Developing a Weighted Measure of Speech Sound Accuracy

Jonathan L. Preston
Haskins Laboratories, New Haven, CT

Heather L. Ramsdell
The University of Memphis, Memphis, TN

D. Kimbrough Oller
The University of Memphis

Mary Louise Edwards
Syracuse University, Syracuse, NY

Stephen J. Tobin
University of Connecticut, Storrs, CT

Purpose: 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, the authors describe a system for differentially weighting speech sound errors on the basis of various levels of phonetic accuracy using a Weighted Speech Sound Accuracy (WSSA) score. The authors then evaluate the reliability and validity of this measure.
Method: Phonetic transcriptions were 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 was validated against existing measures, was used to discriminate typical and disordered speech production, and was evaluated to examine sensitivity to changes in phonetic accuracy over time. Reliability between transcribers and consistency of scores among different word sets and testing points are compared.
Results: Initial psychometric data indicate that WSSA scores correlate with other measures of phonetic accuracy as well as listeners' judgments of the severity of a child's speech disorder. The measure separates children with and without speech sound disorders and captures growth in phonetic accuracy in toddlers' speech over time. The measure correlates highly across transcribers, word lists, and testing points.
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 are compared with normative data. Therefore, it is critical to use measures that are reliable and meaningful and that have known psychometric properties. 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). Many existing measures lack sufficient investigation of reliability and validity or are not sensitive to the nature of the error types that children produce. Consequently, the current study presents a description of, and preliminary psychometric data on, a weighted measure of speech sound accuracy.

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 that many of the commonly used measures are unreliable between different testing occasions (Morris, 2009). This finding 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, more efficient experimental or observational studies can be conducted if the error term is not increased by use of imprecise 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 initial validity and reliability of a relational measure of productive phonology (i.e., one that compares children's productions with the corresponding adult forms).

When a child's productions are inaccurate, it may be helpful to quantify the types of sound errors made in order to be maximally sensitive to mispronunciations of words. 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 do not differentially weight sound errors by error type. Consider the adult target "soap" /sop/. Multiple mispronunciations of this word are possible, including [top], [dop], [ɡop], or [op]. The differences among these errors cannot be captured simply by tallying them because they all involve just one consonant error. For example, [top] involves a change in one feature (manner of articulation); [dop] involves a change in both manner of articulation and voicing; [ɡop] involves an error in manner and voicing, along with an unusual pattern of backing an alveolar to a velar. Analyses that include only correct/incorrect judgments of accuracy would count all of these errors the same. However, it is plausible (and, arguably, preferable) to rank these errors (from most to least accurate: [top], [dop], [ɡop], [op]) and to count them differently. In the present study, we developed the *weighted speech sound accuracy measure* with the purpose of assigning numeric values to each of these error types.

Atypical sound errors (i.e., those that do not commonly occur in typically developing children) may be especially important to capture in a measure of phonetic accuracy, as they may suggest that children have arrived at unusual solutions to satisfy the constraints of their phonological systems (Leonard, 1985). Some unusual errors in consonant production are likely to have a significant impact on intelligibility. For example, Hodson and Paden (1981) reported that unintelligible children produced patterns of backing (e.g., *tea* /ti/ → [ki]) and glottal replacement (e.g., *cup* /kʌp/ → [kʌʔ] or *soap* /sop/ → [hop]); similarly, Dodd and Iacano (1989) also considered these patterns to have a large impact on intelligibility. Other examples of atypical errors include denasalization of nasal consonants, labialization of velars, or production of obstruents for liquids (e.g., Edwards & Shriberg, 1983; Rvachew, Chiang, & Evans, 2007). Such errors have also been found to relate to early phonological awareness skills (Preston & Edwards, 2010; Rvachew et al., 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 of typical and atypical acquisition) would be of value.

Syllable structure changes are also important to consider. For example, Klein and Flint (2006) reported that deletion of final consonants had a greater impact on intelligibility than did stopping of fricatives (a change in manner of articulation) or velar fronting (a change in place of articulation). Similarly, Hodson and Paden (1981) compared 3- to 8-year old intelligible and unintelligible children and observed more consonant deletions and syllable deletions in the speech of unintelligible children. Dodd and Iacano (1989) considered not only consonant deletions but also consonant additions (intrusions) to have a large impact on intelligibility. Thus, when quantifying phonetic accuracy, changes in syllable structure should be weighted heavily.

Additionally, vowels often are not included in the calculation of speech errors in standardized tests or other measures of phonetic accuracy. However, researchers have found vowels to be in error in many children with speech sound disorders (SSDs; Pollock, 1991), including those with childhood apraxia of speech (Crary, 1984). Therefore, vowel errors should 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 errors that are closer to and farther from the target.

Flipsen et al. (2005) reviewed several relational measures thought to index the severity of SSDs. They suggested that speech-language pathologists' (SLPs') 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 the severity ratings. In addition, not all of the measures they describe take into account the degree of phonetic accuracy, and many 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. Researchers have used this scheme as a measure of severity of involvement of speech impairment in many studies. Variants of PCC have been described to overcome some of the limitations of the original PCC, including *PCC—Adjusted* (PCC–A, which does not count common clinical distortions as errors), *PCC—Revised* (PCC–R, which does not count any distortions as errors), and *percent phonemes correct* (PPC, which includes both vowels and consonants in the calculation). As reported by Shriberg et al. (1997), each of the variants of PCC might be appropriate for different clinical or research purposes, but all are intended to be objective measures akin to severity. However, all of these measures are based on examiners' binary (right/wrong) judgments of speech sounds, and none of them permit the assignment of different weights to different types of errors.

Other measures for quantifying phonetic accuracy exist, such as the *phonological mean length of utterance* (PMLU), which is the number of consonants and vowels produced (regardless of accuracy) plus the number of correct consonants, divided by the number of words (Ingram, 2002; Ingram & Ingram, 2001). The PMLU is highly sensitive to the phonological structure of the words sampled (children get credit for attempting longer words) and, therefore, might be easily influenced by a child's expressive vocabulary. Although it quantifies the complexity of the words that children attempt and the adherence to syllable structure, PMLU fails to account for the nature of the consonant errors produced (it relies on right/wrong judgments), and vowel substitutions are not penalized. To date, it does not appear that there has been rigorous psychometric study of the validity and reliability of this measure.

Similarly, Jakielski (1998) developed the *index of phonetic complexity* (IPC) designed to capture the number of complex features (e.g., dorsals, liquids, affricates), word shapes (e.g., final consonants, clusters) and word characteristics (e.g., variation of place of articulation) in each word. This is an independent measure that does not relate the child's production to a target form. Published psychometric data on this measure are limited, although Morris (2009) reported marginally significant test–retest correlations for the IPC in 10 toddlers.

In the present study, we focus on the development of a measure that addresses some of the limitations of currently available transcription-based measures. We focus on preliminary investigation of the psychometric properties of the measure, including validity and reliability.

Validity

Validity can be defined as the extent to which a measure assesses the domain of interest (Kazdin, 2003). Thus, a measure of phonetic accuracy should be carefully constructed on the basis of theoretically relevant and meaningful principles (i.e., *face validity*). 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. To demonstrate *concurrent validity,* the measure should capture the relative degree of impairment, as evidenced by correlations with standardized tests, listener judgments (cf. Flipsen et al., 2005), and other types of existing transcription-based analyses (e.g., PCC). Given the importance of considering the types of errors, as discussed previously, we also believe that a valid measure should correlate with an index of atypical speech errors.

Content validity would be supported by demonstrating that the measure is sensitive to phonological development by capturing small improvements in speech production under the assumption that young children become more phonetically accurate as they get older. In addition, it is also important to show that the measure can differentiate between children with and without SSDs.

Reliability

Reliability can be defined as the extent to which a measure consistently assesses the characteristics of interest (Kazdin, 2003). One issue that frequently arises when using transcription-based measures is *interrater reliability.* Point-by-point reliability between transcribers is often in the 90%–95% range for broad transcription of consonants and approximately 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 they 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 quantifies accuracy of features of the production, the reliability score can be more precise (i.e., both transcribers perceive an alveolar stop followed by a mid-back, rounded vowel).

Another type of reliability is *test–retest reliability,* which is associated with the stability of the measure. It addresses the ability of the measure to capture the construct (i.e., phonetic accuracy) during repeated administrations at different points close in time. Similarly, *alternate form reliability* can be addressed by evaluating whether phonetic accuracy scores derived from one sample of words correlate with accuracy scores derived from a different sample of words produced by the same child on the same occasion. Additionally, we address whether a small sample of a child's speech correlates with a larger, and presumably more stable, sample from the child, which might be thought of as a form of internal consistency.

Goals of the Study

In the present study, we investigated psychometric properties of the new measure that we call the *Weighted Speech Sound Accuracy (WSSA) measure*. In this article, we define the WSSA and present an initial investigation of its validity and reliability using data from children with typical speech (TS) and children with SSDs, sampling a variety of age ranges. A valid measure of phonetic accuracy should be sensitive to conditions in which phonetic accuracy is known to vary and might, therefore, be assessed by determining if values derived from the measure (a) correlate well with existing measures of phonetic accuracy, (b) distinguish between children with and without SSDs at various ages, and (c) are sensitive to growth in phonetic development over time. A reliable measure of phonetic accuracy would (a) show strong correlations between transcribers (interrater reliability), (b) show little change from one sample to another sample that is close in time (test–retest reliability), (c) be reasonably stable across different speech samples (alternate form reliability), and (d) show a strong association between small samples and larger samples from the same speaker (internal consistency). Throughout this article, we 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

In this section, we outline the properties of the measure, define how it is calculated, and describe four groups of children whose speech samples we used to evaluate validity and reliability.

Description of the WSSA Measure

The WSSA rules are based on the match between segments in a listener's phonetic transcriptions of child productions and the corresponding adult forms appropriate to that dialect (i.e., a relational analysis). 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 that word. However, the new measure is intended to be more fine-grained than PCC in that different types of errors are weighted differently. For example, phoneme omissions and unusual errors are weighted more heavily, but errors involving common substitutions are given smaller weights. The WSSA was implemented through use of the Logical International Phonetics Program (LIPP) software (Oller & Delgado, 1999), and it relies on definitions of errors specified in a LIPP alphabet and analysis program (i.e., LIPP Analysis Language [LAL]).[1]

In developing the WSSA, several adaptations were made to the weighted reliability definitions described by Oller and Ramsdell (2006), which were designed to compare two transcribers' renditions of a vocalization for purposes of transcription reliability. For those purposes, the direction of the difference did not matter; for example, if one transcriber perceived [t] and one perceived [k], it did not matter which transcriber perceived which phone. However, for purposes of comparing an adult target with a child's production, we argue that the direction of the difference matters; /k^{h}æp/ → [t^{h}æp] is developmentally quite common, whereas /t^{h}æp/ → [k^{h}æp] is rather uncommon (e.g., Dodd & Iacano, 1989; Edwards & Shriberg, 1983). Thus, many of the consonant rules developed by Oller and Ramsdell (2006) were modified to differentially weight errors based on principles of typical phonological development (e.g., markedness). The weights assigned to vowel errors in the original weighted analysis were maintained in the WSSA, and the calculation procedures (i.e., multiplying the global structural agreement by the featural agreement) were not changed from Oller and Ramsdell (2006). Although there must necessarily be some degree of arbitrariness to the assigned weightings, the multilevel weighting system (*teeny, small, big,* and *huge*) is based on well-founded tenets of phonological theory and, therefore, provides a grounded framework for ranking speech sound errors.

Because many possible types of consonant and vowel errors can occur, a complete review of all permutations is not possible; however, a review of the major concepts is presented. 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 (*global structural agreement*) and in features of the segments that are represented (*featural agreement*). Deletions of phonemes are weighted most

[1] LIPP is a publically marketed software program developed by Intelligent Hearing Systems (see http://www.ihsys.com/site/LIPP.asp?tab=4). The analysis routine for implementing the WSSA calculation in LIPP, along with the accompanying alphabet, can be obtained from the first author.

heavily because they represent changes in syllable and word structure (e.g., *dad* /dæd/ → [dæ]). With respect to features, major changes are weighted heavily (e.g., producing a glottal sound for an orally articulated sound), and minor feature changes are weighted less (e.g., producing a homorganic stop in place of a fricative). Scores can range from 0.0 (*no match in number of segments or in features*) to 1.0 (*complete match of segments and features*).

Transcription and alignment. Following Oller and Ramsdell (2006), *slots* are positions for (vertical) alignments of phones between the target form and the child's production of a word. Transcriptions are entered into the LIPP software program, with an adult form of the target word(s) on the "target" row and the child's production of the target word(s) on the "transcription" row. This target form is a phonetic form that can therefore be adapted to reflect dialectally acceptable variations. For example, reduced vowels in unstressed syllables (e.g., the final syllable of *elephant*) are sometimes realized as a schwa [ə] but may also be realized as a high lax front vowel [ɪ]. Accordingly, the vowel in the target would be adjusted to match the vowel in the child's production so that a penalty is not computed by the WSSA algorithm. Other dialectally acceptable variations can also be allowed simply by modifying the adult form so that it matches the child's production (e.g., because /tɹ/ and /dɹ/ clusters may be produced with affrication, the acceptable target form for *tree* could be either [tɹi] or [ʧɹi], depending on the child's production). Thus, target forms may be modified slightly to accommodate dialect variation so that the algorithm will not calculate these as errors.

Target and child productions can differ in the number of existing segments and in the features of aligned segments. Thus, four principles are followed in comparing the adult target and the child's production in order to standardize the process of alignment (see, e.g., Oller & Ramsdell, 2006; Ramsdell, Oller, & Ethington, 2007). The alignment principles are as follows:

1. *Strict order*: There is no reordering of segments in the transcribed targets or in the child's productions during the process of aligning the two.
2. *Nucleus alignment first*: Vowels are first aligned with other vowels, then consonants are aligned with other consonants.
3. *Matched segment*: Transcriptions with the same number of vowel and consonant segments ordered in the same way are aligned correspondingly so that the vowel and consonant segments match.
4. *Minimal discrepancy*: Segments with maximally similar phonetic features are aligned in order to produce minimal discrepancy between transcriptions of the target form and child production.

In accordance with these guidelines, nuclei were aligned first in our transcriptions. Children rarely add nuclei, but if this occurred, the added nucleus was aligned with other consonants or was included in an empty (orphan) slot. For those instances in which 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, following the minimal discrepancy principle (e.g., [m] was aligned with the target /p/ rather than with /s/ because /m/ and /p/ share place of articulation, whereas /m/ and /s/ share no features).

Calculating the WSSA score. Once the transcriptions are aligned, the WSSA program in LIPP calculates a *global structural agreement* based on phoneme deletions and adjunctions (additions) in the child's production as compared with the adult target. The global structural agreement score represents the proportion of filled (nonempty) slots. Therefore, this score penalizes for sound deletions and adjunctions (i.e., unfilled slots in either the target form or the child's production) by assigning a 0.0 agreement to the slot. The current system weights deletions or additions of 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).

The WSSA algorithm then computes the *featural agreement* score for the slots in which there is a target sound and a produced sound (paired slots), indicating how closely the produced sounds match the target sound. Each paired slot begins with a value of 1.0 and is reduced for any consonant or vowel substitutions based on the consonant–consonant, vowel–vowel, or consonant–vowel alternations described below. The average phonetic accuracy of the sounds that are produced is then computed to derive the featural agreement score (see Appendix A). Thus, if all sounds produced by the child exhibit complete phonetic accuracy when compared with the target (e.g., [su] for /su/), the average feature agreement will be 1.0. If many of the features are in error (e.g., [ɡu] or [ɡo] for /su/), a lower featural agreement score will result. Note that the mean featural agreement for the child's production of [u] for /su/ is 1.0 because the phoneme that is produced is phonetically accurate. (In this case, the deleted /s/ would have been accounted for in the global structural agreement score.) Finally, to derive the WSSA, the global structural agreement score is multiplied by the mean featural agreement score. Featural agreements are defined in the paragraphs that follow.

Consonant substitutions. Consonant substitutions are weighted on the basis of features of place, manner, and voicing. If the child production and the target form of a consonant are exact matches, the featural agreement is 1.0 for that slot. 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). Minor errors in place of articulation result in a small penalty (score reduction), whereas major errors result in a larger penalty. On the basis of 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 (e.g., Edwards & Shriberg, 1983; Hodson & Paden, 1981). Table 1 lists feature weights for consonant substitution errors. Following rules defined by Oller and 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 having five levels of weighting) each could be completely accurate (with a score of 0.333) or, alternatively, could be 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) = 0.167], 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, with only four levels of weighting) could be completely accurate or could be

Table 1. Consonant features and penalties for errors.

Consonant feature (weight)		Penalty	Example		
Manner (0.333)	Huge manner	–0.3333	Plosive	→	Fric. or affric./#____
	Uncommon errors,		Glide	→	Liquid
	damaging to intelligibility		Nasal	→	Non-nasal
			Semivowel	→	Nasal
			Sonorant	↔	Obstruent
	Big manner	–0.25	Plosive	→	Fric. or affric./C or V____
	Less common in		Fric. or affric.	→	Lateral fric. or affric.
	phonological development				
	Small manner	–0.1666	Fric. or affric.	→	Plosive
	Common errors in		Fric.	↔	Affric.
	phonological development		Liquid	→	Glide or tap
	Teeny manner	–0.0833	Nonspecific distortion		
	Minor phonetic errors				
Place (0.333)	Huge place	–0.333	Dorsal	↔	Labial
	Uncommon, very damaging		Glottal	↔	Nonglottal
	to intelligibility				
	Big place	–0.25	Coronal	↔	Labial
	Less common in		Coronal	→	Dorsal
	phonological development		Alveolar	→	Palatal
			Palatal	→	Dental
			Retroflex	↔	Not retroflex
	Small place	–0.1666	Linguadental	↔	Labiodental
	Typical errors in		Dental	↔	Alveolar
	phonological development		Palatal	→	Alveolar
			Dorsal	→	Coronal
	Teeny place	–0.0833	Bilabial	↔	Labiodental
	Phonetic errors in English,		Lips not spread	↔	Lips spread
	based on small changes		Lips not round	↔	Lips round
	in tongue placement		Labialization		
			Blading		
			Tongue advance/retract		
Voicing (0.333)	Huge voicing	–0.3333	Word-initial or medial devoicing		
	Uncommon		Word-final voicing		
	Small voicing	–0.2222	Word-final devoicing		
	Common		Word-initial voicing		
	Teeny voicing	–0.1111	Aspiration of nonaspirated C (e.g., ste → st^{h}e)		
	Phonetic changes				

Note. Fric. = fricative; affric. = affricate; C = consonant; V = vowel.

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 height (high, high lax, mid, mid lax, and low), tongue advancement (front, back, or central), rounding (round or not round), and nasalization (nasal or not nasal). Therefore, 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 receives a score of 1.0 for that vowel, with minor deviations from the target (e.g., /i/ → [ɪ]) 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 that compute errors when vowels are substituted for syllabic or nonsyllabic consonants and vice-versa. The penalties for these errors are mutually exclusive of other penalties (i.e., a substitution error cannot be penalized for both consonant–consonant alternations and consonant–vowel alternations). These errors are generally weighted heavily, as they involve changes in major sound classes (consonants and vowels). The most extreme errors result in a score of 0.0 for that sound. Such errors would include producing an obstruent in place of a vowel or vice versa (e.g., /t/ → [æ]) or a producing a low vowel in place of a high semivowel (e.g., /j/ → [a]). Less severe errors include substituting a vowel for a syllabic liquid (e.g., *bicycle* /baɪsɪkl̩/ → [baɪsɪko]). Table 3 lists penalties for consonant–vowel alternations. Note that some of these errors were 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 includes a provision for scoring such productions.

Speech Samples

To investigate the validity and reliability of the WSSA, we obtained speech samples from children of a variety of ages from several prior studies, including samples of typically speaking (TS) children and those with SSDs. Our goal was to sample a wide range of phonetic accuracy across a range of ages, populations, and speech samples. A brief description of each group is provided, and we refer readers to prior work for further details of the participants, recording parameters, and so forth. Table 4 summarizes the groups and the data available from each group.

Group 1: Preschoolers with SSDs. Forty-four children who had not yet begun kindergarten (ages 4;0–5;9 [years; months]) from upstate/central New York were referred by SLPs as having an SSD of unknown origin. Participants achieved a standard score of below 90 on the Goldman–Fristoe Test of Articulation (2nd ed.; GFTA–2; Goldman & Fristoe, 2000), 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). In a second session, participants were digitally audio recorded naming 125 pictures chosen to elicit many consonant clusters and multisyllabic words. Responses from all 125 words were later transcribed into the LIPP software program by the first author (a certified SLP with expertise in SSDs and 6 years of graduate training in phonological disorders).

Group 2: Young adolescents with and without SSDs naming pictures. A group of adolescents (ages 10;0–15;0) 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 TS adolescents who had no history of speech-language difficulty. All of these participants were recorded and transcribed together by the first author and the fourth author (a clinical phonologist with more than 30 years of experience in phonetic transcription of children's speech) to achieve a consensus (Shriberg, Kwiatkowski, & Hoffmann, 1984; see Preston & Edwards, 2007, for recording and transcription information). Another five children between the ages of 9;0 and 13;0 (one with a lateral lisp and four with errors

Table 2. Vowel feature weights and penalties for errors.

Vowel feature	Weight	Penalties			Example
Height	(0.40)	Huge height	−0.40	4-step height change	/i/ ↔ [ɑ]
		Big height	−0.30	3-step height change	/ɪ/ ↔ [ɑ]
		Small height	−0.20	2-step height change	/i/ ↔ [e]
		Teeny height	−0.10	1-step height change	/ɑ/ ↔ [ɛ]
Advancement	(0.40)	Big front	−0.40	Front ↔ Back	/o/ ↔ [e]
		Small front	−0.20	Front ↔ Central or Back ↔ Central	/i/ ↔ [ə]
Nasalization	(0.1)	Small nasal	−0.10	Not nasal → Nasal	/ɑ/→ [ɑ̃]
Rounding	(0.1)	Small rounding	−0.10	Round ↔ Not round	/ʌ/ ↔ [ɔ]

Table 3. Penalties for consonant–vowel alternations.

Alternation			Penalty
Vowels and syllabic consonant alternation			
Vowel (nonglide)	↔	Syllabic liquid	–0.25
Vowel (nonglide)	↔	Syllabic nasal, tap, trill	–0.05
Vowel (nonglide)	↔	Syllabic voiced obstruent	–0.75
Vowel (nonglide)	↔	Syllabic unvoiced obstruent	–1.0
Vowels and nonsyllabic consonant alternation			
Liquid	↔	Any vowel (nonglide)	–0.5
High semivowel	↔	High vowel	
Mid semivowel	↔	Mid vowel	
Low semivowel	↔	Low vowel	
High or low vowel	↔	Mid semivowel	–0.75
High vowel	↔	Nonsyllabic nasal, tap, trill	
Low semivowel	↔	High vowel	–1.0
High semivowel	↔	Low vowel	
Obstruent	↔	Vowel	

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 between 9 and 15 years of age with known SSDs (11 male, 8 female) and 19 TS young adolescents (8 male, 11 female) with no history of speech problems.

Group 3: Preschoolers with and without SSDs naming pictures. As part of a larger study, 18 male children between the ages of 3;10 and 5;4 participated in an extensive picture-naming task (see Conture, Louko, & Edwards, 1993; Wolk, Edwards, & Conture, 1993). Ten of the children exhibited SSDs (this was confirmed by clinicians, and the children had a GFTA [Goldman & Fristoe, 1986] percentile of < 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. All 120-word speech samples were transcribed online by the fourth author. The audio- and video recordings were later reviewed for the purpose of refining the transcriptions.

Group 4: Spontaneous speech of typically developing toddlers. Transcriptions from a publicly available online dataset—the PhonBank portion of the Child Language Data Exchange System (CHILDES) project (MacWhinney, 2000)—were used. This included samples, provided by child phonologist Barbara Davis, of typically developing English-speaking children from Texas who were recorded 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). These data were used to evaluate the ability of the WSSA to quantify

Table 4. Summary of participants, speech samples, and additional data on phonetic accuracy.

Group	*n*	Speech sample for WSSA	Other measures derived from same speech sample	Other measures available for validity
1. Preschoolers	44 SSD	125-word picture naming	PCC, PCC–R, Atypical Errors per Consonant	GFTA–2, SLP ratings of severity
2. Young adolescents	19 SSD 19 TS	64-word picture naming	PCC, PCC–R	
3. Preschoolers	10 SSD 8 TS	60-word picture naming—Form A 60-word picture naming—Form B	PCC	GFTA
4. Toddlers	14 TS	Connected speech or elicited words	PCC	

Note. WSSA = Weighted Speech Sound Accuracy; SSD = speech sound disorder; PCC = Percent Consonants Correct; PCC–R = Percent Consonants Correct—Revised; GFTA = Goldman–Fristoe Test of Articulation; SLP = speech-language pathologist; TS = typically speaking.

longitudinal change in phonetic accuracy (over at least 10 weeks) and for test–retest reliability (sessions fewer than 15 days apart).

In addition, data were included from a typically developing English speaking boy, here referred to as MR, who was recorded during playtimes, book reading, and mealtimes with an Olympus WS-331M digital voice recorder. He was recorded approximately once or twice per month over 6 months from the ages of 22 to 28 months. Only those productions in which the target word was known (ranging from 13 to 81 utterances per sample) were transcribed by the first author. We included this participant in order to replicate the results from the PhonBank dataset.

Validity Measures

Concurrent validity: Severity. The WSSA algorithm was run on all transcription data to generate a score from 0.0 to 1.0, and these scores were compared with other measures for validation. To assess concurrent validity, PCC was derived from every speech sample. In addition, PCC–R was computed for speech samples from Groups 1 and 3. GFTA–2 scores were available from Group 1 to validate the new measure against a standardized instrument. The earlier version of this instrument, the GFTA (Goldman & Fristoe, 1986), was available for children from Group 3.

From Group 1, 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 novel 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.

To test the validity of the WSSA with respect to clinical judgment, the 125-word speech sample audio files from 20 preschoolers in Group 1 were used to obtain judgments of severity. Each of the 125 words was extracted from the digital sound file and was paired with a spoken number from 1 to 125. These number–word pairs were then concatenated into a single audio file of approximately 5 min 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 SLPs, all of whom were certified by the American Speech-Language-Hearing Association (ASHA), had earned a Certificate of Clinical Competency (CCC), and had clinical experience working with preschoolers. Each SLP, informed only of the child's age and gender, listened to speech samples of five children. The SLPs 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. Thus, they were to use their clinical experience and "trained ear" as a guide for determining severity. The SLPs used the following 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,* and *profound SSD*. These categories were selected to represent an ordering of severity using terms that clinicians might recognize while allowing listeners to make distinctions between children whose severity might differ somewhat. 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 who listened to a child was used for subsequent analysis of perceived severity of the child's SSD; these median scores spanned categories from *normal* to *very severe*.

Construct validity: Distinguishing SSD and TS. Although phonetic accuracy exists along a continuum, clinicians and researchers often choose to consider SSD as dichotomous. Therefore, we examined whether WSSA scores differentiate children with SSD from children with TS. We used Groups 2 and 3 to address whether WSSA scores can differentiate these groups as defined by prior studies.

Construct validity: Growth in phonetic accuracy. We hypothesized that a measure that differentially weights errors will be sensitive to phonological development. Therefore, toddlers who had data spanning at least 10 weeks were included; these toddlers consisted of seven children from the PhonBank dataset and case study MR. A child's data from a session were included and entered into LIPP for the present study based on the following criteria: the child had at least six samples at different times, and each sample had at least 12 different utterances transcribed in which the gloss (target) was known. Hence, unintelligible words were not used. If a child produced a word or phrase multiple times, it was included only twice. In order to capture production variability, we used the first two renditions that were different (e.g., [bau] and [ba] for *ball*). Up to 100 utterances per session were entered into LIPP (with a minimum of 12 different utterances). Appendix B lists the PhonBank samples used.

Reliability Measures

Interrater reliability. From Group 1, the fourth author independently transcribed 25 words from 41 preschoolers (three children were transcribed together and, therefore, were not completed independently; thus, 41 of 44 were included). WSSA scores derived from those transcriptions were compared with WSSA scores derived from the first author's transcriptions. Additionally, 21 participants from Group 2 were transcribed by the second author (who is also a CCC-SLP with 6 years of graduate training in phonetics and phonology), and WSSA scores derived from those transcriptions were compared with scores derived from consensus transcriptions of the first and fourth authors working together.

Test–retest reliability. Participants from Group 1 were recorded speaking the GFTA–2 target words and then were recorded at a second session naming 125 pictures. From the GFTA–2 and the picture-naming task, 20 words overlapped. Therefore, we compared WSSA scores from those 20 words spoken at both sessions, including data from 34 children with samples collected within 15 days. Although this time frame was arbitrary, we expected reasonably high correlations and little significant change in phonetic accuracy for samples collected in 15 days or fewer.

Group 4 data were also used to address test–retest reliability. Participants were included from the PhonBank dataset (plus case study MR); these participants had at least 25 utterances per session in two sessions that were fewer than 15 days apart. If utterances were produced multiple times, the first two different renditions were included. This resulted in test–retest data from 14 toddlers.

Alternate-form reliability. To address *alternate-form reliability* (i.e., the agreement between two samples from the same child), Group 3 data were used. These included 120 words 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. WSSA scores from these alternate forms were compared.

Internal consistency. To address internal consistency, 25 consecutive words from the 125 word list were randomly selected from Group 1 data. Then, WSSA scores for the 25-word subsample were compared with scores for the larger 125-word sample for these children.

Results

Data Analysis

To examine whether parametric statistics were appropriate for the new WSSA measure, we used Kolmogorov–Smirnov tests to compare the data distribution against a null hypothesis of a normal distribution in each of the four groups. These tests failed to reject the null hypothesis of normal distribution (ps > .11). Thus, the data did not violate assumptions of normality.

Because the goal was to demonstrate validity and reliability of the WSSA, these psychometric properties were addressed with converging evidence from various samples. Correlations between WSSA scores and other relevant variables are presented using Pearson's r for samples of 15 or more and for data that met assumptions of normality and interval scale and using Spearman's ρ for samples of fewer than 15 or for ordinal data.

Validity

Concurrent validity: Severity. We examined whether WSSA scores correlated with existing measures of phonological severity, including PCC, GFTA–2 raw scores, and SLP 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 5). 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 the median SLPs' categorical description of the severity of a child's SSD in 20 of these preschoolers (Spearman's $\rho = -0.882$, $p < .001$).

To evaluate whether WSSA scores captured nondevelopmental speech sound errors better than do other measures, we also examined the Pearson's r correlation of the WSSA score and the number of atypical errors per consonant (the novel measure described earlier in this article; Preston, 2008; Preston & Edwards, 2010). As can be seen in Table 5, there was a stronger correlation between this measure and WSSA than 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;0–15;0 with and without SSDs naming 64 pictures. The correlation between WSSA scores and PCC was high for the entire sample ($r = 0.94$, $p < .001$, 95% confidence interval [CI] = 0.89–0.97), demonstrating strong agreement across the broad range of typical and disordered speech. When considering only the 19 participants with SSDs, the correlation was somewhat lower ($r = 0.66$, $p = .002$, 95% CI = 0.26–0.96).

WSSA and PCC scores were also highly correlated in Group 3 ($r = 0.90$, $p < .001$, 95% CI = 0.75–0.96 among all preschoolers, and $r = 0.714$, $p = .001$ among the 10 preschoolers with SSD) and were highly correlated with GFTA percentile rank ($\rho = 0.917$, $p < .001$). In Group 4—the youngest cohort—WSSA scores showed moderate

Table 5. Correlations between WSSA scores and other measures of speech sound accuracy from 44 children with SSDs.

Measure	PCC	PCC–R	GFTA–2 raw score	GFTA–2 standard score	GFTA–2 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

Note. WSSA, PCC, PCC–R, and Atypical Errors per Consonant all derived from the same 125-word picture naming task. All correlations are significant at $p < .02$.

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

correlations with PCC ($r = 0.69$, $p < .001$, 95% CI = 0.54–0.79 for the 68 speech samples).

Construct validity: Distinguishing SSD and TS. To test whether WSSA scores distinguished children with and without SSDs, we compared participants in Group 2 using an analysis of variance (ANOVA), with speech group (SSD and TS) and gender as factors. There was no statistically significant Group × Gender interaction, $F(1, 34) = 0.24$, $p = .631$, $\eta^2_p = .007$, and the main effect of gender was not significant, $F(1, 34) = 0.003$, $p = .954$, $\eta^2_p = 0.000$. In contrast, the main effect of group was statistically significant, $F(1, 34) = 58.2$, $p < .001$, $\eta^2_p = .631$, demonstrating that the TS adolescents differed reliably on their WSSA scores from adolescents with SSD. Moreover, Figure 1 demonstrates that there is no overlap between adolescents with SSD and TS adolescents. WSSA scores ≤ 0.966 (below the dotted line) were associated with SSD, whereas scores > 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, we used Group 3 to evaluate group separation on WSSA scores in preschool boys with and without SSD. Figure 2 represents WSSA scores for both groups. There was a slight overlap between the groups; a cutoff score of < 0.910 (see dotted line in Figure 2) would categorize all of the children with SSD correctly. However, this score would incorrectly categorize one child with TS as having an SSD. His WSSA scores were 0.891 and 0.892 from Lists A and B, respectively. Arguably, because this child inconsistently demonstrated 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 vs. TS) as a fixed effect and speech sample (Lists A and B) as a random effect. Only the main effect of group was statistically significant, $F(1, 29) = 69.0$, $p < .001$, suggesting that these preschoolers with and without SSD differed in their WSSA scores.

Construct validity: Growth in phonetic accuracy. To determine whether 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 got older; thus, there should be positive associations between speech sound accuracy and age. Table 6 presents rank-order correlation coefficients (Spearman's ρ)

Figure 1. Weighted Speech Sound Accuracy (WSSA) scores from a 64-word picture-naming sample plotted against age for 19 adolescents with speech sound disorder (SSD; circles) and 19 typically speaking (TS) 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. yrs = years.

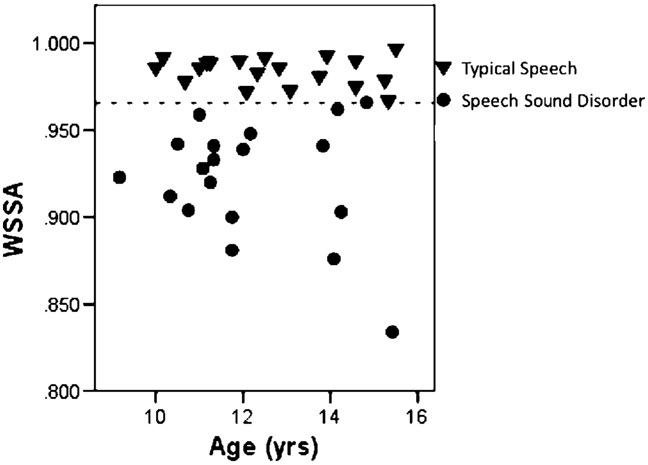
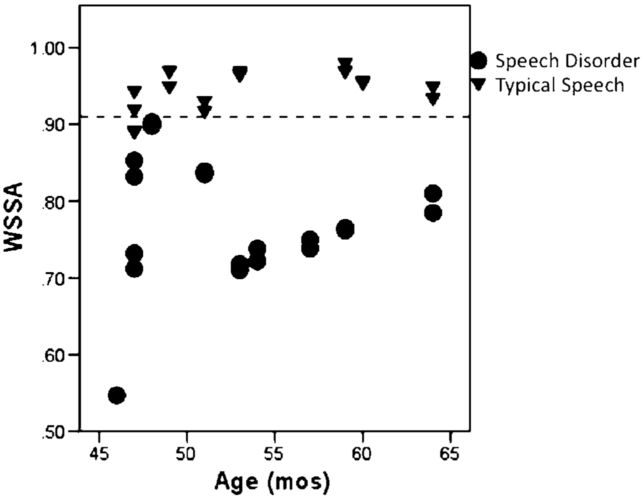


Figure 2. WSSA scores plotted against age for nine preschool boys with SSD (circles) and nine TS preschoolers (triangles). Vertically adjacent symbols represent two speech samples from each child (List A and List B). The dotted line represents a WSSA score of 0.910, below which all children with SSD scored, and above which eight of nine TS preschoolers scored. mos = months.



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. To illustrate, Figure 3 displays WSSA and PCC scores plotted against age for participant MR. In the top panel, PCC scores show a slight decrease in

Figure 3. WSSA and percent consonants correct (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 accuracy that can be accounted for by age.

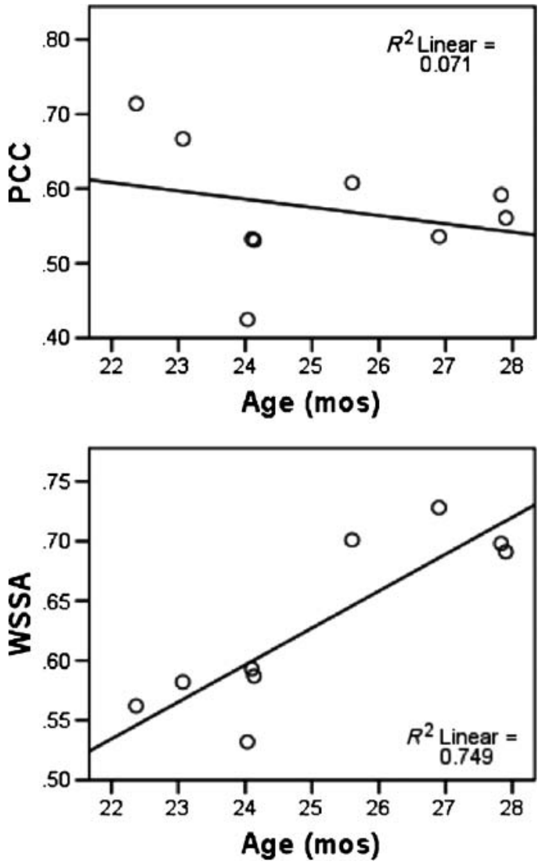


Table 6. Nonparametric correlations between age and two measures of speech sound accuracy.

Child	Age range (mos.)	# 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	.20	.66
Cameron	22.23–35.80	12	.58*	.60*
Charlotte	31.80–35.00	6	.74	−.03
Hannah	23.00–27.23	9	.87**	.07
Rachel	17.13–22.10	11	.21	.07
MR	22.37–26.90	7	.78*	−.2

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

*$p < .05$. **$p < .01$.

accuracy over time. In the bottom 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. Thus, from the same transcriptions, phonetic accuracy was found to increase between 22 and 28 months for this child when using the WSSA score as a measure, but this was not the case when phonetic accuracy was quantified using PCC.

Reliability

Interrater reliability. From Group 1, the first and fourth authors independently transcribed 25 consecutive words from 41 preschoolers, with starting points randomly determined for each child. Correlations between the WSSA scores derived from these transcriptions of two listeners for 41 children was $r = 0.925$ ($p < .001$, 95% CI = 0.86–0.96). Pearson's correlations between the first and fourth author's PCC scores for those same samples was $r = 0.854$ ($p < .001$, 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, the second author transcribed the 64 words from 21 participants (12 TS, 9 SSD). WSSA scores derived from those transcriptions were highly correlated with WSSA scores from consensus transcriptions completed together by the first and fourth authors ($r = 0.930$, $p < .001$, 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 consensus transcriptions completed by the first and fourth authors ($r = 0.932$, $p < .001$, 95% CI = 0.84–0.97).

Test–retest reliability. Participants from Group 1 were recorded speaking the GFTA–2 and, in a separate session, naming 125 pictures. From these two samples, 20 words that were spoken on both occasions were used for test–retest reliability. Correlations between WSSA scores for the 20 words spoken at two test points (within 15 days) were high among the 34 children with SSD ($r = 0.901$, $p < .001$, 95% CI = 0.81–0.95) and were not significantly different, $t(33) = 1.6$ ($p = .119$).

Connected speech samples from 14 children in Group 4 (13 from the Davis PhonBank dataset plus case study MR) who had two samples of at least 25 words recorded within 15 days were also used to address test–retest reliability. WSSA scores from the two sessions correlated highly ($r = 0.948$, 95% CI = 0.840–0.984) and were not significantly different, $t(13) = 1.0$ ($p = .328$).

Alternate form reliability. From 17 children in Group 3, speech samples were split into Lists A and B, each representing 60 of the 120 target words. In Figure 2, two vertically adjacent (often overlapping) symbols represent each child's scores on the two 60-word samples. It is clear that the two sets of words yielded similar WSSA scores. One child with SSD completed only the first 60 words and, therefore, is included only once. Pearson's correlation between WSSA scores derived from List A and List B (each containing 60 words) was $r = 0.990$ ($p < .001$, 95% CI = 0.972–0.996). There was no statistically significant difference between WSSA scores from Lists A and B, $t(16) = 1.16$ ($p = .262$). From those same samples, correlation between PCC scores derived from Lists A and B were high ($r = 0.973$, $p < .001$, 95% CI = 0.93–0.99) and not significantly different, $t(16) = 1.28$ ($p = .218$).

Internal consistency. From the 44 participants in Group 1, 25 words were randomly chosen from the 125-word sample. WSSA scores derived from the 25-word subsamples correlated highly with the child's WSSA score from the 125 words ($r = 0.931$, $p < .001$, 95% CI = 0.867–0.962). PCC derived from the same 25-word subsample also correlated highly with the child's overall PCC from the 125 words ($r = 0.877$, $p < .001$, 95% CI = 0.784–0.931). Thus, the scores from a smaller sample of 25 words were strongly associated with scores from the 125-word sample.

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 that were generally similar to those of PCC. On the basis of cross-sectional data, we have demonstrated that the WSSA is sensitive to disorders of speech production. Longitudinal data demonstrated that the WSSA is sensitive to phonological development. The external validity of the measure is supported by empirical evidence presented from multiple speakers, speech samples, recording environments, recording equipment, and transcribers.

The WSSA scores correlated 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 (a) this instrument scores articulation accuracy based on only a single occurrence of each consonant or cluster in a given word position and (b) scoring procedures do not take into account the nature of the errors or the degree of accuracy. WSSA scores also correlated with the number of atypical sound errors per consonant (Preston & Edwards, 2010), suggesting that an advantage of the WSSA is that it is sensitive to atypical error patterns. In addition, WSSA scores correlated with SLPs' judgments of severity; if SLP judgments of an SSD are a standard by which other more objective metrics should be compared (Flipsen, Hammer, & Yost, 2005), then WSSA and PCC–R appear to be relatively similar in capturing severity. These correlations provide support for the concurrent validity of the WSSA.

WSSA scores correlated highly with PCC scores in preschoolers and adolescents. However, in Groups 2 and 3, correlation coefficients between WSSA and PCC scores were lower when selecting just the children with SSD than when examining all children, possibly because the range of performance was substantially limited (i.e., only low scores rather than a range of low to high scores). Similarly, WSSA and PCC correlations were somewhat lower in toddlers than in the older groups. Thus, the association between WSSA and PCC scores appears to be lower when focusing on children whose speech is less accurate (i.e., SSD or very young children) than when examining the entire range of phonetic accuracy.

WSSA scores were positively associated with age for all of the longitudinal participants, suggesting that these scores are sensitive to phonetic development (see Table 5). In young children, WSSA scores can account for improvements in how well segments in the word are represented, whereas binary judgments of accuracy (i.e., PCC) penalize for even minor errors and might be less sensitive to small changes in phonetic accuracy (note, however, that PCC was developed to quantify severity rather than developmental change). Because conversational samples were used to examine speech sound accuracy in the longitudinal samples, there was 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 quantification of phonetic accuracy. Given the diversity of age ranges and the small number of children with longitudinal data, these results should be viewed with caution. However, these results provide initial support for the notion that WSSA scores are sensitive to small improvements in speech production over time.

WSSA scores were found to correlate strongly between transcribers, suggesting good interrater reliability. WSSA scores in Group 1 derived from two independent transcribers showed a higher correlation than did PCC scores from those same samples (although the 95% CIs did overlap). Slightly better correlations between transcribers for WSSA than for PCC is probably because WSSA scores depend on the accuracy of the features of segments rather than on 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 (i.e., less sensitive to small disagreements) than a system based on binary judgments of the accuracy of a sound. Interrater reliability data from Group 2 suggest nearly identical correlations between transcribers for WSSA and PCC scores.

Stability of WSSA scores was supported by the alternate form reliability, which showed that WSSA scores derived from different words yielded similar results. Internal consistency was supported by the fact that a subsample of 25 words correlated highly with children's larger samples of 125 words. Additionally, test–retest data demonstrated that a child's WSSA scores were highly correlated and showed no systematic change in a 15-day period for conversational samples from typically developing toddlers or from 20-word samples from preschoolers with SSD.

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 (except, e.g., voicing errors).

Note 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 WSSA algorithm if a researcher judges that there is sufficient theoretical, empirical, or experimental reason to do so. The present study, however, demonstrates that the WSSA is valid and reliable. Thus, there are several possibilities for implementing this measure. For example, the WSSA could be used with larger prospective longitudinal studies to evaluate the range of performance among TS children, or it might be used as a tool to track progress in treatment for children with SSDs. The WSSA could also be adapted for speakers of other languages or expanded to capture prosodic factors such as lexical stress. The development of a sensitive, reliable, and viable measure of phonetic accuracy should be considered an ongoing process.

In conclusion, psychometric properties of measures used to quantify speech sound accuracy can and should be investigated. We view the present WSSA to be an improvement over prior measures of speech sound accuracy because it is more sensitive to the types of errors produced by children; it may, therefore, more precisely capture phonetic accuracy as associated with development and disorders. We recognize that no single measure is likely to completely capture every type of error and weight errors in such a way as to achieve 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 National Institute of Child Health and Human Development Grant T32HD07548 awarded to Haskins Laboratories (Carol Fowler, Principal Investigator [PI]) and by 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 National Institute on Deafness and Other Communication Disorders (NIDCD) Research Grant DC00523-01A2 awarded to Syracuse University. Participation of the second and third authors was supported by NIDCD Research Grant DCR01 006099 (D. Kimbrough Oller, PI). Thanks to Barbara Davis and the developers of the CHILDES database.

References

Conture, E. G., Louko, L. J., & Edwards, M. L. (1993). Simultaneously treating stuttering and disordered phonology in children: Experimental treatment, preliminary findings. *American Journal of Speech-Language Pathology, 3,* 72–81.

Crary, M. A. (1984). Phonological characteristics of developmental verbal dyspraxia. *Seminars in Speech and Language, 5*(2), 71–83.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research, 38,* 1199–1211.

Davis, B. L., MacNeilage, P. F., & Matyear, C. L. (2002). Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica, 59*(2–3), 75–107.

Dodd, B., & Iacano, T. (1989). Phonological disorders in children: Changes in phonological process use during treatment. *British Journal of Disorders of Communication, 24,* 333–351.

Edwards, M. L., & Shriberg, L. D. (1983). *Phonology: Applications in communicative disorders*. San Diego, CA: College-Hill.

Flipsen, P., Jr., Hammer, J. B., & Yost, K. M. (2005). Measuring severity of involvement in speech delay: Segmental and whole-word measures. *American Journal of Speech-Language Pathology, 14,* 298–312.

Goldman, R., & Fristoe, M. (1986). *Goldman–Fristoe Test of Articulation*. Circle Pines, MN: AGS.

Goldman, R., & Fristoe, M. (2000). *Goldman–Fristoe Test of Articulation* (2nd ed.). Circle Pines, MN: AGS.

Hodson, B. W., & Paden, E. P. (1981). Phonological processes which characterize unintelligible and intelligible speech in early childhood. *Journal of Speech and Hearing Disorders, 46,* 369–373.

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language, 29,* 713–733.

Ingram, D., & Ingram, K. D. (2001). A whole-word approach to phonological analysis and intervention [Clinical Forum]. *Language, Speech, and Hearing Services in Schools, 32,* 271–283.

Jakielski, K. (1998). *Motor organization in the acquisition of consonant clusters* (Unpublished doctoral dissertation). University of Texas, Austin.

Kazdin, A. E. (2003). *Research design in clinical psychology*. Boston, MA: Allyn & Bacon.

Klein, E. S., & Flint, C. B. (2006). Measurement of intelligibility in disordered speech. *Language, Speech, and Hearing Services in Schools, 37,* 191–199.

Leonard, L. B. (1985). Unusual and subtle phonological behavior in the speech of phonologically disordered children. *Journal of Speech and Hearing Disorders, 50,* 4–13.

MacWhinney, B. (Ed.). (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mawhah, NJ: Erlbaum.

Morris, S. R. (2009). Test–retest reliability of independent measures of phonology in the assessment of toddlers' speech. *Language, Speech, and Hearing Services in Schools, 40,* 46–52.

Oller, D. K., & Delgado, R. E. (1999). *Logical International Phonetic Programs* [Computer program]. Miami, FL: Intelligent Hearing Systems.

Oller, D. K., & Ramsdell, H. L. (2006). A weighted reliability measure for phonetic transcription. *Journal of Speech, Language, and Hearing Research, 49,* 1391–1411.

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

Preston, J. L. (2008). Phonological processing and speech production in preschoolers with speech sound disorders (Doctoral dissertation). Retrieved from ProQuest (Pub. no. AAT 3333580).

Preston, J. L., & Edwards, M. L. (2007). Phonological processing skills of adolescents with residual speech sound errors. *Language, Speech, and Hearing Services in Schools, 38,* 297–308.

Preston, J. L., & Edwards, M. L. (2010). Phonological awareness and speech error types in preschoolers with speech sound disorders. *Journal of Speech, Language, and Hearing Research, 53,* 44–60.

Ramsdell, H. L., Oller, D. K., & Ethington, C. A. (2007). Predicting phonetic transcription agreement: Insights from research in infant vocalizations. *Clinical Linguistics & Phonetics, 21,* 793–831.

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

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.

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders. III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47,* 242–256.

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

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

Wolk, L., Edwards, M. L., & Conture, E. G. (1993). Coexistence of stuttering and disordered phonology in young children. *Journal of Speech and Hearing Research, 36,* 907–917.

Received February 2, 2010

Revision received May 4, 2010

Accepted June 21, 2010

DOI: 10.1044/1092-4388(2010/10-0030)

Contact author: Jonathan L. Preston, Haskins Laboratories, 300 George St., Suite 900, New Haven, CT 06511. E-mail: preston@haskins.yale.edu.

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).

Weighted speech sound accuracy (WSSA): The global structural agreement multiplied by the featural agreement.

Table A-1. 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 = 1.0.
- ɛ → 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.
- v → m, labiodental fricative → bilabial nasal; huge manner deduction of 0.333 and teeny place deduction of 0.0833; featural agreement = 0.583.
- ɪ → ɪ, no deduction; featural agreement = 1.0.
- 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 + 0.583 + 1 + 0.668 + 0.5)/6] = 0.742.

WSSA: 0.667 × 0.742 = 0.495

Table A-2. 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 = 1.0.
- p → p; no deduction; featural agreement = 1.0.
- ɪ → ɪ; no deduction; featural agreement = 1.0.
- n → n; no deduction; featural agreement = 1.0.
- ɚ → ɔ; mid central vowel with r-color → mid back lax vowel; small front deduction of 0.2 and 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 + 0.7)/5] = 0.94.

WSSA: 0.714 × 0.94 = 0.685.

Appendix B. PhonBank files.

Files were selected from http://childes.psy.cmu.edu/browser/index.php?url=PhonBank/English-Davis.

Samples Included in Longitudinal Analysis

Aaron:	Aar05, Aar06, Aar07, Aar08, Aar09, Aar10, Aar11
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
MR:	Not from PhonBank

Samples Included in Test–Retest Analysis

Aar10 & Aar11
Ant11 & Ant12
Ben24 & Ben25
Geo43 & Geo44
Han28 & Han29
Jod11 & Jod12
Kate07 & Kate08
Mar11 & Mar12
Racl40 & Rac41
Nat39 & Nat40
Nic46 & Nic47
Pax48 & Pax49
Row38 & Row39
MR (not from PhonBank)