Journal of Speech and Hearing Disorders, SHRIBERG & KWIATKOWSKI, *Volume 47*, 256-270, August 1982

PHONOLOGICAL DISORDERS III: A PROCEDURE FOR ASSESSING SEVERITY OF INVOLVEMENT

LAWRENCE D. SHRIBERG JOAN KWIATKOWSKI
University of Wisconsin-Madison

Data are presented to support the reliability, validity, and utility of a severity metric for phonological disorders. The metric, percentage of consonants correct (PCC), is readily derived from a continuous speech sample. PCC values are shown to reflect an ordinal severity scale that embraces the constructs of disability, intelligibility, and handicap. PCC values index four levels of "severity of involvement": Mild, mild-moderate, moderate-severe, and severe. The metric provides a means by which instructors, researchers, and speech-language pathologists working in different settings can specify subject descriptions, gauge the effects of intervention programs, and undertake cross-institutional projects. The metric is used as one component of a diagnostic classification system for phonological disorders (Shriberg & Kwiatkowski, 1982a) and within a framework for management of persons with phonological disorders (Shriberg & Kwiatkowski, 1982b), the first and second papers in this series.

Research and practice in speech-language pathology require procedures that quantify the severity and impact of disordered communication. These studies have been concerned with three conceptually independent constructs; disability, intelligibility, and handicap. As suggested in this brief review, research has yet to yield a widely accepted metric for quantifying the severity of involvement of a person with a developmental phonological disorder.[1]

Disability, as interpreted in legal contexts, is a specification of the degree to which status or performance associated with some characteristic is subnormal. For persons with developmental phonological disorders, a disability typically refers to a reduction in articulatory competence for a specified number of sounds at a specified age. Evaluative reviews of the normative developmental data have been presented by Winitz (1969), Ingram (1976), and Bernthal and Bankson (1981). Critique of theoretical and methodological issues in these normative studies is beyond the scope of this paper. Briefly, these data are inadequately referenced to the neuromuscular dimensions that subserve speech development (Netsell, Note 2), they underestimate the phonetic abilities of children (Ferguson & Farwell, 1975; Hare & Irwin, Note 1), and they are invalid relative to contemporary concepts of phonological comprehension, organization, and production (Shriberg, 1980). Although research activities currently are underway to redress these and other problems, speech-language pathologists currently have only a sketchy outline of normative data against which to determine a child's relative disability in phonological performance.

Intelligibility is a construct of interest to many disciplines concerned with speech perception and production. Intelligibility of speech is influenced by many factors, including characteristics of the speaker, the listener, the social context, the message content, and the transmission media. Metrics to assess the intelligibility of persons with motor speech deficits have been developed (e.g., Yorkston & Beukelman, 1981). However, for clinical assessment of intelligibility, speech-language pathol-

[1]The term *developmental phonological disorder* is synonymous here with the traditional term, *functional articulation disorder.* Consideration of relevant definitional issues is presented in Shriberg and Kwiatkowski, 1982a.

0022-4677/82/4703-0256$01.00/0

ogists typically rely on estimates of the percentage of intelligible words in continuous speech. A simple tally of the number of sounds in error does not adequately index intelligibility because (a) the same pattern of errors becomes more intelligible as a listener becomes familiar with the pattern, and (b) other speaker, listener, context, message, and media characteristics interact significantly with communicative effectiveness. For intelligibility statements, speech-language pathologists currently have only subjective means to quantify the consequences for communication of deficits in phonological performance.

Handicap can be defined as the degree to which a disability impedes effective functioning. Although procedures have been reported for rating the handicap of stuttering, vocal pathologies, and other disorders (Darley, Rees, Siegel, Fay, & Newman, 1979), an instrument developed solely for rating the handicap of a developmental phonological disorder has yet to appear in the archival literature. As reviewed elsewhere, studies indicate that children and adults who make articulation errors do suffer social, educational, and vocational consequences (Shriberg, 1980). However, determination of handicap for the purposes of service delivery questions in the public schools typically is made only by appraisal of the input and concerns of the child, the parents, and other involved parties (Public Law 94-142, 1975). Thus, although the potential handicap of a phonological disorder provides the impetus for research in prevention and management, measurement of handicap in this area requires a form of consensual validation.

OVERVIEW

The present study reports the development of a clinical and research metric that obviates the problems described above, yet validly reflects the three constructs—disability, intelligibility, and handicap. The construct, *severity of involvement*, is proposed as a cover term to embrace these three behavioral domains. Results of a series of listener rating studies directed towards validation of a reliable and efficient procedure follow. Development and validation of an ordinal metric, percentage of consonants correct (PCC), is described. This index will be shown to capture a large portion of the variance in ratings of severity of involvement. Also demonstrated are associations between severity of involvement and age, language, and suprasegmentals. Procedures are described for deriving one of four severity adjectives (mild, mild-moderate, moderate-severe, severe) from a child's PCC value.

Construction and Description of Two Stimulus Tapes

An audio tape library of 60 children recorded in previous studies of children with developmental phonological disorders (Shriberg & Kwiatkowski, Note 3, Note 4, 1980) was screened for potentially acceptable continuous speech samples. Criteria for retention on the first screening were: (a) the child had a developmental phonological disorder defined as delayed speech that was not associated with clinical entities such as retardation, cleft palate, or sensorineural hearing loss, (b) the child was between three and nine years of age, (c) the sample was in the form of continuous speech, rather than words, phrases, or sentences from an articulation test or elicited by imitation, (d) the second author was the clinician/interlocutor, (e) the interlocutor had "glossed," or repeated on the tape, exactly what she believed the child intended to say after each child utterance, (f) the tape was recorded on a well-maintained Uher 400 Report L audio recorder on high-quality, low-noise tape, and (g) the tape did not contain distracting noises such as playground noise, microphone cord movement, and so forth. From the available pool, 6-min speech samples from 30 children were selected which maximally met these seven criteria. Ten 20-sec samples and 30 one-min samples were randomly selected from the 30 speech samples. Within each sample, unproductive clinician questions, off-task clinician talk, and lengthy pause time were removed. These excerpts were dubbed from the original tapes onto an Ampex 456 mastering tape by feeding a Sony TC-270 audio recorder into a Crown 8000 Series audio recorder. The total of 40 samples was then passed through a 150 Hz high pass filter and balanced in average peak intensity to within a ± 1 dB of a calibration tone.

To assess possible order effects, two stimulus tapes were prepared from the master tape. Each stimulus tape included first the ten 20-sec speech samples in an identical randomized order. Sample identification numbers recorded by an announcer were dubbed in with 10 sec of silence inserted between each practice sample. The thirty one-minute speech samples were randomized into two orders. One order was dubbed following the 20-sec samples to become Tape A and the other was dubbed after the 20-sec samples to become Tape B. Identification numbers for each speaker were recorded by an announcer and dubbed in with 10 sec of silence inserted between samples.

To assess several types of rater agreement, two of the one-min samples were randomly selected from the 30 samples. On both Tape A and Tape B, these samples were positioned as Child 7 and Child 11, and were dubbed as "Child 27" and "Child 31," respectively. Therefore, in addition to the 10 20-sec practice samples, Tape A and Tape B each contained a total of 32 one-min speech samples.

Table 1 is a summary of the structural characteristics and certain speech and language characteristics of the speech samples. Procedures for calculating the average words per utterance (AWU) index are described in Appendix A. Because procedures used to obtain the continuous speech samples were not completely comparable to those used to obtain a sample from which a conventional mean length of utterance (MLU) (Brown, 1973) could be calculated, the index average words per utterance (AWU) was used. Moreover, the procedures for determining AWU differed from those described by Brown

TABLE 1. Summary of several structural and content characteristics of the speech samples. Samples were from 24 boys and 6 girls. Description of the AWU and PCC indices are provided in the text and in the appendices.

	Age (in mos)	*Time (in sec)*	*Number of Words*	*Number of Utterances*	*(AWU)*	*Number of Consonants Intended*	*(PCC)*
Mean	69	61	47	9.4	5.3	81	69
Standard deviation	13	4.7	16	1.6	2.0	27	11.5
Range	49-102	54-71	19-76	6-13	1.5-9.3	37-149	42-91

(1973). These values describe only the average utterance length in words within each speech sample; they are not meant to be interpreted as an estimate of a child's language complexity. Further discussions of each of these measures, percentage consonants correct (PCC) and average words per utterance (AWU), occur in context later.

Table 2 is a presentation of additional descriptive information relating characteristics of the stimulus speech samples (one min) to characteristics of the original tapes (remaining five min) and other referential sources. To make these comparisons, AWU and PCC values were calculated for all speech samples following the procedures given in Appendix A and Appendix B. Additionally, frequency and proportion of occurrence data for each of the 24 consonants (that is, sounds attempted) was calculated for both the one-min samples and for each of the original tapes. The strong, positive Pearson correlation coefficients in Table 2 support several important assumptions about the representativeness of the stimulus samples in relation to the PCC index proposed in this paper.

First, both the sample AWU's and the sample PCC's correlate highly with the values for each measure calculated for the original tapes. Recall that the randomly selected one-min stimulus samples were lightly edited to remove lengthy pauses, an operation which may have acted to attenuate the correlational statistics to the obtained value (r = .70). A check on the external validity of the stimulus tapes is provided in the proportional occurrence of consonants data in Table 2. The proportional occurrence of the 24 English consonants on the stimulus tapes correlates highly with the proportional occurrence of English consonants in the original tapes (r = .93), in another sample of children with delayed speech (r = .85; Shriberg & Kwiatkowski, 1980), and in a comparable normative sample of first- through third-grade children (r = .79; Mader, 1954). Moreover, coefficients for all combinations of proportions from these four sources are high and positive. Overall, these data are interpreted as strong support for the assumptions that the one-min speech samples are representative and reliable. Specifically, the stimulus tapes were structurally faithful to their respective original tapes; moreover, the proportional distribution of intended consonants is similar in children with both normal and delayed speech.

The following sections describe four ratings studies which utilized the two stimulus tapes. The design was to obtain subjective ratings of the speech-delayed children on a construct termed *severity of involvement*, including raters' anecdotal reasons for their ratings. Data from three other studies using the stimulus tapes are then used to derive objective correlates for the severity of involvement ratings.

Severity of Involvement Ratings

Group I: Clinician Sample

Description. Persons were recruited from two sources to rate the "severity of involvement" of children on the stimulus tapes. The first was a mailing list of 47 public school clinicians who were acting as cooperating practicum supervisors for the University of Wisconsin-Madison, Department of Communicative Disorders. A letter describing, in general terms, purposes and procedures for a study of speech delayed children was sent to each person. Thirty clinicians (64% of those contacted) agreed to participate in the listening-rating task.

The second source of potential raters was a series of workshops on phonological disorders given in Wisconsin

TABLE 2. Descriptive statistics for the 30 speech samples on the stimulus tapes in relation to the original, full-length samples, samples from another group of children with delayed speech, and normative data.

Variable	$\bar{x}$	*SD*	*r*
Average Words Per Utterance (AWU)			
Stimulus Samples	5.31	1.96	
Original Samples	4.35	1.55	
Stimulus/Original			.70
Percentage of Consonants Correct			
Stimulus Samples	69.3	11.5	
Original Samples	68.3	10.3	
Stimulus/Original			.83
Proportional Occurrence of 24 Intended Consonants			
Stimulus/Original			.93
Stimulus/Delayed Speech Group*			.85
Stimulus/Normal Group**			.79
Original/Delayed Speech Group			.85
Original/Normal Group			.91
Delayed Speech Group/Normal Group			.77

*Proportional occurrence of 24 target consonants in continuous speech samples from 10 children with delayed speech (Shriberg & Kwiatkowski, 1980; Table 2).

**Proportional occurrence of 24 target consonants in continuous speech samples from 81 first-third grade children with normal speech (Mader, 1954).

and Illinois. With knowledge of the general purposes of the study, 22 clinicians agreed to participate as listener-raters.

Table 3 describes the composition of these combined groups. The 52 clinicians came from various institutions, had a wide range in number of years of professional expertise and represented a range of experience with young children with delayed speech (by anecdotal reports).

TABLE 3. Descriptive statistics for the 52 speech-language pathologists who rated the stimulus samples on the construct "severity of involvement."

Stimulus Tape	*n*	*Sex* M	F	*Years of Paid Professional Experience* $\bar{x}$	*SD*	*Range*	*Number of Different Cities/ Towns Represented*
Order A	26	4	22	6.8	6.0	<1-24	15
Order B	26	2	24	7.7	7.2	<1-28	17
Total Group	52	6	46	7.3	6.6	<1-28	28

Materials and Procedures. Each volunteer clinician received by mail a package containing a cover letter, a sheet of instructions, an audio cassette copy of either Tape A or Tape B (according to a counter-balanced order), a booklet of response forms (keyed to either Tape A or Tape B) and a stamped, self-addressed return envelope. Appendix C includes a copy of the first page of the response forms booklet and *Instructions to Speech-Language Pathologists.* It is important to inspect these materials closely to understand the "set" the clinicians were given concerning the nature of their rating task as well as the procedural details. Briefly, it was explained that they were to rate the "severity of involvement" of 32 children with "delayed speech." Information on age and sex of each child was provided. They were to assign a severity rating from "3" to "7," which, with intermediate values, yielded a nine-point, equal-appearing interval scale. After making practice ratings on the ten 20-sec samples, they were to listen to each of the 32 one-min samples only once before making their ratings. At the conclusion of the task, they were to list in the space provided, the factors on which they based their ratings of "severity of involvement." The total period of the Group I study, from receipt of the first package by a clinician to the return of the last package, was eight weeks.

Group II: Student Sample

Description. In a course entitled Introduction to Communicative Disorders, 120 students volunteered for a special rating session. When the rating session was held, the sixth week of class, the students were familiar with basic terms and concepts within communicative disorders, but developmental phonological disorders had not been discussed. Approximately 50% of the sample were Freshmen; approximately 85% were female. Students' intended majors ranged broadly across the humanities and sciences, although approximately 50% intended careers in the allied health sciences.

Materials and Procedures. A listening-rating session was held in a large classroom. Instructions were given orally. Students were told that they would be listening to a series of children with "delayed speech." Their task was to rate each child's "severity of involvement." A copy of the same response form used by clinicians described above was passed out; however, the form was abbreviated to allow responses only to the first 15 children. In addition to providing the demographic data described above, students were asked to indicate on their response form whether they were seated in the front or the back half of the classroom (as delineated by the instructor). A copy of Tape B was played to students using a Marantz Model C-105 audio cassette recorder and auxiliary high-quality speaker system. As with the Group I listeners, students first made practice ratings for the ten 20-sec practice samples; they then rated the 15 one-min speech samples. During the first few practice samples, audio intensity was adjusted for audibility and clarity for persons seated at extreme seats from the speaker. Following the listening session, 10 min was provided for students to list the bases for their ratings of "severity of involvement." The entire rating session took 45 min.

Percentage of Intelligible Words Data

Description of Sample. Volunteer judges for a "listening task" were recruited from a senior methods course in communicative disorders and from a group of first-year graduate students. A total of 14 students, 10 seniors and 4 graduate students, volunteered for an intelligibility task that was expected to take several hours.

Materials and Procedures. Intelligibility data were obtained in a 15-station, listening-viewing laboratory. Each student was seated in an individual booth equipped with Koss Pro-4A headphones and a call button that lit a lamp at a central console. Judges were informed of the goals of the study and that their role was to provide intelligibility data on 32 speech samples. A 32-page response form booklet was passed out to each student judge. Each page in the booklet contained a child sample number and boxes for entering a gloss for successive utterances. That is, their task was to enter on the response form the exact words they understood the child to say.

Stimulus Tape B was played on a Sony Model ER-740 audio cassette recorder which fed the earphones in the 14 booths. Comfortable listening levels were set by each clinician on the headphones and amplifiers in each booth during the first few of the 10 practice samples. Clinicians were then informed that the tape would be stopped immediately after each child utterance. The procedure was to write the words that they understood the child to say for each utterance and to press the call button when finished. Practice in using the call button was then given on a series of utterances in two practice samples. The experimenter at the console had transcripts of the speech sample available, thus enabling the tape to

be stopped immediately after each child utterance. Because the interlocuter on the tape glossed each previous utterance, clinicians were continuously updated on each utterance; that is, immediately after clinicians had entered their own gloss for an utterance, they had the opportunity to hear the interviewer's gloss. This procedure was used for obtaining intelligibility data for all 32 samples, with several rest breaks interspersed. The entire listening session took three hours. These data were later reduced and averaged to yield a percentage of intelligible words for each sample.

Suprasegmental Ratings

Description of Sample. Ten persons were recruited to rate the suprasegmental characteristics of the speech sample. These persons were currently in a practicum experience at a diagnostic-treatment center in communicative disorders. Most had participated in an in-service workshop on phonological disorders which had included a unit on suprasegmentals. Among this group, one person had a doctorate in communicative disorders, one was a doctoral candidate, one was an experienced clinical supervisor, and the remaining seven persons were second-year master's students in communicative disorders.

Materials and Procedures. Ratings were accomplished in two stages, a group rating session and individual rating sessions. During the group session, the general goals of the study were described orally. Participants were then given response forms containing spaces for rating the 32 speech samples on six suprasegmentals. They were divided into two superordinate categories, voice and rhythm. The three voice suprasegmentals included pitch, loudness, and quality; the three rhythm suprasegmentals included rate, stress, and phrasing. Instructions for use of a 0, 1, 2 rating system were provided on a separate form, a copy of which is presented as Appendix D. Pilot studies of several alternative rating systems for these six suprasegmentals indicated that judges had difficulty rating the magnitude of deviation from normal, that is, the degree to which a suprasegmental is abnormal. They were more consistently able to decide whether a perceived level of deviance from normal (normal = 0) occurred infrequently (1) in a one-min speech sample, that is, on only one or two utterances, or occurred frequently (2) throughout the sample. Hence, the 3-category system was chosen to reflect the persistence of suprasegmental behaviors throughout a sample. The system does not attempt to index the magnitude of deviation from normal. Preliminary interjudge and intrajudge reliability studies indicated that reliable suprasegmental ratings, like other perceptual rating data in the clinical literature (e.g., hypernasality), would require averaged ratings from a panel of listeners.

The group rating session proceeded as follows. Two copies of Tape B were used to present each speech sample on a Sony Model TC-2 cassette recorder oriented to a semi-circle of the 10 listeners. During the first few practice samples, intensity was adjusted for maximum comfort and quality for each of the listeners. For the first play of each practice sample, listeners made notes in spaces provided on the response forms for each of the six suprasegmentals. After the second play, they entered a 0, 1, or 2 for each of the six suprasegmentals. The first 15 samples were rated in this manner in a session lasting approximately one hour. A copy of Tape B was then given to each judge. Judges were instructed to complete the remaining samples in the same fashion (two plays for each sample) during the next week, using whatever cassette recorder they normally used in the clinic.

RESULTS

Overview

Analysis goals were (1) to determine if severity of involvement ratings reflect the constructs of disability, intelligibility and handicap and, (2) to determine if the PCC metric was, by itself, an adequate index of severity of involvement.

Data analyses were divided into two phases. First, reliability and item-level analyses were accomplished for each of the four data sets—the PCC data, the Severity of Involvement ratings, the Percentage of Intelligible Words data and the Suprasegmental ratings. Second, taken together with other information, all measures were analyzed in several correlation, partial correlation, and multiple correlation models to parcel out significant components of variance in the construct, severity of involvement.

Reliability and Item-Level Analyses of All Measures

Percentage of Consonants Correct (PCC)

Intrajudge Reliability. Calculation of the percentage of consonants correct was done by one person with considerable experience in the transcription of children's speech (JK). Because these children were known to be speech delayed, the response definition was "score as incorrect unless heard as correct." Intrajudge reliability for the PCC measure was assessed by having the judge rescore all 30 stimulus tapes approximately five weeks after the original scoring. The Pearson correlation coefficient between ratings was $r = .97$. Average differences in percentages for each child were 2.1 percentage points with a standard deviation of 1.73 and a range of no difference to differences of five percentage points.

Interjudge Reliability. Interjudge reliability for the PCC values was assessed by having two graduate students studying communicative disorders score five samples randomly selected from the 30. The only constraint on sample selection was that at least 50 target consonants were to be scored by the criterion judge (JK) and the two reliability judges (J_1, J_2). J_1 and J_2 followed the procedures for determining a PCC value as described in Ap-

pendix B. Obtained PCC values for the five samples by the three judges (presented in the order: JK, J_1, J_2) were: Sample 1: 78, 80, 89; Sample 2: 82, 88, 89; Sample 3: 72, 86, 76; Sample 4: 65, 74, 80; Sample 5: 89, 90, 94. The greatest disagreement among any two of the three judges (15%) occurred for Sample 4, whose average of only 59 target consonants scored maximized differences in obtained PCC values (i.e., the smallest denominator). Disagreements were more often due to differences on the gloss of utterances than on whether similarly glossed consonants were correct or incorrect.

Internal Stability. Internal stability of the PCC values was assessed by comparing a sample of one-min values to values calculated for the cumulative minutes in the original five-min sample. Twelve of the original five-min samples were randomly selected for this purpose. The criterion judge calculated PCC values for each minute of tape play; she also kept a cumulative count of the number of target consonants per minute. The results are presented in Table 4. Note that for each child, the PCC values for the first min vary within a maximum of 10 percentage points from values calculated for successive minutes. These data, which are consistent with the correlational findings in Table 2, suggest that a sample of approximately three min will yield about the same PCC value as a sample of five min. Overall, the continuous speech samples average a little over 60 target phonemes per minute, although variability within and across speech samples is considerable.

Summary. These data support several forms of reliability for the PCC metric. Moreover, the interjudge reliability data indicate that the instructional content for teaching clinicians how to derive a PCC value, as presented in Appendix B, is functionally sufficient. Appendix B also includes a summary table (Table 9) containing selective statistical data on continuous speech samples.

Severity of Involvement Ratings

Effect of Order of Presentation. The Pearson correlation coefficient between mean ratings of the 30 children by the 26 Order A judges and the 26 Order B judges in Group I was $r = .96$. The differences in the magnitude between each set of ratings means averaged only .30 scale points. These data indicate that order of presentation of the speech samples was not associated with severity ratings. For the purposes of the second phase of the analysis, therefore, ratings from all 52 judges were calculated to yield one overall mean rating for each speech sample.

Interjudge Reliability. Interjudge agreement for severity ratings in Group I and Group II was not formally assessed because differences in the judges' interpretation of the construct "severity of involvement" were expected. An overall estimate of interjudge agreement is provided by the average standard deviation and range of ratings in each group. For Group I, the speech-language clinicians, the average standard deviation across the 32 speech samples (including the two reliability samples) was .87 scale points. For Group II, the average standard deviation was .76 scale points. Thus, for both groups, the majority (66%) of judges rated speech samples within 1.52-1.74 scale points of one another on the nine-division scale.

Intrajudge Reliability. Severity ratings for the two children whose speech samples appeared twice on the stimulus tape were inspected for each judge. Of the 104 sets of ratings in Group I (two children × 52 judges), 47.3% of judges rated children with the same scale value on the second listening as on the first, and an additional 39.7% of judges rated children within ±1 scale point value on the two occasions. These data support the relative stability of judges' ratings throughout the task.

TABLE 4. Internal stability of the PCC values in 12 randomly selected continuous speech samples. Total number of consonants and the PCC values are cumulated at the end of each minute of the 5-min speech samples.

	Five-Minute Continuous Speech Samples									
	First		*Second*		*Third*		*Fourth*		*Fifth*	
*Subject**	*Total No. Consonants*	*PCC*	*Total No. Consonants*	*PCC*	*Total No. Consonants*	*PCC*	*Total No. Consonants*	*PCC*	*Total No. Consonants*	*PCC*
1	63	37	157	44	218	44	308	43	404	44
2	31	42	77	45	101	46	143	46	177	45
3	64	58	160	56	256	55	313	56	390	54
4	50	60	101	60	142	60	179	59	238	62
5	70	63	99	65	150	60	194	59	247	57
6	45	64	97	71	178	75	259	72	327	73
7	33	67	92	65	145	66	209	66	312	64
8	105	68	187	70	274	70	370	72	518	73
9	60	70	94	69	160	71	231	65	294	68
10	69	74	116	66	193	63	233	64	271	65
11	62	79	137	72	201	75	258	77	340	76
12	84	81	149	80	239	81	312	81	349	82
$\bar{x}$	61.3		122.2		188		250.8		322.3	
SD	20.6		34.6		51.9		65.8		89.5	

*Subjects are arranged in increasing PCC values for the first minute.

Comparison of Ratings in Group I and Group II. Ratings for the first 15 samples obtained from the 52 experienced clinicians in Group I were compared to the ratings of the children by the 120 college students in Group II. The obtained Pearson correlation coefficient (r = .86) indicates that the two ratings were highly similar. Inspection of the magnitude of the ratings' values indicates also that ratings were virtually identical between these two groups. Thus, experienced clinicians and essentially naive listeners were similar in their ratings of these speech samples on the construct "severity of involvement."

Anecdotal Rationale for Severity Ratings in Group I and Group II. Analysis of the judges' written rationale for their ratings was accomplished in similar fashion for the 52 experienced clinicians (Group I) and the 120 college students (Group II). An item analysis format was developed which tallied for each judge, the number and rank-ordered occurrence of factors cited as determinants of their ratings. For example, if the phrase "how well I could understand the child" was listed first by a judge, it received one tally and was entered as that rater's most important factor (rank = "1"). Superordinate categories were readily suggested for the diversity of phrases and terms most often used. For example, phrases such as the one quoted above were subsumed under the category label "intelligibility." A category titled "Suprasegmentals" was derived for raters' terms such as "voice characteristics," "intonation," "tone of voice," "voice quality," and so forth. The college students used more colloquial terms and phrases (e.g., "pronunciation," "level of cooperation with the lady"), but these comments, too, could be readily subsumed by higher-order technical terms. Some raters in both groups listed as many as two pages of comments and as many as 15 variables on which they purportedly based their severity ratings. Some raters in both groups organized their comments into binary decision trees for rating the severity of involvement of each sample.

Although the lists of criteria and decision logics for rating "severity of involvement" were diverse in length and form, a rank-ordered list of the same five superordinate terms emerged from both studies. These factors are listed in Table 5. A scattering of other factors, such as "enthusiasm," "child's interest in answering questions," "child's level of fluency," were cited by a few clinicians. However, the five factors listed and rank-ordered in Table 5 were the only ones cited by more than 10% of all judges.

Intelligibility Ratings

Interjudge Reliability. An estimate of the interjudge agreement for the intelligibility ratings accomplished by the 14 speech-language clinicians can be obtained by inspecting the average standard deviation and range of standard deviations across the 30 speech samples. These data indicate that, on the average, 66% of judges agreed with one another on the percentage of intelligible words within ±8.8 percentage points of the mean value for each child; the range of standard deviations was ±4.4 to ±14.3 percentage points.

Intrajudge Reliability. Intelligibility percentages for the two children whose samples were repeated were inspected to determine the intrajudge agreement for the intelligibility ratings. Across the 28 pairs of intelligibility percentage comparisons (two children × 14 judges), the average difference between the first and the reliability ratings was 5.7 percentage points. These data indicate that judges were reliable for the purposes of this study.

Suprasegmental Ratings

Interjudge Reliability. As expected from preliminary studies, agreement on the use of 0, 1, and 2 was not high among judges for the six suprasegmental categories. Across the 180 sets of ratings (six suprasegmentals × 30 samples), percentage of exact agreement on the use of 0, 1, or 2 among the 10 judges ranged from 40% to 100%, with a mean of 66.5%. Hence, for any one suprasegmental, the majority of judges were, on the average, in agreement. Overall, judges agreed more on ratings of

TABLE 5. Rank-ordered importance of factors underlying "severity of involvement" ratings as cited by raters in Group 1 and Group 2.

Group 1: Speech-language pathologists (n = 52)			*Group 2: Undergraduate Students (n = 120)*		
*Overall Rank**	*Factor*	*Percentage of Citations at Rank or Above*	*Overall Rank*	*Factor*	*Percentage of Citations at Rank or Above*
1st	Intelligibility	79	1st	Intelligibility	83
2nd	Age	64	2nd	Age	52
3rd	Articulation	65	3rd	Language	56
4th	Language	54	4th	Articulation	39
5th	Suprasegmentals	44	5th	Suprasegmentals	14

*Overall rank was determined by assigning "1" to the factor cited most frequently as the most important determinant of severity ratings, assigning "2" to the factor cited most frequently as either the first or second most important determinant of severity ratings, and so forth. Hence, the percentage of citations at rank or above values are in some places, higher for factors ranked lower overall.

loudness (mean agreement = 80% of judges); however, more children received the numeral 0 on this suprasegmental than any other. Comparison of the mean and modal ratings among the three categories for each of the six suprasegmentals indicated that mean ratings would best represent the central tendency of the group's ratings. Accordingly, mean ratings were used for the second phase of the data analysis.

Intrajudge Reliability. Intrajudge characteristics were assessed in two ways. First, for each suprasegmental, a judge's use of 0, 1, 2 across the 30 samples was compared to the mean and standard deviation of the total group of 10 judges. Each judge's "underuse" and "overuse" of 0, 1, or 2 was then determined by plotting the number of times for the six suprasegmentals that ratings were beyond ±1 standard deviation from the mean. In this way, it was determined that certain judges tended to overuse and/or underuse one or more of the three values. These data on intrajudge characteristics partially explain the range of interjudge agreements across the six suprasegmentals as described above. Specifically, these data demonstrate that 8 of 10 judges were biased towards overusing or underusing 0, 1, or 2 across the six suprasegmentals. For example, some judges rarely rated children a 2 on any of the six suprasegmentals, while other judges consistently rated suprasegmentals 1.

The second form of intrajudge reliability inspected each judge's ratings on the second occurrence of the two speech samples. Across the 12 mean percentage of agreement figures (two children × six suprasegmentals), the average was 77% agreement in the use of either 0, 1 or 2 on both occasions. Comparison of the two children indicated that agreement was tied to the actual behaviors of the child. As is customary, test-retest agreement was highest at the extremes of the scale, in the present case, when the child's actual behaviors clearly were a 0 or a 2.

Effect of Order of Presentation. Mean suprasegmental ratings for the two children whose samples were repeated also afford a check on the possible effect of order of presentation of samples. These two samples were positioned relatively early in the order of samples (Child 7 and Child 11) and were repeated relatively late in the tape (Child 27 and Child 31). The means for the 12 pair-wise comparisons (two children × six suprasegmentals) varied from perfect agreement to differences of .6 of a scale point. Overall, the average of the mean differences in ratings was .187 or approximately two-tenths of a scale point. These data indicate that the central tendency value did not shift as judges progressed through the stimulus tapes. For these two children, the three voice suprasegmentals were slightly more stable (mean retest difference = .13 scale points) than the three rhythm suprasegmentals (mean retest difference = .28 scale points).

Effect of Group/Individual Ratings. Finally, the design allowed for inspection of the possible effects of the group session, in which an audio tape recorder was used, compared to individually completed ratings, in which personal tape recorders were used. Inspection of the interjudge agreement data tallied separately for the first 15 sets of ratings (group session) and the last 15 sets of ratings (individual) failed to indicate observable differences in interjudge agreement as calculated above.

Statistical Description of Associations Among Measures

The goal of the second phase of the data analysis was to develop an explanatory model for the construct, "severity of involvement." Given the modest number of speech samples in relation to the number of potentially interesting independent variables, correlation, partial correlation, and multiple correlation analyses were deemed to be appropriate and adequate for this goal. Arcsin transformation of all percentage data (PCC and intelligibility scores) were performed, although they were not strictly necessary, given that the values were well distributed and that no percentages were "0" or "100." Inspection of the correlation coefficients calculated with both untransformed percentages and with the Arcsin transformation scores indicated that the two yielded virtually identical values.

TABLE 6. Zero-order Pearson correlation coefficients among 12 descriptive variables for the 30 speech samples.*

	Pitch	*Loudness*	*Quality*	*Phrasing*	*Stress*	*Rate*	*Age*	*Sex*	*AWU*	*PCC*	*Intelligibility*
Loudness	−.05										
Quality	.45	−.02									
Phrasing	−.03	−.30	.10								
Stress	.33	.09	.00	.31							
Rate	.22	.12	.12	.63	.34						
Age	.21	.23	.12	.03	.01	.12					
Sex	.13	.13	−.04	−.01	.13	.27	.06				
AWU	−.31	−.13	.08	.52	.09	.18	.03	−.25			
PCC	−.36	.16	.09	−.10	−.18	−.03	−.07	−.16	.20		
Intelligibility	−.23	−.19	−.42	−.27	.05	−.16	−.07	.08	−.24	.42	
Severity	.45	−.01	.25	.28	.24	.31	.43	.13	−.01	−.62	−.74

*Higher values on the six suprasegmentals and severity ratings scales corresponded to poorer performance; higher values on the PCC index and intelligibility corresponded to better performance.

Correlational Findings

Table 6 is a summary of the zero-order correlational data for the 12 variables to be studied. Only certain of the intercorrelations are logically of interest.

The first set of coefficients that warrant inspection are intercorrelations among the six suprasegmentals (upper 15 coefficients). High positive intercorrelations between any two suprasegmental variables would indicate either that children tend to exhibit similar ratings on these variables or that the two suprasegmentals might be assessing a common factor. Overall, these 15 coefficients suggest that neither was the case. The six variables are not highly intercorrelated, with the exception of the ratings for phrasing and rate (r = .63) and pitch and quality (r = .45) which share 40% and 20% (r^2) of common variance, respectively. None of the other variables share more than 12% common variance. These data indicate that the six suprasegmental ratings may be viewed as essentially independent aspects of the suprasegmental domain and, therefore, that they should be retained as individual factors in subsequent analyses.

The second set of entries in Table 6 that warrant attention are the coefficients for each of the 11 independent variables with the severity ratings. As indicated across the bottom row in Table 6, Intelligibility (r = −.74) and the PCC index (r = −.62) are most highly associated with severity ratings, sharing 55% and 38% common variance, respectively, with severity ratings. Sex, AWU, and loudness are essentially uncorrelated with severity ratings, with the remaining variables sharing less than 20% common variance with severity ratings.

These coefficients provide a quantitative parallel to the rank-ordering of factors underlying the "severity of involvement" judgments of the clinicians (Group I) and the students (Group II). Recall (see Table 5) that both groups included among the bases for their severity judgments: Intelligibility, Age, Language, Articulation, and Suprasegmentals. Four of these five factors (assuming that PCC reflects "Articulation" competence) are correlated with the severity ratings listed in Table 6. Moreover, their ordering in terms of the magnitude of association (r) roughly parallels the rank-ordered anecdotal data in Table 5. The lack of association of the AWU index with Severity ratings may reflect two considerations. First, AWU values in these one-min samples were only moderately correlated with AWU in the original samples (.70), most plausibly because of sample homogeneity, because pause time was removed for efficiency, and because of sampling error. Second and more importantly, AWU (and MLU) is just one index of language performance. Interestingly here, a moderate positive association between AWU and Phrasing (r = .52; Table 6) suggests that increased AWU may be costly for this suprasegmental, a possibility proposed by Shriner, Holloway, and Daniloff (1969). In any case, perhaps other sorts of analyses in the syntactic, semantic, or pragmatic domains could tease out the relative contributions of "language" variables to raters' perceived severity of involvement ratings for speech-delayed children.

As developed at the outset of this paper, intelligibility ratings are problematic for clinical purposes because of a variety of speaker, listener, context, content, and media factors. As shown in Table 6, the intelligibility data were only moderately correlated with the articulation proficiency index, PCC (r = .42; r^2 = 18%); in fact, this association is no higher than the correlation of intelligibility with quality. These data affirm the assumption that speech intelligibility reflects a complex of factors in addition to articulation proficiency. For the goals of this paper, then, data analysis attempted to determine the independent and summative contribution of all other independent variables to the severity of involvement ratings. That is, we attempted to determine how accurately a child's severity of involvement rating could be predicted by factors other than intelligibility.

Partial Correlation and Multiple Correlation Analyses

To assess the independent associations of each of the eight independent variables with severity ratings, coefficients were calculated with the effects of all other variables removed. These partial correlation coefficients are summarized in Table 7. These data suggest that approximately 78% of the variance in severity ratings was accounted for by children's status on these eight variables. A child's PCC value and age combined to account for nearly 65% of the variance, with quality and rate ratings contributing the majority of the additional variance. The assumption is that the remaining sources of variance (approximately 22%) are to be found in language variables not captured in this study, and in other variables mentioned occasionally by clinicians and students in their anecdotal comments.

Multiple correlation analyses were computed to confirm statistically the additive effects of each source of variance in determining severity ratings. Table 8 summarizes these data for an analysis of variance by regression. When the eight factors are placed in a 4-step model, with the suprasegmentals divided into the two

TABLE 7. Partial correlation coefficients for eight independent variables with severity ratings.

Variable	*Partial Correlation rp*	*Percentage of Variance Accounted For*
Percentage of Consonants Correct (PCC)	−.6658	43.01
Age	.4688	21.98
Pitch	.0227	.05
Loudness	.0166	.03
Quality	.2898	8.40
Phrasing	.0963	.93
Stress	.0559	.31
Rate	.1666	2.78
Total		77.49%

major groups, voice and rhythm, a significant proportion of variance is gained at each step. The final multiple correlation between all variables and severity ratings is .81. Again, the assumption is that the remaining sources of variance are to be found in language variables and in the other variables mentioned occasionally by raters.

TABLE 8. Multiple correlation coefficients for eight independent variables with severity ratings at each of four steps in an analysis of variance by regression.

Step	Variable	*Multiple Correlation (r)*	*F*	*df*	*p*
1	Percentage of Consonants Correct (PCC)	.6199	17.48	1,28	< .001
2	Age	.7326	15.64	2,27	< .001
3	Voice: Pitch, Loudness, Quality	.7758	7.26	5,24	< .001
4	Rhythm: Phrasing, Stress, Rate	.8115	5.06	8,21	< .005

Classification Analysis

To this point, the data indicate that although a child's age and, to some degree, his or her suprasegmental characteristics influence severity ratings, the PCC index (a measure of articulation proficiency) clearly is the major predictor. On the strength of these correlational data, an attempt was made to classify the severity data purely on the basis of PCC values. That is, were the PCC values "robust" enough to correctly classify a significant proportion of children on an ordinal scale of severity of involvement? Figure 1 is the crossplot that resulted from several "best-fit" trial solutions.

As illustrated in Figure 1, the severity ratings are classified and labeled as four ordinal divisions, based on the terms used on the original scale: mild = 3-3.5; mild-moderate = 3.5-5.5; moderate-severe = 5.5-6.5; severe = 6.5-7.0 (see scale in Appendix C for association between these adjective descriptors and scale point labels available to raters). As shown in Figure 1, the severity classification of 20 of 30 speech samples (including data points that fall on classification boundaries) are accurately classified ("hits") when the PCC index is parcelled into four sectors: mild = 85-100%; mild-moderate = 65-85%; moderate-severe = 50-65%; and severe = less than 50%. An additional three samples (Child 2, 7, 9) fall within .05 of a rating scale point of the correct classification and an additional three children (Child 12, 14, 17) fall within 3 percentage points on the PCC of their severity ratings. That is, the severity ratings of Child 7 and Child 9, for example, place them as mild-moderate, whereas the PCC cut-off values would convert to the adjective—mild. Similarly, Child 2's severity rating was moderate-severe, whereas the PCC cut-off value would convert to a rating of mild-moderate. These values for the scores which fell just outside of the category boundaries are well within the measurement error of the severity ratings. Moreover, the misclassifications by up to 3% points on the PCC are well within the error of measurement on these one-min samples (see Table 3 for internal stability data). Taken together, 27 of the 30 children's severity ratings (90%) can be accurately or reasonably accurately predicted by a child's PCC value alone. Only three children (Child 13, 20, 26) were grossly misclassified. The data sets for these children were examined to explore why the PCC index might have failed to classify them appropriately.

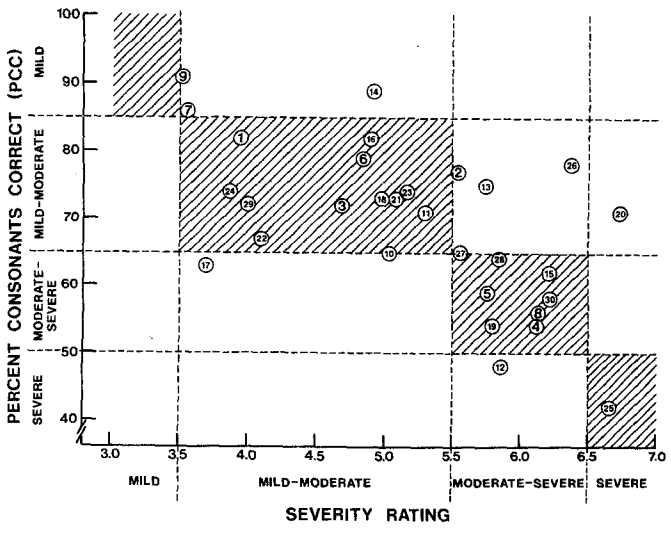


FIGURE 1. Classification of the 30 speech samples into four severity adjectives (mild, mild-moderate, moderate-severe, severe) by means of their Percentage of Consonants Correct (PCC) values.

Inspection of Figure 1 indicates that for each of the three children whose PCC values did not yield a correct severity adjective, the situation was similar; each was rated lower by the severity judges. Child 13 and Child 26, whose PCC's of 75% and 78% would convert to severity values of mild-moderate, were actually rated as moderate-severe; Child 20, whose PCC of 71% would convert to mild-moderate, was actually rated as severe. What factors in the data set might have accounted for a lower rating in each of these three cases? Two were suggested which are consistent with the statistical analyses just presented. Child 13 and Child 26 were the third- and second-oldest children, respectively, of the 30 children. Child 20 and Child 26 received the poorest (1.9) and the second poorest (1.4) average ratings, respectively, on the suprasegmental Phrasing. These two factors—age and suprasegmental performance, could have influenced raters towards more severe involvement than indicated by these children's articulation performance as quantified by the PCC measure. Age, as indicated by anecdotal comments, was considered by raters—older children were considered more severely involved. In a similar fashion, poor suprasegmental performance, as evidenced here for Phrasing, was cited by judges as important in their ratings. Child 20, who was rated as "severely" involved, had a PCC value of 71 which converts to mild-moderate. Inspection of his AWU

and suprasegmental ratings sheds light on this clear "miss," however. His AWU of 8.1, third highest in the group, indicated that he spoke in long utterances. However, his rhythm suprasegmental ratings overall were the poorest in the group: Phrasing = 1.9; Stress = 1.4; Rate = 1.4. For this child, the assumption is that his rhythm problems were so pronounced in his longer utterances that he received severity ratings well below his segmental performance (PCC).

DISCUSSION

The goal of this study was to develop a procedure to assess the severity of a developmental phonological disorder. Results suggest that a procedure may be followed that appears to have construct validity, reliability, and clinical utility. Specifically the procedure yields a severity description from mild to severe, that captures the quantitative and qualitative correlates of disability, intelligibility and handicap (construct validity). The procedure is also based primarily on an articulation task which requires only correct-incorrect judgments of a continuous speech sample, as opposed to phonetic transcription (interjudge reliability, intrajudge reliability, sample stability, and internal stability). It can be used repeatedly with the same child by one examiner or clinician for clinical or research purposes (utility).

The recommended procedures to classify a child's delayed speech as mild, mild-moderate, moderate-severe, or severe may be summarized as follows.

1. Tape record a continuous speech sample of a child following sampling procedures such as those described in Shriberg and Kwiatkowski (1980). Any means that yield continuous speech from the child are acceptable, provided that the clinician glosses each child-utterance immediately. The clinician can tell the child that his exact words will be repeated onto the "tape machine" so that the clinician is sure to "get things right." Children adapt to this very rapidly if the clinician is skillful at conversing with children.
2. Calculate a Percentage of Consonants Correct from the audio tape following the procedures described in Appendix B. Assign the appropriate severity adjective.
3. Score the six suprasegmental variables from the audio tape following the procedures described in Appendix D.
4. The PCC rating may not accurately reflect the child's perceived severity of involvement to the extent that (a) the child is considerably older, (b) the child's suprasegmentals are markedly involved, (c) the child's interpersonal and/or language performance are markedly deviant. Each of these factors can be weighted to lower the severity adjective assignment, but generally only if the PCC value is close to the bottom of the range for the assigned severity descriptor (Child 20 was an exception). For most children, the PCC value alone should accurately index their perceived "severity of involvement" as defined in this paper.

ACKNOWLEDGMENTS

This study was made possible with the volunteer efforts of a great many people. We are grateful to all raters in the groups, the clinic and public school speech-language pathologists in Illinois and Wisconsin, the undergraduate, masters, and doctoral students at the University of Wisconsin-Madison and colleagues in the Department of Communicative Disorders and at the Harry A. Waisman Center on Mental Retardation and Human Development. Many individuals in these groups provided significant input beyond their specific ratings tasks. We would also like to express thanks to Robin Chapman, Jon Miller, and Anne Smith for their very helpful comments on the draft of this manuscript. Finally, we also thank the administrative and secretarial staff of the Department of Communicative Disorders for their competent support throughout these studies.

REFERENCE NOTES

1. HARE, G., & IRWIN, J. *Consonant acquisition in children aged 21-24 months.* Paper presented at the American Speech-Language and Hearing Association National Convention, San Francisco, November, 1978.
2. NETSELL, R. *The acquisition of speech motor control: A perspective with directions for research.* Paper presented at the conference on Language Behavior in Infancy, Santa Barbara, October, 1979.
3. SHRIBERG, L., & KWIATKOWSKI, J. *Phonological programming for unintelligible children in early childhood projects.* Paper presented at the American Speech and Hearing Association National Convention, Chicago, November, 1977.
4. SHRIBERG, L., & KWIATKOWSKI, J. *Natural process analysis for children with severely delayed speech.* Paper presented at the American Speech and Hearing Association National Convention, San Francisco, November, 1978.

REFERENCES

BERNTHAL, J., & BANKSON, N. *Articulation disorders.* Englewood Cliffs, NJ: Prentice-Hall, 1981.

BROWN, R. *A first language.* Cambridge: Harvard University Press, 1973.

DARLEY, F., REES, M., SIEGEL, G., FAY, W., & NEWMAN, P. *Evaluation of appraisal techniques in speech and language pathology.* Reading, MA: Addison-Wesley, 1979.

DAVIS, E. The development of linguistic skill in twins, singletons with siblings, and only children from age five to ten years. *University of Minnesota institute of child welfare monograph,* (No. 14), University of Minnesota Press, Minneapolis, 1937 (cited in Templin, 1957).

FERGUSON, C., & FARWELL, C. Words and sounds in early language acquisition: English initial consonants in the first fifty words. *Language,* 1975, *51,* 419-439.

INGRAM, D. *Phonological disability in children.* New York: Elsevier, 1976.

MADER, J. The relative frequency of occurrence of English consonant sounds in words in the speech of children in grades one, two, and three. *Speech Monographs,* 1954, *21,* 294-300.

MILLER, J. Assessing language production in children: Experimental procedures. Baltimore: University Park Press, 1981.

MINES, M., HANSON, B., & SHOUP, J. Frequency of occurrence of phonemes in conversational English. *Language and Speech,* 1978, *21,* 221-241.

U.S. CONGRESS. *Education for All Handicapped Children Act,* (PL 94-142), 20 U.S.C. 1401 et seq, 1975.

SHRIBERG, L. Developmental phonological disorders. In T. Hixon, L. Shriberg, & J. Saxman (Eds.), *Introduction to communication disorders.* Englewood Cliffs, NJ: Prentice-Hall, 1980.

SHRIBERG, L., & KWIATKOWSKI, J. *Natural process analysis (NPA): A procedure for phonological analysis of continuous speech samples.* New York: John Wiley, 1980.

SHRIBERG, L., & KWIATKOWSKI, J. Phonological disorders I: A diagnostic classification system. *Journal of Speech and Hearing Disorders,* 1982, *47,* 226-241. (a)

SHRIBERG, L., & KWIATKOWSKI, J. Phonological disorders II: A conceptual framework for management. *Journal of Speech and Hearing Disorders,* 1982, *47,* 242-256. (b)

SHRINER, T., HOLLOWAY, M., & DANILOFF, R. The relationship between articulatory deficits and syntax in speech defective children. *Journal of Speech and Hearing Research,* 1969, *12,* 319-325.

TEMPLIN, M. *Certain language skills in children.* Minneapolis: University of Minnesota Press, 1957.

YORKSTON, K., & BEUKELMAN, D. *Assessment of intelligibility of dysarthric speech.* Tigard, OR: C. C. Publications, 1981.

WINITZ, H. *Articulatory acquisition and behavior.* Englewood Cliffs, NJ: Prentice-Hall, 1969.

Received January 23, 1981
Accepted October 13, 1981

Requests for reprints may be addressed to Lawrence D. Shriberg, Department of Communicative Disorders, 1975 Willow Drive, Madison, WI 53706.

APPENDIX A
Procedures to Calculate Average Words Per Utterance (AWU)

Counting Rules

1. An utterance is defined as "... the child 'comes' to a complete stop, either letting the voice fall, giving interrogatory or exclamatory inflection, or indicating clearly that he [does] not intend to complete the sentence." (Templin's, 1957, p. 75) adaptation of Davis (1937).
2. Unintelligible words are not counted.
3. Parts of words that were repeated are not counted, e.g., *corn-corn flakes.*
4. Filler words are not counted, e.g., *um, uh, oh.*
5. Bound morphemes are not counted as another unit, e.g., *reading* = one word.
6. Contractions of subject/verb are counted as two words, e.g., *it's, we're.* Contractions that are negatives or possessives are counted as one word, e.g., *don't, Pat's.*
7. Compound nouns are counted as one word, e.g., *blackboard.*
8. Occurrences of *yes* and *no* are counted as follows:
   a. count every token if *yes* or *no* occurs as part of a longer utterance.
   b. count only one token if more than one occurs prior to or following an utterance in which it also occurs within the utterance.

Calculation of AWU

The average words per utterance (AWU) for each sample is calculated by dividing the total number of words counted in the sample by the total number of utterances in the sample.

APPENDIX B
Procedures to Calculate Percentage of Consonants Correct (PCC)

Sampling Rules

1. Consider only intended (target) consonants in words. Intended vowels are not considered.
   a. Addition of a consonant before a vowel, e.g., *on* [hɔn] is not scored because the target sound /ɔ/ is a vowel.
   b. Post-vocalic /r/ [feɪr] *fair* is a consonant, but stressed and unstressed vocalics [ɝ], [ɚ], as in *furrier* [fɝiɚ] are vowels.
2. Do not score target consonants in the second or successive repetitions of a syllable, e.g., *ba-balloon*—score only the first /b/.
3. Do not score target consonants in words that are completely or partially unintelligible or whose gloss is highly questionable.
4. Do not score target consonants in the third or successive repetitions of adjacent words unless articulation changes. For example, the consonants in only the first two words of the series [kæt], [kæt], [kæt] are counted. However, the consonants in all three words are counted if the series were [kæt], [kæk], [kæt].

Scoring Rules

1. The following six types of consonant sound changes are scored as incorrect:
   a. deletions of a target consonant;
   b. substitutions of another sound for a target consonant, including replacement by a glottal stop or a cognate;
   c. partial voicing of initial target consonants;
   d. distortions of a target sound, no matter how subtle;
   e. addition of a sound to a correct or incorrect target consonant, e.g., *cars* said as [karks].
   f. initial /h/ deletion (he [i]) and final n/ŋ substitutions (ring [rin]) are counted as errors only when they occur in stressed syllables; in unstressed syllables they are counted as correct, e.g., feed her [fidɚ]; running [rʌnin].
2. Observe the following:
   a. The response definition for children who obviously have speech errors is "score as incorrect unless heard as correct." This response definition assigns questionable speech behaviors to an "incorrect" category.
   b. Dialectal variants should be glossed as intended in the child's dialect, e.g., *picture* "piture"; *ask* "aks", etc.
   c. Fast or casual speech sound changes should be glossed as the child intended, e.g., *don't know* "dono"; *and* "n", etc.
   d. Allophones should be scored as correct, e.g., *water* [wɑrɚ], *tail* [teɪl].

Calculation of PCC

The Percentage of Consonants Correct (PCC) for a speech sample is calculated by the formula:

$$\text{PCC} = \frac{\text{Number of Correct Consonants}}{\text{Number of Correct Plus Incorrect Consonants}} \times 100.$$

Some Descriptive Statistics for Continuous Speech Samples

Procedures for obtaining and tape recording continuous speech samples are described in detail in Shriberg and Kwiatkowski (1980) and Miller (1981). Depending on how it is collected, the same speech sample can often be used for syntactic, semantic, pragmatic, and/or phonological analyses.

During several studies we have collected information about continuous speech samples that can be useful for a variety of assessment questions in speech and language disorders. Table A is a presentation of these assorted findings. Taken together with the data presented earlier in this paper (Tables 2 and 4), these data indicate that speech samples are extremely stable. That is, structural characteristics remain similar, whether the unit of analysis is the consonant, the morpheme, the word, the canonical form, the part of speech or the utterance. Moreover, this stability yields stable derivatives that can be useful in research or clinical tasks. Here are just a few examples of the utility of the information in Table A.

a. Three-minute samples can yield approximately 25-40 in-

TABLE A. Some descriptive statistics for continuous speech samples.

Word and Morpheme Data			*Consonant, Morpheme, Word and Utterance Derivatives*			*Proportional Occurrence of Consonants in Speech Samples*			
	$\bar{x}$	*SD*		$\bar{x}$	*SD*	*Sounds*	*Adults (f)*	*Normal Children (e)*	*Delayed Speech Children (a)*	*(c)*
Number of Intelligible Words			*Number of Utterances*			*Nasals*				
(b) 6 min. samples	228	46	(c) 5 min. samples	47	8.7	m	5.11	4.63	8.1	5.6
3 min. (derived)	114		3 min. (derived)	28		n	11.49	13.14	9.9	11.7
(c) 5 min. samples	191	79	(d) 3 min. samples	28	7.3	ŋ	1.85	1.61	.3	2.5
3 min. (derived)	115						18.45	19.38	18.30	19.80
(d) 3 min. samples	90	41	*Number of Consonants*			*Glides*				
			(c) 5 min. samples	342	129	w*	4.81	5.33	2.0	4.8
Percent Monosyllable Words			3 min. (derived)	205		j	1.87	.77	1.1	2.2
(b) 6 min. samples	24	4.6	(d) 3 min. samples	159	72		6.68	6.10	3.1	7.0
			Average Number of Words			*Stops*				
			Per Utterance (AWU)			p	3.07	2.73	3.8	3.9
Parts of Speech (b)			(c) 5 min. samples	4.1	1.5	b	3.24	2.97	1.5	3.5
Percent Nouns	29.5	5.7	(d) 3 min. samples	3.2	1.4	t	9.88	11.74	23.8	13.7
Percent Verbs	25.0	5.5				d	7.80	10.25	7.6	5.8
			Number of Morphemes			k	5.30	4.25	7.4	6.0
Percent Adjectives	8.8	2.9	*Per Utterance*			g	2.02	2.38	2.5	4.1
Percent Adverbs	8.0	3.4	(c) 5 min. samples	4.4	1.6		31.31	34.32	46.60	37.00
Percent Pronouns	7.0	3.3	(d) 3 min. samples	3.3	1.5					
Percent Propositions	6.4	2.7				*Fricative/ Affricates*				
	84.7		*Number of Consonants*			f	2.65	1.83	1.1	2.4
			Per Utterance			v	2.97	1.91	2.0	1.2
Canonical Forms (a)			(c) 5 min. samples	7.3	2.3	θ	1.19	.93	1.0	.9
(based on total number of different words)			(d) 3 min. samples	5.7	2.3	ð	5.37	6.40	1.6	4.1
						s	7.88	6.50	9.7	7.1
CVC	41.8	4.3	*Number of Consonants Per Word*			z	4.70	3.70	5.8	3.0
CV	18.6	4.2	(c) 5 min. samples	1.8	.3	ʃ	.95	.84	1.0	1.5
VC	11.5	5.0	(d) 3 min. samples	1.8	.2	ʒ	.15	.01	.0	.0
CVCV	5.3	2.7				tʃ	.85	.55	.3	.7
			Number of Consonants			dʒ	.95	.69	.0	.6
CVCC	3.3	1.4	*Per Morpheme*			h	2.23	3.33	1.6	4.2
CCVC	2.5	3.6	(c) 5 min samples	1.7	.3		29.89	26.69	24.1	25.7
	83.0		(d) 3 min. samples	1.7	.2					
Number of Morphemes						*liquids*				
(c) 5 min. sample	205	81	*Number of Morphemes Per Word*			l	6.21	5.55	3.1	5.6
3 min. (derived)	123		(c) 5 min. samples	1.1	.1	r	6.61	7.83	4.5	5.2
(d) 3 min. sample	94	43	(d) 3 min. samples	1.1	.1		12.82	13.38	7.6	10.8

*combines /w/ and /ʍ/

(a) 10 children with moderately to severely delayed speech (Shriberg & Kwiatkowski, 1977)
(b) 12 children with moderately to severely delayed speech (Shriberg & Kwiatkowski, 1980)
(c) 30 children with mild to severely delayed speech (present study)
(d) 20 children with mild to severely delayed speech (Hodson, personal communication)
(e) 81 normal speaking first-third grade children (Mader, 1954)
(f) 26 adults (Mines, Hanson & Shoup, 1978)

telligible WPM, depending on the child's AWU and the procedures used for eliciting continuous speech.

b. Regardless of a child's AWU, an average of 1.8 consonants occurs per word. Thus, for example, to obtain 180 consonants would require on the average, 100 words or 3-4 minutes of continuous speech.

c. Only approximately 6% of a child's spontaneous words are monosyllabic words containing either an initial or a final cluster (i.e., either CCVC or CVCC). If a particular clinical or research task requires a sample that includes 9-10 such words for a child whose AWU averages 3.5, how many minutes of continuous speech are needed? Answer: It would take approximately 5 min to accumulate 9 monosyllabic words of these canonical forms. As indicated in Table A, children average 8-9 utterances per minute. Hence, an AWU of 3.5 will yield approximately 30 WPM; 6% of 30 words = 1.8 WPM or nine words in 5 min.

d. If a child's only speech errors are distortions of the sibilants /s/ and /z/, what is the minimum PCC a child could score in continuous speech? Answer: approximately 84%. As indicated in Table A, the proportional occurrence of /s/ is 7-10%, /z/ = 3-6%; total = maximum of 16%—subtracted from 100 = 84%. Notice that a PCC of 84% places a child close to the "mild" category of involvement as defined in the text. In contrast, note that a child who had every fricative and affricate in the sample incorrect would lose approximately 26% points from the PCC index.

These examples are only to illustrate potential uses of the PCC index in relation to severity ratings and for other purposes. The mean data are based on limited samples, but they should provide at least a first approximation for specific applications.

PCC Scoring Form

The Percentage of Consonants Correct (PCC) Scoring Form in Figure A has proven to be adequate for clinical needs. Space is provided for sorting the consonant data into initial, medial, and final position, if such are useful for a particular clinical or research purpose. If these columns are used, the consonant must actually be the first or the last intended sound in a word to be considered initial or final, respectively. Hence, /w/ in *away* is a medial consonant as are /l/ and /s/ in *blast*. This form allows ready comparison of a child's consonant performance to the proportional occurrence of consonants data provided in Table 9.

Shriberg and Kwiatkowski,1981

PERCENTAGE CONSONANTS CORRECT (PCC) SCORING FORM

Child_____
DOB_____ Age at Sampling Date_____
Sampling Date_____ PCC Scoring Date_____
Sampling Clinician_____
Scoring Clinician_____

Severity Adjective (circle):

PCC	Adjective
>85%	Mild
65%-85%	Mild-Moderate
50%-65%	Moderate-Severe
<50%	Severe

Key: + Correct; ○ Incorrect; Other:____

Consonant Class	Consonant Sound	Initial	Medial	Final	Number of Consonants Correct	Total Number of Consonants	Percent Consonants Correct
Nasals	/m/						
	/n/						
	/ŋ/						
Glides	/w/						
	/j/						
Stops	/p/						
	/b/						
	/t/						
	/d/						
	/k/						
	/g/						
Fricatives/ Affricates	/f/						
	/v/						
	/θ/						
	/ð/						
	/s/						
	/z/						
	/ʃ/						
	/ʒ/						
	/tʃ/						
	/dʒ/						
	/h/						
Liquids	/l/						
	/r/						

Notes: _____

Number of Consonants Correct	Total Number of Consonants	Percentage of Consonants Correct (PCC)

FIGURE A. Percentage of Consonants Correct (PCC) Scoring Form.

APPENDIX C
Severity Studies Materials

INSTRUCTIONS FOR SPEECH-LANGUAGE PATHOLOGISTS

Goal of Study

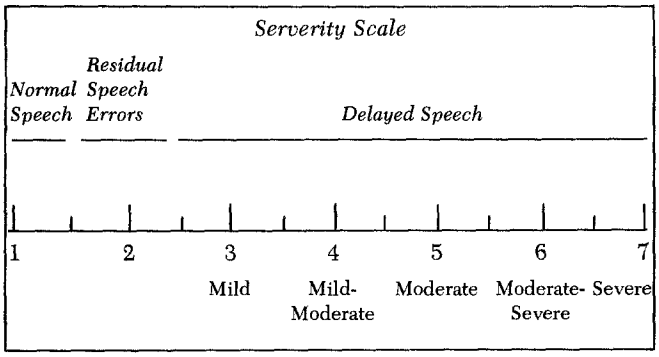
The goal of this project is to determine how speech-language pathologists rate "severity of involvement" for children with developmentally delayed speech. The project will collect severity ratings from several dozen speech-language pathologists working in schools in several states. Because little is known in this area, there are no "right" or "wrong" ways to rate "severity of involvement." The goal of the project is to determine the correlates of the concept. We hope to develop a better understanding of how "severity of involvement" can be used in assessment and management of children with developmentally delayed speech.

Procedures

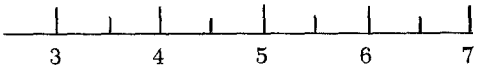
1. Please fill out the demographic data at the top of the response forms. This information is needed only to describe listeners as a group.
2. Look at the "severity scale" below the demographic data. Your task will be to circle the scale point that you think best describes the "severity of involvement" of each child you will hear on the tape. You must circle an actual scale point or half point, rather than spaces in between points.
3. For this project, all children have delayed speech. Therefore, you will be using the portion of the scale only from 3-7. That is, we have not included children with normal speech or children with only residual errors on /r/, /s/, /l/, and so forth. These children would be rated 1-2.5 on the scale.
4. Practice using the scale. Obtain an audio cassette recorder and set the tone control (if available) to the treble position. Play Side A of the tape. You will hear 10 practice samples lasting approximately 20 sec each. These samples were randomly chosen—they may or may not cover the full range of severity ratings (from 3-7) as you view the concept "severity of involvement." Listen to these 10 samples and think about which rating you would give to each child. You may replay the 10 items if you like until you feel ready to begin.
5. When you are ready to begin, proceed to the first of 32 children you are to rate. Each sample lasts approximately one minute. Notice that the clinician repeats what she thinks the child intended to say after each utterance by the child. Listen to the entire one minute sample and then make your rating on the severity scale. Remember to circle only one point along the scale (do not draw a circle around space in between points). Please listen to each child only once—do not replay any samples.
6. When finished with the ratings task, please describe the basis(es) you used to make your ratings. Remember, there are no "right" or "wrong" views of the concept "severity of involvement." We are interested in your views. Please be as candid and descriptive as you can in response to this question.

RESPONSE FORMS

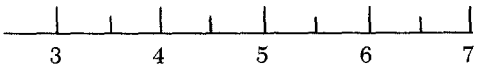
Name: _____ Total Number of Years of Paid Professional Experience: _____
City: _____ State: _____



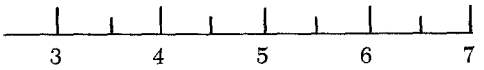
Child 1
Age: _____
Sex: _____



Child 2
Age: _____
Sex: _____



Child 3
Age: _____
Sex: _____



APPENDIX D
Directions for Coding Suprasegmentals

Directions for Scale Values

0 = normal, appropriate for the linguistic, pragmatic context
1 = slight to pronounced deviations from normal occur on a few utterances in the sample (less than 10-15%)
2 = slight to pronounced deviations from normal occur often during the sample (more than 10-15%)

Description of the Six Parameters

Parameter	*Assessment Question*
Voice Characteristics	
Pitch	Is the pitch of an utterance too low or too high?
Loudness	Is the loudness of an utterance too soft or too loud?
Quality	Is the quality of an utterance too breathy, harsh, hypernasal, etc.?
Rhythm Characteristics	
Phrasing	Are phrases appropriately divided (i.e., grouped)? Do pauses occur appropriately?
Stress	Are words appropriately emphasized relative to their canonical, syntactic, semantic, and pragmatic forms?
Rate	Are syllables, words, or phrases appropriately timed or are they said too slow, too fast, or variably too slow-too fast?

Phonological Disorders III: A Procedure for Assessing Severity of Involvement

Lawrence D. Shriberg, and Joan Kwiatkowski
J Speech Hear Disord 1982;47;256-270