boy, MR, who was recorded during playtimes, book reading, and mealtimes with an Olympus WS-331M digital voice recorder. He was recorded approximately once per month over six months, from the age of 22 to 28 months. Only productions in which the target word was known (13 to 81 utterances per sample) were transcribed by the first author. This participant was included to replicate the results from Group 4 (which included different transcribers).

Results

Because the focus is on demonstrating validity and reliability of the WSSA, fundamental questions about the psychometric properties of the measure are presented and addressed with converging lines of evidence from various samples. Correlations between WSSA scores and other relevant variables are presented. Pearson's *r* is used for samples of 15 or more and for data that meet assumptions of normality and interval scale, and Spearman's ρ is presented for samples of less than 15 or ordinal data.

Do WSSA Scores Correlate with Other Measures of Speech Sound Accuracy?

To validate the WSSA, we examine whether WSSA scores correlate with other measures of phonological severity, including PCC, GFTA-2 raw scores, and SLP's judgments. Cross-sectional data from the 44 preschoolers with SSDs in Group 1 indicated that each child's WSSA score was correlated with other measures of speech sound accuracy including PCC and PCC-R from the same 125-word speech sample (see Table 4). Additionally, WSSA scores from the samples correlated with raw scores (number of sound errors), standard scores, and percentile scores on the GFTA-2. The WSSA was strongly related to a SLP's categorical description of the severity of a child's SSD in 20 of these preschoolers (Spearman's $\rho = -0.882$ $p < 0.001$).

To evaluate whether the WSSA captures non-developmental speech sound errors better than other measures do, we also examined the Pearson's r correlation of the WSSA score and a novel measure termed "atypical errors per consonant" (Preston, 2008; Preston & Edwards, 2010). As can be seen in Table 4, the WSSA is more strongly correlated with this measure than is any other index of speech sound accuracy, indicating that the WSSA is more sensitive than other measures to atypical speech errors. Group 2 included 38 adolescents ages 9 to 15 with and without SSDs naming 64 pictures. The correlation between WSSA scores and PCC was high for the entire sample ($r = 0.94$, $p < 0.001$, 95% CI 0.89-0.97), demonstrating strong agreement across the broad range of typical and disordered speech. However, when considering only the 19 participants with SSDs, the correlation was lower ($r = 0.66$, $p = 0.002$, 95% CI 0.26-0.96). WSSA and PCC scores were also highly correlated in Group 3 ($r = 0.90$, $p < 0.001$, 95% CI 0.75-0.96 among all preschoolers; $r = 0.714$, $p = 0.001$ among the 10 preschoolers with SSD), in addition WSSA scores were correlated with GFTA percentile rank ($\rho = 0.917$, $p < 0.001$). In Group 4, the youngest cohort, WSSA scores showed moderate correlations with PCC ($r = 0.69$, $p < 0.001$, 95% CI 0.54 - 0.79 for the 68 speech samples).

Do WSSA Scores Distinguish Children with and without SSD?

To test whether WSSA scores distinguish children with and without SSDs, Group 2 (young adolescents) were compared with an Analysis of Variance (ANOVA), using speech group (SSD and TS) and gender as factors. There was no interaction between group and gender (F [1, 34] = 0.235, $p = 0.631$, $\eta^2_p = 0.007$), and the main effect of gender was not significant (F [1, 34] = 0.003, $p = 0.954$, $\eta^2_p = 0.000$). In contrast, the main effect of group was statistically significant (F [1, 34] = 58.2, $p < 0.001$, $\eta^2_p = 0.631$), demonstrating that the adolescents who were TS and SSD differed reliably on their WSSA scores. Moreover, Figure 1 demonstrates that there is no overlap between adolescents with SSD and those with TS. WSSA scores less than or equal to 0.966 (below the dotted line) were associated with SSD, while scores greater than 0.966 (above the dotted line) were associated with TS. (Note, however, that one TS participant's WSSA score was 0.967 and one SSD participant's score was 0.966).

Similarly, Group 3 was used to evaluate group separation on WSSA scores in preschool boys with and without SSD. Figure 2 represents WSSA scores for both groups. Speech samples were split into lists A and B, each representing 60 of the 120 target words, so each child has two vertically adjacent (often overlapping) symbols representing his scores on the two 60-word samples. It is clear that for most preschoolers the two sets of words yield similar WSSA scores. One child with SSD completed only the first 60 words, and therefore is included only once. There is slight overlap between the groups; a cutoff score of less than 0.910 (dotted line) would categorize all of the children with SSD correctly. However, this score would incorrectly categorize one child with TS as having a SSD. His WSSA scores were 0.891 and 0.892 from lists A and B, respectively. Arguably, because this child did demonstrate some unusual speech sound patterns (e.g., nasalization), it is possible that his speech might not be developing normally. Statistical evaluation of group differences were conducted using a mixed model ANOVA, testing group (SSD versus TS) as a fixed effect and speech sample (lists A and B) as a random effect. The main effect of group was highly significant (F [1, 28.89] = 69.0, $p < 0.001$), suggesting that these preschoolers with and without SSD differed in their WSSA scores.

Are WSSA Scores Consistent across Different Samples for the Same Child?

From the 44 participants in Group 1, 25 words were randomly chosen from the 125 word sample. WSSA scores derived from the 25-word sub-samples correlated highly with the child's WSSA score from the 125 words ($r = 0.931$, $p < 0.001$, 95% CI 0.867.-0.962). Using PCC derived from the 25-word sub-sample, correlation with the child's overall PCC from the 125 words was $r = 0.877$ ($p < 0.001$, 95% CI 0.784.-0.931). Thus, the WSSA score from a smaller sample of 25 words was strongly associated with scores from the 125-word sample.

From 17 children in Group 3, Pearson's correlation between WSSA scores derived from word list A and word list B (each containing 60 words) was $r = 0.990$ ($p < 0.001$, 95% CI 0.972 - 0.996). There was no significant difference between WSSA scores from Forms A and B (t [16] = 1.16, $p = 0.262$). From those same samples, correlation between PCC scores derived from lists A and B were $r = 0.973$ ($p < 0.001$, 95% CI 0.93-0.99).

Is there Between-Transcriber Agreement on WSSA Scores?

From Group 1 (preschoolers with SSD), both PCC and WSSA scores were calculated for the 125-word speech samples. A second transcriber (the fourth author) transcribed 25 consecutive words from 41 preschoolers from the group, with starting points randomly determined for each child. Correlations between the WSSA scores derived from these transcriptions of two listeners for 41 children was $r = 0.925$ ($p < 0.001$, 95% CI 0.86 – 0.96). Pearson's correlations between the first and fourth author's PCC scores for those same samples was $r = 0.854$ ($p < 0.001$, 95% CI 0.74-0.92). Thus, transcriptions of the same 25-word speech samples yielded high between-transcriber agreement for both measures.

From Group 2 (young adolescents), the second author (a speech-language pathologist with several years of experience in phonetic transcription) transcribed the 64 words from 21 participants (12 TS). WSSA scores derived from those transcriptions were highly correlated with WSSA scores from transcriptions completed together by the first and fourth authors ($r = 0.930$, $p < 0.001$, 95% CI 0.83 - 0.97). PCC scores from the second author's transcriptions of those 21 participants correlated highly with PCC scores derived from the transcriptions completed by the first and fourth authors ($r = 0.932$, $p < 0.001$, 95% CI 0.84 - 0.97).

Are WSSA Scores Sensitive to Growth in Children's Phonetic Accuracy?

To determine if the WSSA score captures growth in a child's phonological development, longitudinal data from Group 4 were used. The expectation was that children would demonstrate increasingly accurate speech as they get older; thus, there should be positive associations between speech sound accuracy and age. Table 5 presents rank-order correlation coefficients (Spearman's ρ) for each of the seven participants from the PhonBank dataset, plus the case study MR. In all cases, the correlation coefficient between age and WSSA was positive. However, for three of the eight children the coefficient between age and PCC was negative, suggesting that (for these three children) the overall trend was for PCC scores to decrease with age, although the WSSA score was found to increase with age. Additionally, WSSA scores showed a stronger association with age than did PCC scores in six of eight children (Ben and Cameron being the exceptions). Given the diversity of age ranges and small number of samples, these results should be viewed with caution. However, these results provide initial support for the notion that the WSSA scores might be more sensitive to small phonetic improvements in speech production, and thus more strongly associated with phonological development than are PCC scores.

To illustrate the difference between WSSA and PCC, Figure 3 displays WSSA and PCC scores plotted against age for participant MR. In the top panel, it is evident that age accounts for a significant proportion of variance in phonetic accuracy (as defined by the WSSA scores), as would be expected. However, PCC scores (lower panel) show a slight decrease in accuracy over time. Thus, from the same transcriptions, phonetic accuracy is found to increase between 22 and 28 months for this child when using the WSSA score as a measure, but this is not the case when using PCC as a measure of phonetic development.

DISCUSSION

We have provided justification and empirical evidence for implementing a weighted measure of speech sound accuracy to quantify phonetic accuracy in toddlers, preschoolers and adolescents, including children with SSDs. The measure was found to have reasonable psychometric properties, properties that were generally similar to those of PCC. Based on both longitudinal and cross-sectional data, we have demonstrated that the WSSA is sensitive to phonological development and to disorders of phonological production. The external validity of the measure is supported by empirical evidence presented here from multiple speakers, speech samples, recording environments, recording equipment, and transcribers.

The WSSA scores correlate with several measures of speech sound accuracy, providing additional support for its validity as a measure of the accuracy of phonetic production. The moderate correlations with raw and standard scores from the GFTA-2 are expected, given that this instrument scores articulation accuracy based on only a single occurrence of each consonant or cluster in a given word position. Similar to other standardized articulation measures, scoring for the GFTA-2 does not take into account the nature of the errors or the degree of accuracy. WSSA scores also correlate with speech-language pathologists’ (SLPs’) judgments of severity. If SLP judgments of severity of a SSD are a standard by which other more objective metrics should be compared (Flipsen, et al., 2005), then WSSA and PCC-R appear to be relatively similar in capturing severity.

WSSA scores were found to correlate strongly between transcribers. In Group 1, PCC derived from the same speech samples by two independent transcribers were found to correlate somewhat less well than WSSA scores derived from those same samples (although the 95% CI overlap). Slightly better correlations between transcribers for WSSA than for PCC is probably due to the fact that WSSA scores assign values to the accuracy of the features of segments, rather than scoring entire segments as correct or incorrect, as in PCC. Because transcribers may agree on most of the features they hear, a measurement system that is based on features might be more robust and sensitive to small disagreements than a system based on binary right/wrong judgments of the accuracy of a sound. However, reliability data from Group 2 suggest nearly identical correlations between transcribers for WSSA and PCC scores.

For the majority of children in the longitudinal cohort (75%), the WSSA scores were more strongly associated with age than were PCC scores (Table 5). If phonetic development advances with age, it appears that intermediate steps in phonetic accuracy made by children as they come closer to target (adult) forms might not be adequately captured by considering only binary (right/wrong) accuracy. In young children, PCC penalizes for even minor errors, whereas WSSA scores can account for improvements in how well the word is represented. Indeed, it is surprising that not all children showed stronger correlation coefficients between age and WSSA scores than between age and PCC scores. Given that conversational samples were used to examine speech sound accuracy in the longitudinal samples, there is no consistency across samples or across children. It is possible that repeated assessments of the same corpus would be more sensitive to growth. It is also possible that modifications in the WSSA algorithm might improve its ability to capture growth in phonetic accuracy.

The present project demonstrates a general measure of phonetic accuracy and provides evidence validating its use. We have demonstrated that the WSSA is predictable and reliable, and that there are several possibilities for implementing this measure, both clinically and in research. However, we recognize that, although the quantification of errors used here is based on rank-orders derived from developmental phonological principles, the actual numeric value assigned to the errors is arbitrary. Because there are many degrees of freedom in speech production that could be captured, other researchers might reach different conclusions about how particular errors should be weighted. Also, because the WSSA is based on a linear (sound-by-sound) analysis, it does not take into account errors that might be due to assimilations, nor does it differentiate errors by word position (with the exception of voicing errors). It is important to point out that the WSSA is not intended to encompass every nuance in speech development or in speech errors. Fortunately, the weights applied to each type of error can be adjusted in the LIPP program if a researcher judges there is sufficient theoretical, empirical, or experimental reason to do so. Future studies could use different weighting systems, depending on the needs of the research. Additionally, the WSSA could be used with larger prospective longitudinal studies to evaluate the range of performance among typically speaking children, or it might be used as a tool to track progress in treatment for children with SSDs. In addition, the WSSA could be adapted for speakers of other languages in which developmental phonological patterns differ, or expanded to capture prosodic factors such as lexical stress. The development of a sensitive, reliable, and clinically-viable measure of phonetic accuracy should be considered an ongoing process.

In conclusion, psychometric properties of the standard measures used for quantification of speech sound accuracy can and should be investigated. We view the current WSSA to be an improvement over prior measures of speech sound production accuracy because it is more sensitive to the types of errors produced by children. We recognize that no single measure is likely to completely capture every type of error and to weight errors in such a way that there would be universal agreement. Nonetheless, the WSSA, illustrated in the present study, represents a step toward more accurate and reliable measures of speech sound accuracy.

Acknowledgments

This work was supported by NICHD grant T32HD7548-8 to Haskins Laboratories (C. Fowler, PI), and an American Speech-Language-Hearing Foundation Early Childhood Language Development Grant awarded to the first author (Group 1 data). Collection of Group 2 data was funded by NIDCD research grant DC00523-01A2 to Syracuse University. Participation of authors Ramsdell and Oller was supported by NIDCD research grant DCR01 006099 (D. K. Oller, PI). Thanks to Barbara Davis, as well as the developers of the CHILDES database.

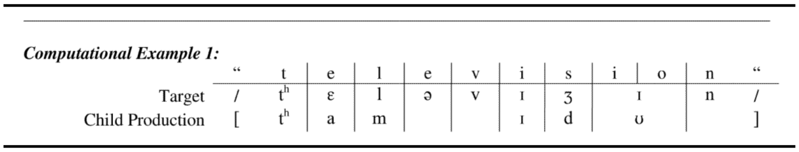
Appendix A: Calculation of the Weighted Speech Sound Accuracy

Definitions

Global structural agreement: the proportion of segment slots (represented by columns below) in the aligned utterances in which both transcriptions (the target and child production) include a segment.

Featural agreement: the proportion of phonetic information shared in segments that are present in the same slot in the two transcriptions (the target and child production).

WSSA: the global structural agreement multiplied by the featural agreement.



Computational Example 1:

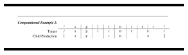
	“	t	e	l	e	v	i	s	i	o	n	“
Target	/	t^h	ɛ	l	ə	v	ɪ	ʒ	ɪ		n	/
Child Production	[	t^h	a	m			ɪ	d	ʊ			]

Global structural agreement: There are 9 slots with 6 shared (the child omitted 3 of the target sounds), so the global structural agreement is (6/9) = 0.667.

Featural agreement:

- t → t, no deduction, featural agreement = 1
- ε→ a, mid front lax vowel → low central vowel, teeny height deduction of 0.1 and small front deduction of 0.2, featural agreement = 0.7
- l → m, semivowel → nasal, huge manner deduction of 0.333 and big place deduction of 0.25, featural agreement = 0.417
- ɪ → ɪ, no deduction, featural agreement = 1
- ʒ → d, palatoalveolar fricative → alveolar plosive, small manner deduction of 0.166 and small place deduction of 0.166, featural agreement = 0.668
- ɪ → ʊ, high front lax vowel → high back lax vowel, big front deduction of 0.4 and small rounding deduction of 0.1, featural agreement = 0.5
- The mean featural agreement for all of the paired slots represents the featural agreement. The mean of the featural values for the 6 paired slots is [(1+0.7+0.417+1+0.668+0.5)/6] =0.714

WSSA: 0.667 * 0.714 = 0.476.



Computational Example 2:

	“	s	p	l	i	n	t	e	r	“
Target	/	s	p	l	ɪ	n	t	ɚ		/
Child Production	[	s	p		ɪ	n		ɔ		]

Global structural agreement: There are 7 slots with 5 shared (the child omitted 2 of the target sounds), so the global structural agreement is (5/7)= 0.714.

Featural agreement:

- s → s, no deduction, featural agreement = 1
- p → p, no deduction, featural agreement = 1
- ɪ → ɪ, no deduction, featural agreement = 1
- n → n, no deduction, featural agreement = 1
- ɚ → ɔ, mid central vowel with r-color → mid back lax vowel, small front deduction of 0.2, teeny height deduction of 0.1, featural agreement =0.7
- The mean featural agreement for all of the paired slots is [(1+1+1+1+0.8)/5].=0.94

WSSA: 0.714 * 0.94 = 0.685.

Appendix B: PhonBank Files included in Group 4

Aaron: Aaron05, Aaron06, Aaron07, Aaron08, Aaron09, Aaron10, Aaron11

Anthony: Ant03, Ant04, Ant08, Ant09, Ant10, Ant11, Ant12, Ant13

Ben: Ben21, Ben22, Ben23, Ben24, Ben25, Ben33

Cameron: Cam40, Cam41, Cam42, Cam43, Cam44, Cam45, Cam46, Cam47, Cam48, Cam49, Cam50, Cam51, Cam52

Charlotte: Cha44, Cha45, Cha46, Cha47, Cha48, Cha49

Hannah: Han18, Han19, Han20, Han21, Han22, Han23, Han24, Han25, Han26, Han27, Han28, Han29

Kate: Kate04, Kate05, Kate06, Kate07, Kate08, Kate09, Kate10

Rachel: Rac32, Rac33,Rac34, Rac35, Rac36, Rac37, Rac38, Rac39, Rac40, Rac41,Rac42

REFERENCES

Conture EG, Louko LJ, Edwards ML. Simultaneously treating stuttering and disordered phonology in children: Experimental treatment, preliminary findings. American Journal of Speech-Language Pathology 1993;3:72–81.

Crary MA. Phonological characteristics of developmental verbal dyspraxia. Seminars in Speech and Language 1984;5(2):71–83.

Davis BL, MacNeilage PF. The articulatory basis of babbling. Journal of Speech & Hearing Research 1995;38(6):1199. [PubMed: 8747814]

Davis BL, MacNeilage PF, Matyear CL. Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. Phonetica 2002;59(2-3):75–107. [PubMed: 12232462]

Dodd, B., editor. Differential diagnosis and treatment of children with speech disorder. 2 ed.. Whurr Publishers; Philadelphia: 2005.

Dodd B, Iacano T. Phonological disorders in children: Changes in phonological process use during treatment. British Journal of Disorders of Communication 1989;24:333–351. [PubMed: 2627548]

Edwards, ML.; Shriberg, LD. Phonology: Applications in communicative disorders. College-Hill Press; San Diego, CA: 1983.

Flipsen P Jr. Hammer JB, Yost KM. Measuring severity of involvement in speech delay: Segmental and whole-word measures. American Journal of Speech-Language Pathology 2005;14(4):298–312. [PubMed: 16396613]

Goldman, R.; Fristoe, M. Goldman Fristoe Test of Articulation. Second Ed.. AGS; Circle Pines, MN: 2000.

Kazdin, AE. Research Design in Clinical Psychology. Allyn and Bacon; Boston: 2003.

Leonard LB. Unusual and subtle phonological behavior in the speech of phonologically disordered children. Journal of Speech and Hearing Disorders 1985;50(1):4–13. [PubMed: 3974210]

Louko LJ, Edwards ML, Conture EG. Phonological characteristics of young stutterers and their normally fluent peers: Preliminary observations. Journal of Fluency Disorders 1990;15(4):191–210.

MacWhinney, B., editor. The CHILDES project: Toos for analyzing talk. Third ed.. Lawrence Erlbaum Associates; Mawhah, NJ: 2000.

Morris SR. Test-Retest Reliability of Independent Measures of Phonology in the Assessment of Toddlers’ Speech. Language, Speech, and Hearing Services in Schools 2009;40(1):46–52.

Oller, DK.; Delgado, RE. *Logical International Phonetic Programs* [Computer program: Windows Version]. Intelligent Hearing Systems; Miami, FL: 1999.

Oller DK, Ramsdell HL. A weighted reliability measure for phonetic transcription. Journal of Speech, Language, & Hearing Research 2006;49(6):1391–1411.

Pollock K. The identification of vowel errors using traditional articulation or phonological process test stimuli. Language, Speech and Hearing Services in Schools 1991;22:39–50.

Preston, JL. Dissertation Abstracts International. Vol. 69. 2008. Phonological processing and speech production in preschoolers with speech sound disorders (Doctoral dissertation, Syracuse University, 2008); p. 10

Preston JL, Edwards ML. Phonological processing skills of adolescents with residual speech sound errors. Language, Speech and Hearing Services in Schools 2007;38:297–308.

Preston JL, Edwards ML. Phonological awareness and speech error types in preschoolers with speech sound disorders. Journal of Speech, Language & Hearing Research 2010;53

Rvachew S, Chiang P-Y, Evans N. Characteristics of speech errors produced by children with and without delayed phonological awareness skills. Language, Speech and Hearing Services in Schools 2007;38(1):60–71.

Shriberg LD, Austin D, Lewis BA, McSweeny JL, Wilson DL. The percentage of consonants correct (PCC) metric: Extensions and reliability data. Journal of Speech, Language, and Hearing Research 1997;40:708–722.

Shriberg LD, Kwiatkowski J. Phonological disorders III: A procedure for assessing severity of involvement. Journal of Speech and Hearing Disorders 1982;47:242–256. [PubMed: 7186560]

Shriberg LD, Kwiatkowski J, Hoffmann K. A procedure for phonetic transcriptions by consensus. Journal of Speech & Hearing Research 1984;27(3):456–465. [PubMed: 6482415]

Shriberg LD, Lof GL. Reliability studies in broad and narrow phonetic transcription. Clinical Linguistics & Phonetics 1991;5(3):225–279.

Wolk L, Edwards ML, Conture EG. Coexistence of stuttering and disordered phonology in young children. Journal of Speech and Hearing Research 1993;36:907–917.

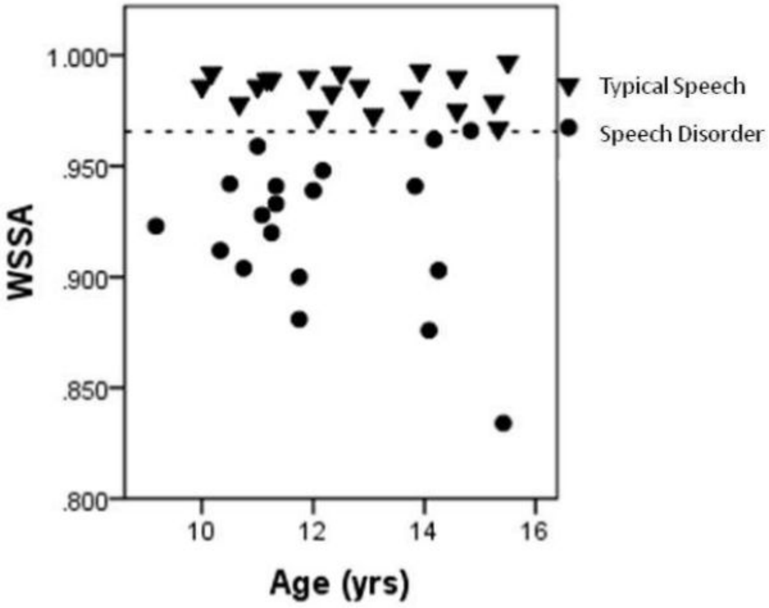


Figure 1.

WSSA scores from a 64-word picture-naming sample plotted against age for 19 adolescents with SSD (circles) and 19 typically speaking adolescents (triangles). The dotted line represents a WSSA score of 0.966. All 19 adolescents with SSD scored at or below this value, and all TS adolescents scored above this value.

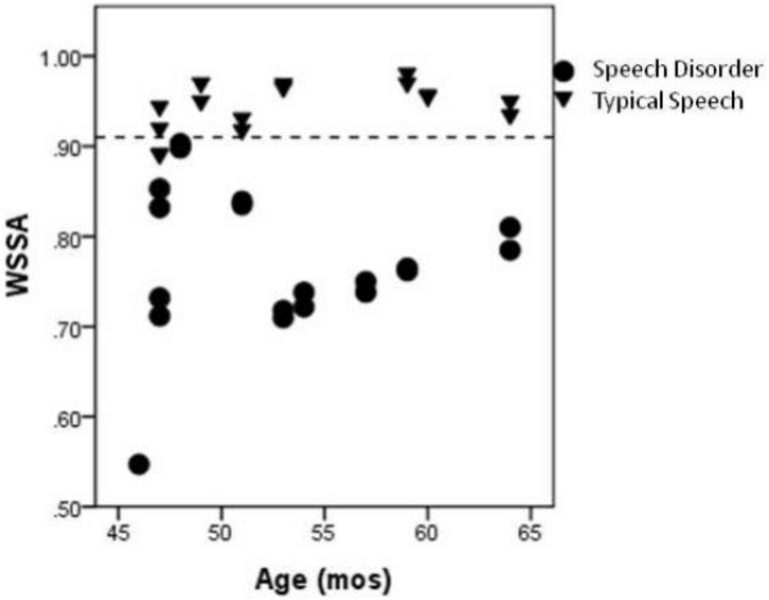


Figure 2.

WSSA scores plotted against age for 9 preschool boys with SSD (circles) and 9 typically speaking preschoolers (triangles). Vertically adjacent symbols represent two speech samples from each child (list A and list B). The dotted line represents a WSSA score of 0.910, below which all 9 children with SSD scored, and above which 8 of 9 typically speaking preschoolers scored.

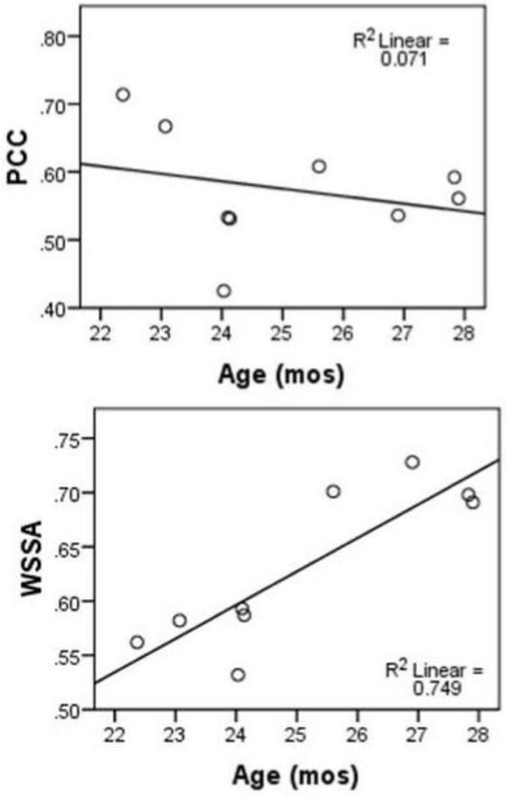


Figure 3.

WSSA and PCC scores derived from the same transcriptions of the speech of a typically developing child, MR. WSSA scores gradually increase with age, but PCC scores do not show this trend. The solid line represents the linear regression derived from these samples, and R^2 represents the proportion of variance in phonetic development that can be accounted for by age.

Table 1

Consonant features and penalties for errors

CONSONANT FEATURE (Weight)		Penalties	Examples
Manner (0.333)	**Huge Manner** -uncommon errors, damaging to intelligibility	−.3333	Plosive → Fric. or Affric. / #____ Glide → Liquid Nasal → Non-Nasal Semivowel → Nasal Sonorant → Obstruent
	Big Manner - Less common in phonological development	−.25	Plosive → Fric. or Affric. / C or V____ Fric. or Affric. Lateral Fric. or Affric. →
	Small Manner -Common errors in phonological development	−.1666	Fric. or Affric. → Plosive Fricative ↔ Affricate Liquid → Glide or Tap
	Teeny Manner -minor phonetic errors	−.0833	Nonspecific distortion
Place (0.333)	**Huge Place:** -Uncommon, very damaging to intelligibility	−.333	Dorsal → Labial Glottal → Non-Glottal[*]
	Big Place - Less common in phonological development	−.25	Coronal → Labial Coronal → Dorsal Alveolar → Palatal Palatal → Dental Retroflex → Not Retroflex
	Small Place - Typical errors in phonological development	−.1666	Linguadental → Labiodental Dental → Alveolar Palatal → Alveolar Dorsal → Coronal
	Teeny Place -Phonetic errors in English, based on small changes in tongue placement.	−.0833	Bilabial → Labiodental Lips not spread → Lips spread Lips not round → Lips round Labialization Blading Tongue Advance/Retract
Voicing (0.333)	**Huge Voicing** -Uncommon	−.3333	Word-Initial or Medial Devoicing Word-Final Voicing
	Small Voicing -Common	−.2222	Word-Final Devoicing Word-Initial Voicing
	Teeny Voicing -Phonetic changes	−.1111	Aspiration of nonaspirated C (e.g., ste → st^{h}e)

[*] (exception for /t/ → [ʔ] /___# due to English dialect)

Table 2

Vowel feature weights and penalties for errors

Vowel Feature	Weight	Penalties			Example
Height	(0.40)	Huge Height	−.40	4 step height change	/i/ ↔ [a]
		Big Height	−.30	3 step height change	/ɪ/ ↔ [a]
		Small Height	−.20	2 step height change	/i/ ↔ [e]
		Teeny Height	−.1	1 step height change	/a/ ↔ [ε]
Advancement	(0.40)	Big Front	−.40	Front ↔Back	/o/ ↔ [e]
		Small Front	−.20	Front↔Central or Back↔Central	/i/ ↔ [⍰]
Nasalization	(0.1)	Small Nasal	−.10	Not Nasal → Nasal	/a/ → [ã]
Rounding	(0.1)	Small Rounding	−.10	Round ↔ Not Round	/Λ/ ↔ [⍰]

Table 3

Penalties for consonant-vowel alternations

Alternations	Penalty	Example
Vowels and Syllabic Consonant Alternation		
Vowel (non-glide) ↔ Syllabic Liquid	−.25	
Vowel (non-glide) ↔ Syllabic Nasal, tap, trill	−.5	
Vowel (non-glide) ↔ Syllabic voiced obstruent	−.75	
Vowel (non-glide) ↔ Syllabic unvoiced obstruent	−1.0	
Vowels and Non-syllabic Consonant Alternation		
Liquid ↔ Any vowel (non-glide)	−.5	
High semivowel ↔ High vowel		
Mid semivowel ↔ Mid vowel		
Low semivowel ↔ Low vowel		
High or low vowel ↔ Mid semivowel	−.75	
High vowel ↔ Nonsyllabic nasal, tap, trill		
Low semivowel ↔ High vowel	−1.0	
High semivowel ↔ Low vowel		
Obstruent ↔ Vowel		

Table 4

Correlations between WSSA scores and other measures of speech sound accuracy from 44 children with SSDs

	PCC	PCC-R	GFTA-2 Raw Score	GFTA-2 Std Score	GFTA2 Percentile	Atypical Errors Per Consonant	Median SLP Severity Rating[†]
WSSA	.85	.91	−.78	.74	.60	−.68	−.88
PCC		.95	−.89	.82	.74	−.54	−.82
PCC-R			.82	.82	.72	−.61	−.89
GFTA-2 Raw Score				−.90	−.78	.54	−.74
GFTA-2 Standard Score					.88	−.49	−.65
GFTA-2 Percentile						−.43	−.54
Atypical Errors per Consonant							.71

Notes: All correlations are significant at $p < 0.02$. WSSA, PCC, PCC-R, and Atypical Errors per Consonant all derived from the same 125 word picture naming task.

[†] Nonparametric correlations (Spearman’s rho) are provided for the SLP severity rating because of the small sample ($n = 20$) and the categorical nature of the data. All other correlations based on Pearson’s r with $n = 44$.

Table 5

Nonparametric correlations between age and two measures of speech sound accuracy

Child	Age range (mos.)	Num. Sessions	Age-WSSA Corr.	Age-PCC Corr.
Aaron	22.40-25.10	7	.64	.14
Anthony	21.13-33.88	7	.07	−.41
Ben	21.01-28.07	6	.200	.66
Cameron	22.23-35.80	12	.58*	.60*
Charlotte	31.80-35.00	6	.74	−.030
Hannah	23.00-27.23	9	.87**	.07
Rachel	17.13-22.10	11	.21	.07
MR	22.37-26.9	7	.78*	−.2

Notes: Participants drawn from the PhonBank dataset, with the exception of MR. Correlation coefficients are nonparametric (Spearman's ρ).

* p <0.05

** p <0.01