JSLHR, Volume 40, 708–722, August 1997

The Percentage of Consonants Correct (PCC) Metric: Extensions and Reliability Data

Lawrence D. Shriberg
Diane Austin
University of Wisconsin–Madison

Barbara A. Lewis
Case Western Reserve University
Cleveland, OH

Jane L. McSweeny
David L. Wilson
University of Wisconsin-Madison

Research in normal and disordered phonology requires measures of speech production that are biolinguistically appropriate and psychometrically robust. Their conceptual and numeric properties must be well characterized, particularly because speech measures are increasingly appearing in large-scale epidemiologic, genetic, and other descriptive-explanatory database studies. This work provides a rationale for extensions to an articulation competence metric titled the Percentage of Consonants Correct (PCC; Shriberg & Kwiatkowski, 1982; Shriberg, Kwiatkowski, Best, Hengst, & Terselic-Weber, 1986), which is computed from a 5- to 10-minute conversational speech sample. Reliability and standard error of measurement estimates are provided for 9 of a set of 10 speech metrics, including the PCC. Discussion includes rationale for selecting one or more of the 10 metrics for specific clinical and research needs.

KEY WORDS: phonology, articulation, speech disorders, assessment, measurement

A lucid tutorial review by Kent, Miolo, and Bloedel (1994) provides comparative analysis of 19 procedures that researchers and clinicians have used to assess intelligibility of speech in children. Although the focus is on intelligibility assessment, rather than on measures that index severity of speech involvement, the procedures included in this tutorial and the discussion of relevant psychometric issues provide good coverage of the state of the art in the measurement of speech disorder in children. Kent and colleagues divide their evaluative analysis of assessment procedure into five primary categories, including those that emphasize phonetic contrast analyses, phonological analyses, word identification, scaling methods, and phonetic accuracy in continuous speech. In the present context, the most salient observation these authors underscore is "the relatively little work that has been done to evaluate the reliability and validity of the procedures developed to date" (p. 90).

The present work addresses reliability and validity issues for an approach to speech assessment that Kent and colleagues subsumed under the category of phonetic accuracy in continuous speech. This work continues the directions suggested in an earlier article providing rationale and validity data for several speech and prosody-voice measures for genetics research and other descriptive studies in developmental phonological disorders (Shriberg, 1993). Each metric is derived from a conversational speech sample, including a procedure to obtain an index of intelligibility in natural conversational speech. The present work adds information to the earlier report in the following areas: (a) descriptions

 1092-4388/97/4004-0708

of several new articulation competence indices that provide methodological alternatives to the *Percentage of Consonants Correct* (*PCC*) metric (Shriberg & Kwiatkowski, 1982; Shriberg et al., 1986), (b) point-to-point transcriber agreement and standard error of measurement estimates for 9 of the 10 speech measures (excepting the Intelligibility Index), and (c) rationale for selecting one or more of the 10 measures for clinical and research questions. A companion work provides information on a clinical classification measure and provides lifespan reference data for all 10 speech measures (Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997).

Extensions to the Percentage of Consonants Correct (PCC) Metric

The Percentage of Consonants Correct (PCC) metric expresses the percentage of intended consonant sounds in a conversational sample that were articulated correctly. Since its initial (Shriberg & Kwiatkowski, 1982) and updated (Shriberg et al., 1986) development, variants of this approach have been used by investigators pursuing diverse questions in child phonology. Typically, applications have been consistent with methods used in the validation studies. However, sometimes procedures have differed. For example, PCCs have sometimes been used to represent the percentage of correct consonants on an articulation test or on some other nonconversational speech task. Because the proportional distribution of intended consonants in such samples differs from the distribution of consonants in continuous conversational speech, PCCs based on different sampling contexts are not directly comparable. Also, reported PCCs have sometimes been based on broad phonetic transcription rather than on the response definitions and conventions for narrow phonetic transcription used in the validation study. For such applications, too, it would be inappropriate to convert the resulting percentage of consonants correct into the four severity classifications suggested in the validation study (> 90% = *mild*, 65%–85% = *mild-moderate*, 50%–65% = *moderate-severe*, and < 50% = *severe*).

Six methodological concerns and suggestions have been expressed by colleagues who have used the PCC for clinical research. The following discussion considers these concerns and suggestions, prompting rationale and procedures for several new metrics to meet specific clinical research needs.

Concern 1: PCC Scores Require a Conversational Speech Sample

Colleagues have expressed the following two concerns with the requirement of a conversational speech sample as the basis for a PCC score.

Representativeness

The first concern is with the lack of standardization relative to the responses evoked in conversational speech. Several colleagues have suggested that a standard set of stimulus materials for a continuous conversational speech sample would provide content stability within and across research laboratories.

A perspective documented in Shriberg and Kwiatkowski (1985) and Morrison and Shriberg (1992) and advocated in Shriberg (1993) is that the speech data obtained from conversational samples are linguistically and psychometrically robust, regardless of semantic content. It is these findings that have motivated the data presented in the companion paper (Shriberg et al., 1997); these data are based on conversational speech samples from 836 persons between the ages of 3 and 40+ years. The breadth of conversational topics represented in the corpora within and across ages is viewed as support for the external validity of the descriptive statistics for each metric.

Productivity

A second and more frequent concern is that, although conversational speech samples may yield representative data, on an individual basis they are not invariably productive or efficient with all children. For some children, speech rates (i.e., words per minute) may be exceedingly slow and speech may be too unintelligible for the examiner to gloss.

Experience with over 1,000 conversational samples from children suggests that the crucial factor in the productivity of speech sampling is the examiner's skill in evoking and glossing speech. Specifically, the ability to obtain linguistically rich speech samples from sometimes reluctant talkers seems more to reflect an examiner's ability to converse (traditionally, *rapport*) than the types of stimulus materials or prescribed topics used to evoke speech (cf. Shriberg & Kwiatkowski, 1985). However, it is useful to underscore a constraint discussed in earlier reports. There is a small proportion of children who are not candidates for conversational speech sampling because of their severe phonological and/or discourse deficits. For such children, the use of a set of interest-appropriate pictured stimuli to evoke conversational speech is generally successful, and, as suggested above, a standard set of materials for such purposes would be quite useful. More generally, the challenge to design and validate materials that could be used for speakers of all ages and many other relevant demographic characteristics is formidable, but not insurmountable.

Concern 2: PCC Scores Reflect Weighted Performance on All 24 English Consonants

Two concerns have been expressed about the computational structure of the PCC.

Developmental Sound Classes

The first structural concern is that, as an index of speech competence based on all 24 English consonants, a PCC score may obscure important differences associated with only certain sounds or only certain subgroups of sounds. To address these needs, the *speech profile* approach described in the earlier report provides three subscale scores in addition to the total PCC score (Shriberg, 1993). Subscale percentages are computed for three *developmental sound classes* termed the *Early-8* (m,b,j,n,w,d,p,h), *Middle-8* (t,ŋ,k,g,f,v,tʃ,dʒ), and *Late-8* (ʃ,θ,s,z,ð,l,r,ʒ) consonant sounds. As shown in the reference data reported in the companion article (Shriberg et al., 1997), children with speech delay typically have nearly all of the Early-8 English consonant sounds correct, only some of the Middle-8 sounds correct, and few of the Late-8 sounds correct. Such descriptive detail, which is unavailable in the total PCC score, is typically of central interest in clinical research settings.

When the 24 English consonants are treated as one response class, as they are in the total PCC score, other statistically, theoretically, or clinically significant differences in articulation competence might be obscured. The speech profile approach provides a number of additional subscale percentage tallies, including percentages by individual sounds, by singletons and clusters, by class and manner features, by error types, and by absolute and relative percentages for each error type (cf. Shriberg, 1993). When available in addition to total PCC scores, data at this level provide for fine-grained inspection of phonological status in cross-sectional and longitudinal studies. As shown later, subscale scores are available for 5 of the 10 measures described in this paper.

Percentage of Consonants in the Inventory (PCI)

A second structural question about the PCC concerns its weighting of errors on sounds by the frequency of occurrence of each sound in conversational speech. The validity study indicated that the frequency of occurrence of a consonant error in conversational speech was highly correlated with the percept of "severity of involvement," with errors on more frequently occurring consonants counting more than errors on less frequently occurring consonants.

An alternative perspective on speech-sound acquisition is concerned with the number or percentage of sounds mastered, rather than the per-sound percentage correct. Such inventories depict which target consonants a child has been observed to produce correctly, regardless of how often they are correctly articulated in conversational speech. Typically such relational inventories (cf. Stoel-Gammon & Dunn, 1985) are presented in some type of place-manner format arranged for ease of developmental analysis.

The PCI was developed to meet the need for a quantitative, relational inventory of consonants mastered, particularly for questions involving very young children. The denominator of the PCI is the number of English consonants (out of the total of 24) intended in a sample, with single attempts at a consonant weighted 0.5 and two or more attempts weighted 1.0. The numerator is the number of consonants for which one or more correct articulations occurred, again with a 0.5 weight for a single correct production and a 1.0 weight for two or more correct productions. Thus, each consonant can contribute a maximum of 1.0 to the total percentage calculation across all consonants attempted in the sample.

Given the focus on the consonant inventory ("phonetic" inventory is a misnomer, because such inventories typically reflect only a child's recognizable consonants; "phoneme" inventory is also inaccurate unless the sounds have been attested as contrastive), only omissions and substitutions are considered errors for the PCI calculations (see below). That is, distorted sounds are nevertheless considered to be *in* a child's consonant inventory. In addition to the total PCI, subscale PCI percentages are calculated to reflect inventories for the Early-8, Middle-8, and Late-8 consonants. As above, for children with limited discourse skills or severely reduced intelligibility, a set of age-appropriate stimuli may be needed to evoke a sufficient sample for a PCI.

Concern 3: PCC Scores Weight Distortions as Heavily as Omissions and Substitutions

A third concern is that the PCC metric gives equal weight to the three types of speech-sound errors—omissions, substitutions, and distortions. Justification for weighting all error types equally, as opposed to the weighting schemes used in some published articulation tests, rests on information from the validation studies. As described in Shriberg and Kwiatkowski (1982), it was the total percentage of consonants *correct* that was significantly associated with perceived "severity of involvement" in conversational samples evaluated by groups of clinically experienced and inexperienced listeners. That is, the variance in severity of involvement captured by the PCC is statistically associated with the percentage of correctly articulated sounds, not with error-type characteristics.

For some clinical research questions, however, it is useful to differentiate speech-sound omission and substitution errors from speech-sound distortion errors. For example, a figure in Shriberg (1993) illustrated how the standard deviations of PCC scores overlap at each age for 3-, 4-, and 5-year-old speech-normal children compared

to children referred for speech delay. The overlap is due to the high percentage of distortion errors in young children with both normal and delayed speech acquisition when speech is narrowly transcribed (using diacritics in addition to phoneme symbols). That is, although young children with normal speech do not have age-inappropriate omission and substitution errors—errors that define speech delay in the Speech Disorders Classification System reviewed in the companion paper (Shriberg et al., 1997)—they do have a sufficient proportion of common distortion errors to overlap the high end of the PCC scores for children with speech delay.

Articulation Competence Index (ACI)

Rationale for an alternative to the PCC metric termed the *Articulation Competence Index* (*ACI*), which differentially weights distortion errors, was reported in Shriberg (1993). The ACI is calculated as follows:

$$ACI = \frac{PCC + RDI}{2}$$

where RDI = the Relative Distortion Index. The Relative Distortion Index is the percentage of a speaker's errors that are distortions. Although the ACI successfully separates distributions of children acquiring speech normally from those with speech delay, several constraints were noted in Shriberg (1993) and have been noted by colleagues who have attempted to use the ACI. First, weighting distortions by adding the RDI to the PCC is arbitrary, rather than motivated by specific findings. Second, the convention that ACI = PCC for PCC scores above 95 yields nonlinearity in the distribution of ACI scores (cf. Shriberg, 1993). Third, children with similar ACI scores could have quite different component percentages of PCC and RDI scores. Fourth, and perhaps most troublesome, the ACI does not differentiate among types of consonant distortions, specifically common versus uncommon clinical distortions.

Two alternatives to the PCC metric have been developed to meet these four concerns: the *Percentage of Consonants Correct–Adjusted* (*PCC-A*) and the *Percentage of Consonants Correct–Revised* (*PCC-R*). Differences between the new measures and the PCC are shown in the first three rows of Table 1.

Percentage of Consonants Correct

As shown in the first row in Table 1, only correctly articulated consonants are scored as correct on the PCC. Each of three other possible classes of responses—common clinical distortions, uncommon clinical distortions, and deletions or substitutions—are scored as incorrect. Common clinical distortions (cf. Shriberg, 1993, Appendix) include the following sound changes: (a) labialized and velarized /l/; (b) labialized, velarized, and derhotacized /r/,/ɜ˞/, and /ɚ˞/; and (c) dentalized and lateralized /s/, /z/, /ʃ/, /ʒ/, /tʃ/, and /dʒ/. Common clinical distortions are observed in young children acquiring speech normally, in children with speech delay (SD), and, importantly, in children with residual errors (RE). It is this latter criteria—that some distortions persist as residual errors—that defines the class of common clinical distortions. Uncommon clinical distortions are simply all other clinically relevant distortions (see list in Shriberg, 1993, Appendix).

Percentage of Consonants Correct–Adjusted (PCC-A)

A speaker's sound changes on the *Percentage of Consonants Correct–Adjusted* (*PCC-A*) are calculated in the same way as on the PCC. As shown in the second row of Table 1, however, common clinical consonant distortions are also scored as correct on the PCC-A. Thus, a PCC-A score "ignores" this class of allophones (i.e., common clinical distortion errors) on speech-sound targets that are otherwise phonemically correct. The result is that PCC-A is specifically sensitive to children with speech delay, who, as defined by most researchers, have phoneme deletions and substitutions. For example, if a child's consonant errors consisted of *only* common clinical distortions, the PCC score would be less than 100%, whereas the PCC-A would be 100%.

Percentage of Consonants Correct–Revised (PCC-R)

The *Percentage of Consonants Correct–Revised* (*PCC-R*) is also calculated in the same way as the PCC. However, as shown in the third row of Table 1, both common and uncommon clinical consonant distortions are scored as correct. Thus, a child who had 10% common and 10% uncommon clinical consonant distortions—but no consonant deletions or substitutions—would score 80% on the PCC, 90% on the PCC-A, and 100% on the PCC-R. The difference between a speaker's PCC-A and PCC-R scores (in the above example, 10%) indicates the percentage of all distortion errors that are uncommon clinical distortions. Rationale for choosing among these three measures or the others described below for clinical and research questions is provided in a later section.

Concern 4: PCC Scores Index Only Consonant Articulation

Percentage of Vowels/Diphthongs Correct (PVC)

A fourth concern colleagues have expressed about the PCC as an index of articulation competence is that it reflects only consonant production. To provide additional

Table 1. Scoring differences among the Percentage of Consonants Correct (PCC) and six other measures of articulation competence in conversational speech.

Metric	Correctly articulated sounds		Common clinical distortions		Uncommon clinical distortions		Deletions and substitutions	
	Consonants	Vowels	Consonants	Vowels	Consonants	Vowels	Consonants	Vowels
Percentage of Consonants Correct (PCC)	+	–	0	–	0	–	0	–
Percentage of Consonants Correct–Adjusted (PCC-A)	+	–	+	–	0	–	0	–
Percentage of Consonants Correct–Revised (PCC-R)	+	–	+	–	+	–	0	–
Percentage of Vowels Correct (PVC)	–	+	–	0	–	0	–	0
Percentage of Vowels Correct–Revised (PVC-R)	–	+	–	+	–	+	–	0
Percentage of Phonemes Correct (PPC)	+	+	0	0	0	0	0	0
Percentage of Phonemes Correct–Revised (PPC-R)	+	+	+	+	+	+	0	0

Note. + Scored as correct. 0 Scored as incorrect. – Not scored.

information on a speaker's articulation of the 17 vowels and diphthongs of American English, the *Percentage of Vowels/Diphthongs Correct* (*PVC*) profile was added to the speech profile series (Shriberg, 1993). As shown in the fourth row in Table 1, the metric is computed in the same way as the PCC, with vowel/diphthong deletions, substitutions, and all distortions considered incorrect. The speech profile format also provides individual percentages by sound, feature class, rhotic versus nonrhotic vowels, diphthong type (phonemic, nonphonemic), error type, and absolute versus relative errors. Two concerns limiting the usefulness of the PVC have been addressed in three new metrics described next.

Percentage of Vowels/Diphthongs Correct–Revised (PVC-R)

Rationale for the *Percentage of Vowels Correct–Revised* (*PVC-R*) parallels the concern discussed above for distortions on consonant targets. Because all vowel/diphthong distortions are scored as errors on the PVC, it has poor discrimination among error patterns consisting of distortions only (e.g., [ɜ˞], where [◡] indicates a derhotacized r-colored vowel) versus those that include substitutions as well. The PVC-R is calculated in the same way as the PVC, but as shown in Table 1, all distortions are considered correct. Thus, a child with only distortions on vowels/diphthongs will score lower than 100% on the PVC, but 100% on the PVC-R.

Percentage of Phonemes Correct (PPC) and Percentage of Phonemes Correct–Revised (PPC-R)

A problem with use of the PCC or PCC-R and the PVC or PVC-R is that none of the measures alone reflect a total index of all American English phonemes. To meet the need of a combined consonant and vowel/diphthong index, the method proposed by Dollaghan (1995) and Dollaghan, Biber, and Campbell (1993) is useful. Dollaghan and colleagues proposed a *Percentage of Phonemes Correct* (*PPC*) measure to reflect the percentage of both consonants and vowels/diphthongs articulated correctly on responses to a nonsense repetition task assessing phonological memory. For their research needs, these authors scored all distortions as correct. To be consistent with the terminology used in the present context, however, the Percentage of Phonemes Correct (PPC) as defined in Table 1 scores only correct responses on consonant and vowel/diphthong phonemes as correct (i.e., distortions are considered incorrect). The *Percentage of Phonemes Correct–Revised* (*PCC-R*) scores both correctly articulated speech sounds and all speech-sound distortions as correct (see Table 1).

Concern 5: PCC Scores Are Unadjusted for Age and Gender

A fifth concern with PCC scores is that they are not adjusted for potential individual differences associated with age and gender—relevant correlates in developmental, epidemiologic, and descriptive-explanatory studies of traits and behaviors. It would be useful to be able to account for possible age or gender differences in articulation competence. For example, in situations where age and gender cannot be entered as covariates in an analysis it would be useful to be able to convert raw scores to *z* scores or some other standardized score. Pending the possibility of demographically representative normative data, a companion paper (Shriberg et al., 1997) presents an example of what are termed *lifespan reference data* assembled from a clinical research database of conversational speech samples for Age × Gender groups between the ages of 3 and 40+ years. The database includes descriptive statistics, *z* scores, and standard scores for each of 10 speech measures.

Concern 6: PCC Scores Lack a Standard Error of Measurement

A sixth concern with PCC scores is that they lack well-developed estimates of a standard error of measurement. Although the test-retest reliability and internal stability of PCC scores were documented in Shriberg and Kwiatkowski (1982), this study did not include an estimate of the standard error of measurement for phonetic transcription agreement. Rather, using comparisons of PCC scores of samples completed by different transcriptionists, the only suggestion was that a speaker's PCC was reliable within approximately four percentage points. Thus, cautious interpretation was recommended for evaluating the reliability of PCC scores within four percentage points of one another or any score within four percentage points of a classification boundary (i.e., mild > 85%, mild-moderate 65%–85%, moderate-severe 50%–65%, and severe < 50%).

In phonology and other areas of communication science, estimates of the reliability of phonetic transcription or acoustic analyses are nearly always computed by point-to-point percentage of agreement. The advantage of agreement percentages, especially when corrected for guess rates, is that they estimate the reliability of data at the level of each sound, word, or other unit of analysis. What they do not directly provide, however, is an estimate of the reliability of the total or summative score an individual receives on the domain of interest.

In contrast to point-to-point percentage of agreement, the conventional reliability statistic used in test and measures development is a correlation coefficient (Pearson *r* or Spearman ρ), which expresses the degree of association between two sets of scores. The obtained correlation is, in turn, used to compute a standard error of measurement (SEM = $SD\sqrt{1-r}$). The computed standard error of measurement expresses a 68% probability that a person's true score will be within the SEM (in either direction) from the obtained score. SEMs provide important information about measures in a test protocol—information that is especially useful in providing guidance in interpreting the significance of effect sizes. The following section provides point-to-point transcriber reliability estimates and SEM values for 9 of the 10 speech measures (i.e., excluding the Intelligibility Index).

Transcription Agreement, Classification Agreement, and Standard Error of Measurement Estimates

Method

Subjects

Table 2 is a summary of demographic and reliability information for 33 children and adults whose conversational speech samples were used in a series of transcriber and metric reliability studies. The 33 speech samples were randomly selected from a database of children and adults whose speech was recently assessed in two collaborative and two local studies. A standard protocol for speech sampling and recording was used in each of the studies (cf. Shriberg & Kwiatkowski, 1994). The protocol included the use of high-quality audiocassette tape recorders with matching external microphones. The goal was to assemble, from the larger database of several hundred samples, an approximately 10% sample representing speakers with diverse demographic (age, gender, geographic community) and speech (normal acquisition, mild to severe speech delay, residual errors) backgrounds. Subset 1 includes 11 speakers, age 4 years 11 months to 44 years 7 months, randomly selected from a speech genetics study conducted with families in the Cleveland, Ohio, area (Lewis & Shriberg, 1994). Subset 2 includes 10 children, age 5 years 6 months to 6 years 6 months, randomly selected from a group of 5- to 7-year-old children assessed in an epidemiologic study of language disorders conducted in three geographic centers (including three social strata: urban, suburban, and rural) in Iowa (Tomblin, 1991). Subset 3 includes seven children, age 3 years 9 months to 5 years 9 months, randomly selected from a longitudinal study of speech-delayed children (Shriberg, Gruber, & Kwiatkowski, 1994) drawn primarily from the Madison, Wisconsin, area. Subset 4 includes five children from Wausau, Wisconsin, randomly selected

Table 2 Reliability estimates for broad and narrow phonetic transcription and Speech Disorders Classification System (SDCS) classification.

					Reliability of phonetic transcription				Reliability of SDCS classification[a]	
					% Broad agreement		% Narrow agreement		Transcriber (T)	
Subset	Sample number	Age (years:months)	Gender	Number of words used	Consonants	Vowels	Consonants	Vowels	T1	T2 or T3
1	1	4:11	M	256	92.5	87.6	81.2	79.4	{NSA-}	NSA-
	2	5:11	M	238	94.3	87.4	83.7	78.4	{NSA-}	SD
	3	6:5	M	221	97.4	84.4	95.5	82.6	NSA-	NSA-
	4	8:10	M	176	97.1	89.6	87.9	80.9	{NSA-}	{NSA-}
	5	10:6	F	270	86.4	77.8	65.3	65.8	RE-B2	RE-B2
	6	12:4	M	186	93.4	82.8	69.9	76.6	RE-B1	[RE-B2]
	7	14:0	M	209	98.1	85.7	95.6	83.0	NSA	NSA
	8	14:8	M	178	97.6	87.4	90.7	82.1	NSA	{NSA-}
	9	19:7	F	203	96.8	86.6	94.0	85.6	NSA	NSA
	10	35:7	M	193	97.7	85.2	96.8	84.5	NSA	NSA
	11	44:7	M	216	98.2	89.1	84.8	86.3	NSA	NSA-
			Total	2,346	95.5	85.7	86.1	80.4		
			Mean	213.3	95.4	85.8	85.9	80.5		
			SD	31.0	3.6	3.3	10.5	5.7		
2	1	5:6	M	208	90.2	89.1	79.9	81.3	{NSA-}	NSA-
	2	6:0	M	217	96.5	88.5	89.0	82.7	{NSA-}	NSA-
	3	6:0	F	218	97.4	87.9	88.8	84.1	NSA-	{NSA-}
	4	6:0	M	256	91.4	87.3	78.2	80.6	NSA-/SD	NSA-/SD
	5	6:0	M	189	91.9	85.3	74.9	79.4	NSA-/SD	NSA-/SD
	6	6:2	M	188	97.1	90.8	87.4	79.1	{NSA-}	{NSA-}
	7	6:3	F	218	94.8	86.3	78.5	83.8	{NSA-}	QRE-B
	8	6:3	M	198	95.3	83.3	80.7	78.6	QRE-B	NSA-/SD
	9	6:5	F	222	96.1	93.4	87.0	86.5	NSA-	NSA-
	10	6:6	M	183	95.9	84.0	80.4	66.0	{NSA-}	NSA-
			Total	2,097	94.6	87.7	82.5	80.5		
			Mean	209.7	94.7	87.6	82.5	80.2		
			SD	21.7	2.6	3.1	5.1	5.6		
3	1	3:9	M	80	86.5	89.0	69.1	61.4	{NSA-}[+]	NSA-
	2	3:10	M	173	89.7	79.1	75.6	66.8	NSA-/SD	QSD
	3	4:3	M	152	92.4	91.5	77.0	71.9	{NSA-}	NSA-/SD
	4	4:5	F	152	93.7	81.3	76.1	67.3	NSA-/SD	NSA-/SD
	5	4:10	M	243	88.4	88.4	70.1	80.0	SD	SD
	6	5:4	M	136	88.5	85.7	73.7	70.2	{NSA-}	NSA-/SD
	7	5:9	M	112	90.3	88.9	77.3	73.3	SD	SD
			Total	1,048	90.0	86.0	73.9	71.2		
			Mean	149.7	89.9	86.3	74.1	70.1		
			SD	51.2	2.5	4.5	3.3	5.9		
4	1	3:10	M	241	80.9	85.7	71.3	77.9	SD	SD
	2	3:11	M	191	88.5	85.8	70.1	79.3	SD	SD
	3	4:5	F	153	81.5	80.9	71.8	66.0	SD	SD
	4	5:2	M	151	90.5	89.9	78.1	84.2	NSA-	NSA-
	5	5:10	M	247	92.6	88.9	81.0	82.4	{NSA-}	NSA-
			Total	983	87.3	86.5	74.7	78.7		
			Mean	196.6	86.8	86.2	74.5	78.0		
			SD	46.2	5.3	3.5	4.8	7.1		
		SUMMARY:	Mean	196.2	92.7	86.5	80.6	77.8		
			SD	42.7	4.6	3.5	8.5	7.0		

[a]Essentially each classification consists of a stem and optional affixes and bracket qualifiers: NSA = Normal Speech Acquisition. SD = Speech Delay. NSA/SD = Between NSA and SD. RE = Residual Errors. See companion paper (Shriberg et al., 1997) for description of each classification.

from a cross-sectional study of speech-sound stimulability (Lof, 1994).

Transcription

Two persons with many years of research experience in narrow phonetic transcription of conversational speech (designated Transcribers 1 and 2) and one person with approximately one year of experience (Transcriber 3) independently transcribed samples using similar procedures and protocols (cf. Shriberg, 1986). Transcriber 1 completed all 33 speech samples; Transcriber 2 completed Subsets 1, 2, and 3; and Transcriber 3 completed Subset 4. The transcribers were not familiar with the speakers on each audiocassette recording and were informed only of the speaker's age and gender.

Analysis Procedures: Transcription Agreement

The *PEPAGREE* program (Shriberg & Olson, 1988) was used to calculate transcribers' broad and narrow phonetic agreement for consonants and vowels/diphthongs. The concordance procedures in this software require that comparisons be made on the same intended word forms. The procedure uses one transcriber's gloss as the standard, so that the program compares a second transcriber's transcription of a word to the standard transcriber's transcription of the same word. Accordingly, the PEPAGREE program cannot be used to calculate the reliability of the Intelligibility Index, a measure that requires a different set of reliability and validity procedures discussed elsewhere (Kwiatkowski & Shriberg, 1992; Shriberg & Kwiatkowski, 1982; Shriberg & Lof, 1991; Weston & Shriberg, 1992) and in the previously noted review by Kent et al. (1994).

First, Transcriber 1's gloss (orthographic transcription) and intended phonological forms (phonetic representation of the gloss) were copied and distributed to the other transcriber. The other transcriber then listened to the appropriate section of tape and completed narrow phonetic transcription of the speaker's realized forms. Third, the computer files were used to obtain *consonant agreement* and *vowel/diphthong agreement* outputs for each pair-wise comparison, as shown in Table 2. Total agreement percentages were calculated for each subset by adding the numerators and denominators used to calculate pair-wise agreement, dividing the sum of the numerators by the sum of the denominators, and multiplying by 100 to get a percentage. Additionally, means and standard deviations of the pair-wise percentages are reported for each subset, as well as a summary mean and standard deviation for all 33 pair-wise comparisons.

Analysis Procedures: Standard Error of Measurement

The computer files used for the standard error of measurement (SEM) data (to be reported in Table 3) were the set of original and reliability transcripts. Thus, transcripts for the SEM analyses used the same initial gloss, but all disagreements in phonetic transcription and word forms were entered as transcribed. Scores for each of the 9 metrics on each of the transcript files were then obtained using the appropriate PEPPER reports. Again, scores for the Intelligibility Index could not be obtained because the same gloss was used for both files. Finally, the SEM was calculated using the distributional statistics and correlation coefficients for each measure, as described in a later section.

Sample Size

As shown in Table 2, the number of words transcribed for each modified speech sample ranged from 80 words (Subset 3, Sample 1) to 270 words (Subset 1, Sample 5). Over all 33 samples, the mean was 196.2 words, with a standard deviation of 42.7 words. It should be noted that although these transcripts are long enough for reliability estimates, longer transcripts are especially useful for classification by the Speech Disorders Classification System (SDCS; see Shriberg et al., 1997). For example, in one longitudinal study in which rich conversational samples were needed for both speech and language analysis, the original transcripts for each speech sample averaged over 500 words.

Point-to-Point Transcription Agreement

The sixth through ninth columns in Table 2 are the point-to-point interjudge transcription agreement data computed by the PEPAGREE program. The entries in these columns suggest the following observation about interjudge agreement when based on point-to-point comparison.

The overall findings for broad and narrow phonetic transcription of consonants and vowels/diphthongs are consistent with findings for a larger sample of transcribers and speakers as reported in detail in Shriberg and Lof (1991). In that study (p. 267), broad transcription for consonants averaged 93% across transcriber consensus teams, and narrow transcription averaged 74%—nearly 20 percentage points lower. Findings for the present study, which used different conversational speech samples, were quite similar for consonants, averaging 92.7%. The average overall agreement in the present study for narrow phonetic transcription, 80.6%,

is approximately seven percentage points higher than the 74% overall figure reported in Shriberg and Lof. The difference is readily traced to speaker characteristics. For Subsets 1 and 2, the two samples of essentially normally speaking children and adults, the mean narrow phonetic transcription agreement percentages for consonants were 85.9% and 82.5%, respectively. For Subsets 3 and 4, the two samples of essentially speech-delayed speakers, the mean narrow phonetic transcription agreement percentages for consonants were 74.1% and 74.5%, respectively—in close agreement with the estimates for speech-delayed speakers reported in Shriberg and Lof. Thus, there is a nearly 20-percentage-point difference in transcriber agreement between broad and narrow phonetic transcription when most speakers have some type of speech involvement. This gap is reduced as proportionally more of the speakers have essentially normal speech competence. As shown in Table 2, the estimates for vowels are similar to consonants in magnitude and patterning for broad versus narrow transcription of normal versus speech-delayed speakers.

As concluded in Shriberg and Lof (1991), the reliability of broad transcription seems to be adequate for clinical research questions, whereas the reliability of narrow phonetic transcription may be inadequate for some assessment purposes. This difference in the reliability of broad and narrow phonetic transcription will be important to the upcoming examination of the standard error of measurement estimates.

Reliability of SDCS Classifications

As described in Shriberg (1993) and extended in the companion paper (Shriberg et al., 1997), the Speech Disorders Classification System (SDCS) is a qualitative measure of articulation competence. The rightmost set of columns in Table 2 provides telling information on the consequences of differences in phonetic transcription for clinical classification of children's speech status. As underscored above, the magnitudes of point-to-point percentages of agreement do not directly reflect the reliability of articulation competence scores based on phonetic transcription. The latter can be estimated only by pair-wise comparisons such as shown for the SDCS outcomes in Table 2.

A primary observation about the SDCS outcomes in Table 2 is that, notwithstanding the lowered point-to-point interjudge agreement percentage data shown to the left, there is reasonably good agreement on SDCS classifications based on samples from different transcribers. Ignoring all brackets (which, as described in the companion paper, qualify the level of support for each classification; see Shriberg et al., 1997), 23 of the 33 (70%) SDCS outcomes for Transcriber 1 agree exactly with those of Transcriber 2 or Transcriber 3. Of the 10 remaining disagreements, 3 disagreements in SDCS classification involve a disparity only in the "–" or the "+" affix with the NSA stem (see companion paper). The remaining seven differences (18% of the total of 33 comparisons for Transcribers 1, 2, and 3) involve differences in the stems of SDCS classifications—differences that have more serious consequences for clinical research questions. Note that these four subsets were deliberately assembled to provide the most difficult test of SDCS classification agreement. If all of the subsets had included only children with significant speech delay, as in Subset 4, the overall percentage of agreement in SDCS outcomes would have been greater.

What these SDCS classification data demonstrate, as a precursor to the SEM data to be reported, is that point-to-point percentage of agreement data are not sufficient to estimate the reliability of a classification or score. In the present situation, for example, consider the agreement between two transcribers for Subset 4. Whereas their point-to-point narrow phonetic transcription agreement on consonants and vowels ranged from only 66.0% to 84.2%, their SDCS classification agreement was nearly perfect.

Standard Error of Measurement Estimates

As reviewed previously, a standard error of measurement (SEM) reflects a 68% probability that a person's true score will be within the calculated SEM (in either direction) from the obtained score. If there is need for a more conservative confidence interval, two SEMs above and below the score can be used with 95% confidence that the true score falls within this interval.

Table 3 provides SEM data for the articulation competence metrics. As noted previously, procedures and data on the reliability of the Intelligibility Index reported elsewhere differ substantially from those used to assess the reliability of phonetic transcription and are not amenable to SEM computation at this time. The columns for each subset in Table 3 include the average mean, average standard deviation, rank-order correlation, and the SEM for each pair-wise comparison. Columns on the right include a summary of the range of SEMs across the four subsets and summary SEMs calculated on all 33 samples. Rank-ordered Spearman Rho coefficients were used for the correlational statistic at the subsets level, in consideration of low cell sizes (5–11 pairs of scores per sample). For the SEM computed across all samples, the 33 pair-wise comparisons met customary criteria for parametric analyses using the Pearson correlation coefficient. Three characteristics of the data in Table 3 warrant comment.

Table 3. Standard error of measurement estimates.

	Subset 1 (n = 11)[a]				Subset 2 (n = 10)[a]				Subset 3 (n = 7)[a]				Subset 4 (n = 5)[b]					Pooled Samples (n = 33)			
Metric	***M***	***SD***	**ρ**	**SEM**	***M***	***SD***	**ρ**	**SEM**	***M***	***SD***	**ρ**	**SEM**	***M***	***SD***	**ρ**	**SEM**	**SEM Range**	***M***	***SD***	***r***	**SEM**
PCC																					
Early[c]	96.2	4.5	.82	1.9	96.6	2.8	.88	1.0	92.8	2.8	.71	1.5	92.3	4.9	.63	3.0	1.0 –3.0	95.0	4.1	.84	1.6
Middle	94.0	8.6	.75	4.3	91.1	5.0	.67	2.9	71.6	10.0	.75	5.0	67.8	20.2	1.00	0.0	0.0–5.0	84.4	15.1	.95	3.4
Late	78.5	21.7	.92	6.1	65.6	21.4	.95	4.8	31.8	16.3	.50	11.5	32.1	23.2	.87	8.4	4.8–11.5	42.6	28.5	.93	7.5
Total	90.7	9.4	.91	2.8	86.4	7.7	.89	2.6	70.5	7.4	.74	3.8	68.3	12.7	1.00	0.0	0.0 –3.8	81.7	12.9	.95	2.9
PCC-A																					
Early[c]	96.2	4.5	.82	1.9	96.6	2.8	.88	1.0	92.8	2.8	.71	1.5	92.3	4.9	.63	3.0	1.0–3.0	95.0	4.1	.84	1.6
Middle	94.0	8.6	.75	4.3	91.2	4.9	.67	2.8	71.7	10.2	.75	5.1	67.8	20.2	1.00	0.0	0.0–5.1	84.5	15.1	.96	3.0
Late	83.2	17.2	.92	4.9	83.0	11.8	.89	3.9	46.7	16.3	.89	5.4	42.0	27.9	.80	12.5	3.9–12.5	69.2	25.4	.93	6.7
Total	92.0	8.1	.86	3.0	91.3	4.9	.88	1.7	74.5	7.1	.85	2.7	71.3	14.1	.98	2.0	1.7–3.0	84.9	12.1	.95	2.7
PCC-R																					
Early	96.9	3.9	.78	1.8	97.4	2.5	.49	1.8	93.6	2.8	.50	2.0	93.3	4.6	.67	2.6	1.8–2.6	95.8	3.8	.81	1.7
Middle	94.5	8.2	.85	3.2	91.9	4.4	.61	2.7	73.3	10.3	.76	5.0	69.5	20.6	.90	6.5	2.7–6.5	85.4	14.6	.96	2.9
Late	85.4	16.6	.95	3.7	83.7	11.9	.89	3.9	47.8	16.5	.89	5.5	43.7	28.9	.90	9.1	3.7–9.1	70.6	25.5	.95	5.7
Total	93.1	7.7	.91	2.3	92.0	4.7	.90	1.5	75.6	6.8	.85	2.6	72.6	14.6	.90	4.6	1.5–4.6	86.0	11.9	.96	2.4
ACI																					
Early	88.4	19.0	.82	8.1	89.8	17.7	.88	6.1	65.0	21.4	.45	15.9	69.7	26.2	.87	9.4	6.1–15.9	81.0	22.5	.74	11.5
Middle	77.5	25.2	.64	15.1	60.9	21.3	.31	17.7	39.7	8.9	.75	4.5	37.7	13.6	.90	4.3	4.3–17.7	58.4	25.4	.68	14.4
Late	70.2	27.4	.95	6.1	58.3	15.4	.87	5.6	27.8	10.7	.96	2.1	26.5	18.6	.90	5.9	2.1–6.1	51.0	27.0	.93	7.1
Total	76.1	22.6	.92	6.4	65.8	13.1	.38	10.3	43.9	6.1	.75	3.0	43.0	12.7	.95	2.8	2.8–10.3	61.1	21.1	.77	10.1
PCI																					
Early	100.0	0.0	CND	—	99.1	2.3	.67	1.3	100.0	0.0	CND	—	99.4	1.4	CND	—	—	99.6	1.5	.56	1.0
Middle	96.5	6.4	.58	4.1	95.8	6.7	.80	3.0	83.5	14.2	.70	7.8	75.5	27.8	.67	16.0	3.0 –16.0	90.4	15.2	.86	5.7
Late	97.8	5.3	.50	3.7	97.5	5.4	.99	0.5	64.9	18.0	.75	9.0	74.8	27.2	1.00	0.0	0.0 –9.0	87.2	20.3	.76	9.9
Total	98.1	2.9	.96	0.6	97.5	4.0	.83	1.6	83.7	8.9	.93	2.4	84.3	16.7	.98	2.4	0.6–2.4	92.8	10.3	.87	3.7
PVC	97.5	3.0	.87	1.1	97.1	2.1	.90	0.7	95.7	1.6	.02	1.6	95.7	2.1	.18	1.9	0.7–1.9	96.7	2.4	.80	1.1
PVC-R	98.1	2.6	.82	1.1	98.5	1.6	.61	1.0	96.9	1.8	.76	0.9	96.3	1.9	.74	1.0	0.9–1.1	97.7	2.1	.82	0.9
PPC	93.5	6.6	.89	2.2	90.7	5.2	.94	1.3	80.6	4.5	.70	2.5	79.5	8.1	1.00	0.0	0.0–2.5	87.8	8.3	.96	1.7
PPC-R	95.2	5.5	.86	2.1	94.6	3.1	.84	1.2	84.2	4.1	.70	2.2	82.3	9.2	.95	2.1	1.2–2.2	90.7	7.7	.95	1.7

[a]Transcriber 1 and Transcriber 2
[b]Transcriber 2 and Transcriber 3
[c]Early-8 subscales are identical for the PCC and PCC-A

First, as computed on the total scores for each of the nine metrics, the magnitude of the SEMs are generally adequate for the purposes for which measures of articulation are used in clinical and research contexts. Confidence intervals based on these SEMs should be small enough to retain the clinical and theoretical validity of statistically significant effect sizes.

Second, as computed for the subscales (Early-8, Middle-8, Late-8) of each of the primary consonant metrics (PCC, PCC-A, and PCC-R) the magnitudes of the SEMs are, in some subsets, fairly large. The magnitudes of the SEMs at the subscale level are predictably associated with the difficulty level of sounds and children's articulation competence. Thus, as shown in Table 3, SEMs are lowest for Early-8 consonant sounds, and they are highest for the Late-8 consonant sounds, on which children and adults may have distortion errors that are hard to transcribe reliably. Overall, SEMs are lowest for the predominantly speech-normal children and adults in Subsets 1 and 2 and are highest for the predominantly speech-delayed children in Subsets 3 and 4. However, the highest SEM in Table 3 is 17.7, which was obtained on the Middle-8 sounds for Transcribers 1 and 2 on the ACI scores for normally speaking children in Subset 2. Because of the large role played by distortion errors in the calculation of an ACI and the nonlinearity of the measure, many other SEM estimates for this metric are also considerably larger than SEMs for the other metrics.

Third, and extending the above observation, the considerable influence of narrow phonetic transcription on the magnitude of SEMs is observed across the three PCC measures: PCC, PCC-A, and PCC-R. As shown in Table 3, the SEMs are generally highest for the PCC, which includes all distortion errors. SEMs are generally lowest for the PCC-R, which scores all distortions as correct (see Table 1). As noted above, such standard error of measurement information is central to the interpretation of effect sizes obtained with these metrics. Moreover, as discussed next, such information might play a significant role in the selection of a metric to reflect the articulation competence of a speaker or one or more groups of speakers.

Suggestions for Selecting a Conversational Speech Measure

The primary goal of this report has been to make available, for clinical and research questions, information on the PCC and nine other metrics of articulation competence in conversational speech. This final section includes several suggestions for reliability and validity concerns, including several considerations for selecting a measure to meet specific assessment needs and constraints.

Reliability Concerns

The point-to-point transcription reliability findings in this paper are consistent with earlier literature indicating that reliable transcription of speakers across the lifespan is a primary assessment concern. Until there are well-developed validity data for acoustics-aided procedures and speech recognition technologies, clinicians and researchers must recognize the consequences of reduced reliability associated with the perceptual skill of phonetic transcription. As discussed, transcription disagreement reduces the reliability of speech metrics, which may eventually reduce the validity of clinical and research interpretations and recommendations. Four suggestions address these concerns.

1. *Consensus transcription should be used whenever feasible*. When once-calibrated transcribers are left to transcribe speech over long periods of time, without the benefit of some cross-checking of perceptions with colleagues, a phenomenon termed *transcriber drift* typically occurs (cf. Shriberg & Lof, 1991). Therefore, some type of periodic calibration of transcription skills should be scheduled to increase the reliability of clinical decisions based on phonetic transcription, especially in clinical settings where an examiner's transcription skills may be based solely on training received in an undergraduate phonetics class. When feasible, the optimum arrangement includes the checks and balances obtained when transcription is accomplished using some type of consensus system for two or more transcribers (e.g., Shriberg, Kwiatkowski, & Hoffmann, 1984).

2. *Narrow phonetic transcription should be used as the basis for response definitions*. This suggestion appears to be counter to the findings that reliability is significantly lower in narrow phonetic transcription than broad transcription. The rationale, however, is that the validity of broad transcription is problematic if response definitions have not included percepts developed in narrow phonetic transcription. This concept is well known to clinical phonetics instructors. That is, starting students with the experience of narrow phonetic transcription enhances their ability to differentiate between speech-sound substitutions and speech-sound distortions (e.g., w/r vs. derhotacized /r/; cf. Shriberg, 1995 and Shriberg & Kent, 1995). Although it is difficult to use many diacritics reliably, and although some metrics may not require allophone-level detail, diacritic-level percepts are necessary to identify and discriminate articulatory detail. In the present report, for example, distinctions between common clinical distortions (e.g., dentalized [̪], lateralized [̴])and common nonclinical distortions (e.g., palatalized [ʲ], lengthened [ː]) are fundamental concepts

underlying the speech metrics and eventual clinical classification categories. More generally, the distinction between substitutions for /s/, /l/, or /r/, versus distortions of these sounds, plays a central role in linguistic descriptions of normal and disordered speech. Thus, it is important to underscore the significant role that a transcription system plays in clinical speech measures, including the present measures. To summarize, although some questions in child phonology require metrics that categorize distortions as correct (e.g., PCC-R; see below), narrow phonetic transcription percepts are needed to distinguish among subtypes of distortions and between distortions and substitutions. It is crucial for cross-laboratory comparisons using the present or other metrics that such descriptions be based on the same set of phonetic and diacritic symbols and on common transcription systems providing rules for their use.

3. *Reliability constraints should be addressed when selecting conversational speech measures*. As described in this report, the standard error of measurement of an instrument (which includes terms reflecting the spread of scores on the measure and the reliability of measurement) provides information that is seldom considered in contemporary speech research. Historically, as research in the past few decades has eschewed formal tests and measures in favor of more naturalistic speech samples and individualized linguistic analyses, psychometric issues at the level of test scores have received less attention. Reliability data for metrics based on conversational speech, such as the standard error of measurement, split-half reliability, or test-retest reliability, have been infrequently reported—replaced by point-to-point estimates of interjudge and intrajudge phonetic transcription agreement.

The suggestion here is that the decision process for selecting a conversational speech metric for a given clinical research question should include consideration of the metric's standard error of measurement. Specifically, Table 3 provides SEM estimates for nine metrics, including subscale scores (Early-8, Middle-8, and Late-8 developmental sound classes) for five of the metrics. As reviewed previously, the magnitude of SEMs in Table 3 varies considerably within and among metrics, ages, speech status, and (of course) reliability of the examiners if the reliability term in the SEM is transcription reliability. Although there may be good reason on validity grounds to select a specific measure for a clinical research question (see below), on reliability grounds it may be necessary to select the next best alternative. For example, in order to retain sensitivity to speech change in longitudinal designs, it may be necessary to use total PCC scores rather than Late-8 PCC scores because of the large SEMs associated with Late-8 PCC scores and small expected effect sizes. In large-scale database studies, selection of a measure should be guided by close comparisons of the variances (standard deviation) and SEM for each candidate measure or subscale of a measure, together with estimates of transcriber agreement calculated for each measure.

Validity Concerns

The following four suggestions address validity concerns when selecting and interpreting speech measures from conversational speech samples. For each of the 10 metrics for which reference data and reliability data have been assembled, which metrics are appropriate for certain clinical research questions and subject characteristics? As indicated in Table 4 and described next, selection of an appropriate measure is associated with two considerations: needs and constraints. That is, selection is motivated by the specific needs of the assessment and constrained by the speech status and age/gender status of the speakers.

1. *Select the measure(s) that meet the specific need(s) of the assessment*. Clearly, the primary consideration in selecting one or more of the 10 measures is the specific need of the assessment. As indicated in Table 4: (a) The PCC, PCC-A, PCC-R, and ACI are appropriate when the interest is limited to consonants. (b) The PCI is appropriate to quantify the number of sounds in a child's phonetic inventory, typically for very young or severely speech-delayed children. (c) The PVC and PVC-R are appropriate for a focus on vowels and diphthongs. (d) The PPC and PPC-R are appropriate when there is need for one metric reflecting articulation competence on all speech sounds. (e) The Intelligibility Index is appropriate when the interest is in intelligibility.

2. *Select the measure that meets the age and speech status of the speakers*. As suggested at the outset of the paper, the PCC seems to have been useful for a variety of clinical and research applications with 3- to 6-year-old children who have speech delay. However, when some speakers in a group or comparison groups are older or do not have speech delay, it is not a good metric to represent or to compare articulation competence. In addition to the reliability constraint associated with the transcription of distortions, the PCC has a validity constraint because of the weight given to clinical distortion errors. A speaker with only clinical distortion errors can score as low on a PCC as a speaker with mild speech delay—the latter a more phonologically significant problem involving omissions and substitutions. Moreover, the ACI proposed in Shriberg (1993) also does not usefully meet measurement needs when the speech status of speakers is heterogeneous, because of several problems described earlier, and the ACI may have large SEMs associated with its weighting of distortions. These considerations lead to the three recommendations listed in Table 4.

Table 4. Recommendations for selecting a metric of articulation competence in conversational speech.

Recommendations	Alternative metrics of articulation competence[a]									
	PCC	PCC-A	PCC-R	ACI	PCI	PVC	PVC-R	PPC	PPC-R	II
1. Select the measure(s) that meet the specific need(s) of the assessment:										
a. consonants	✓	✓	✓	✓						
b. phonetic inventory					✓					
c. vowels/diphthongs						✓	✓			
d. consonants and vowels/diphthongs								✓	✓	
e. intelligibility										✓
2. Select the measure(s) that match the age and speech status of the speakers:										
a. 3–6 years; all speech-delayed	✓									
b. diverse ages; all have some speech involvement		✓								
c. diverse ages; diverse speech status			✓							
3. Use subscale scores for increased sensitivity to one or more developmental sound classes:										
a. Early-8	✓	✓	✓	✓	✓					
b. Middle-8	✓	✓	✓	✓	✓					
c. Late-8	✓	✓	✓	✓	✓					
4. Use *z* scores or standard scores to adjust for age or gender differences and for reference comparisons:										
a. *z* scores are more compact	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
b. standard scores are more transparent	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

[a]PCC: Percentage of Consonants Correct. PCC-A: Percentage of Consonants Correct–Adjusted. PCC-R: Percentage of Consonants Correct–Revised. ACI: Articulation Competence Index. PCI: Percentage of Consonants in the Inventory. PVC: Percentage of Vowels Correct. PVC-R: Percentage of Vowels Correct–Revised. PPC: Percentage of Phonemes Correct. PPC-R: Percentage of Phonemes Correct–Revised. II: Intelligibility Index.

First, the PCC is an appropriate selection when all speakers are between 3 and 6 years old and all have speech delay. Reliability constraints associated with consonant distortions notwithstanding, the PCC provides the most information reflecting all three error types: omissions, substitutions, and clinical distortions. Moreover, the severity adjectives validated in the original study—mild, mild-moderate, moderate-severe, and severe—can be used advisedly to characterize severity of involvement at the ordinal level of measurement.

Second, the PCC-A is recommended as an appropriate metric when all speakers have speech involvement, but are of diverse ages. By considering common clinical distortion errors correct, the PCC-A provides a metric of speech involvement that remains sensitive to all distortions excepting common clinical distortion errors. Thus, in a cross-sectional or longitudinal study of speakers ranging from 3 to 18 years old, for example, the PCC-A would be sensitive to speaker differences in the percentage of omissions, substitutions, and all uncommon clinical distortions. Such sensitivity could be necessary to tease out effects otherwise not apparent in the PCC or PCC-R metrics.

Third, the PCC-R is recommended as the most appropriate metric for comparisons involving speakers of diverse ages and of diverse speech status. By scoring all consonant distortions correct, the PCC-R has two advantages over the PCC, PCC-A, and ACI: lower SEMs associated with gains in transcription reliability and greater sensitivity to true involvement because of the focus on deletion and substitution errors. As shown in the reference data in the companion paper (Shriberg et al., 1997), because both normal and delayed speech includes a high proportion of correct vowels, the PCC-R is more sensitive to differences in phonological involvement than the PPC-R. Of course, as above, if vowels/diphthongs were of specific interest, the PVC-R or PPC-R would be recommended for the same reasons.

3. *Use subscale scores for increased sensitivity to one or more developmental sound classes*. As discussed, there

are significant differences in articulation competence when calculated for subscales for the Early-8, Middle-8, and Late-8 developmental sound classes. For certain clinical and research questions, sensitivity to such differences might be necessary to isolate relevant effects. For example, in longitudinal studies and treatment designs, total scores on a metric may not be sensitive to significant changes at the level of one or more of the three developmental sound classes. Use of one subscale score in a univariate design or two or more scores in a multivariate design allows an investigator to more fully exploit the power of the hundreds of tokens obtained in conversational speech samples.

4. *Use* z *scores or standard scores to adjust for age or gender differences and for reference comparisons*. The companion paper (Shriberg et al., 1997) provides lifespan reference data for each of the 10 measures. As discussed previously, use of *z* scores and standard scores is recommended only when raw scores are not appropriate for an applied or research question. In contrast to raw scores, which reflect competence on a criterion-referenced trait or behavior, standardized or normative-referenced scores reflect a person's rank order relative to others' competence on the trait or behavior. In most situations, it is preferable to index criterion-referenced behavior, but in some clinical and research situations standardized scores are useful or required. Such situations include contexts in which it is necessary to compare samples that differ in age or gender and in which there is need to compare performance to some external reference. Moreover, as suggested in Table 4, *z* scores provide a more compact value for computation, whereas differences indicated by standard scores may be more transparent to communicate. The lifespan reference data provided in the companion paper provide only preliminary guidance for these needs. Researchers requiring well-standardized normative data on any of the 10 measures should attempt to collect demographically appropriate samples for specific needs.

Acknowledgments

Preparation of this article was supported by research grants R01 DC00496-08 and R01 DC00528-07 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. We thank the following people who provided research expertise, research assistance, and/or thoughtful discussion about measurement issues associated with the PCC and the SDCS at different stages of this project: Paula Buckwalter, Michael Chial, Mary Elbert, Peter Flipsen Jr., Lisa Freebairn, Frederic Gruber, Barbara Hodson, Doris Kistler, Joan Kwiatkowski, Gregory Lof, Julie Masterson, Karen Pollock, Bruce Pennington, Carmen Rasmussen, Nancy Records, Dorothy Ross, Hollis Scarborough, Nicholas Schork, Bruce Tomblin, Carol Widder, and Xuyang Zhang.

References

Dollaghan, C. A. (1995, June). *Phonological working memory and new phonological learning*. Paper presented at the Symposium on Research in Child Language Disorders, Madison, WI.

Dollaghan, C. A., Biber, M., Campbell, T. (1993). Constituent syllable effects in a nonsense-word repetition task [research note]. *Journal of Speech and Hearing Research, 36*, 1051–1054.

Kent, R. D., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology, 3*(2), 81–95.

Lewis, B. A., & Shriberg, L. D. (1994, November). *Life span interrelationships among speech, prosody-voice, and nontraditional phonological measures*. Miniseminar presented at the Annual Convention of the American Speech-Language-Hearing Association, New Orleans, LA.

Lof, G. L. (1994). *A study of phoneme perception and speech stimulability*. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35*, 259–273.

Shriberg, L. D. (1986). *PEPPER: Programs to examine phonetic and phonologic evaluation records*. Hillsdale, NJ: Lawrence Erlbaum.

Shriberg, L. D. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research, 36*, 105–140.

Shriberg, L. D. (1995, December). *Some perspectives on teaching the undergraduate phonetics course*. Paper presented at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Speech Disorders Classification System (SDCS): Extensions and lifespan reference data. *Journal of Speech, Language, and Hearing Research, 40*, 723–740.

Shriberg, L. D., Gruber, F. A., & Kwiatkowski, J. (1994). Developmental phonological disorders III: Long-term speech-sound normalization. *Journal of Speech and Hearing Research, 37*, 1151–1177.

Shriberg, L. D., & Kent, R. D. (1995). *Clinical phonetics* (2nd ed.). Boston: Allyn & Bacon.

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47*, 256–270.

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

Shriberg, L. D., & Kwiatkowski, J. (1994). Developmental phonological disorders I: A clinical profile. *Journal of Speech and Hearing Research, 37*, 1100–1126.

Shriberg, L. D., Kwiatkowski, J., Best, S., Hengst, J., & Terselic-Weber, B. (1986). Characteristics of children

with phonologic disorders of unknown origin. *Journal of Speech and Hearing Disorders, 51*, 140–161.

Shriberg, L. D., Kwiatkowski, J., & Hoffmann, K. A. (1984). A procedure for phonetic transcription 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 and Phonetics, 5*, 225–279.

Shriberg, L. D., & Olson, D. (1988). *PEPAGREE: A program to compute transcription reliability*. Waisman Center Research Computing Facility, University of Wisconsin-Madison.

Stoel-Gammon, C., & Dunn, C. (1985). *Normal and disordered phonology in children*. Baltimore: University Park Press.

Tomblin, J. B. (1991). *Epidemiology of specific language impairment* (Grant No. N01-DC-1-2101). Bethesda, MD: National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Received March 8, 1996

Accepted November 29, 1996

Contact author: Lawrence D. Shriberg, PhD, The Phonology Project, Waisman Center on Mental Retardation and Human Development, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705. Email: shriberg@waisman.wisc.edu

Copies of the lifespan reference data and additional information are available at the Phonology Project Web site: http://www/waisman.wisc.edu/phonology/

The Percentage of Consonants Correct (PCC) Metric: Extensions and Reliability Data

Lawrence D. Shriberg, Diane Austin, Barbara A. Lewis, Jane L. McSweeny, and David L. Wilson

J Speech Lang Hear Res 1997;40;708-722