CLINICAL LINGUISTICS & PHONETICS, 1991, VOL. 5, NO. 3, 225–279

Reliability studies in broad and narrow phonetic transcription

LAWRENCE D. SHRIBERG and
GREGORY L. LOF

University of Wisconsin-Madison, USA

(*Received 16 November 1990; accepted 18 March 1991*)

Abstract

A 16-category framework is proposed to review the sources of variance in studies of phonetic transcription reliability. The same framework is used to analyse transcription agreement data collected in the course of a project in child phonology, including 22 reliability estimates from five consensus transcription teams who transcribed eight subject groups. Detailed agreement data at the level of consonants, vowels and diphthongs, feature classes, and diacritics are presented for each of the 16 categories, including such traditional measurement variables as sampling mode (continuous speech; articulation tests), agreement type (intra-judge; inter-judge), and level of transcription (broad; narrow). Tabular and plotted data are deliberately presented at the lowest feasible levels for readers interested in specific questions at these levels. A total of 16 generalizations about transcription reliability are derived from descriptive and inferential statistical findings. The primary conclusion is that for certain clinical and research tasks in communicative disorders, broad phonetic transcription appears to be reliable, whereas narrow transcription may be unreliable.

Keywords: Phonetic transcription, reliability, phonological disorders

Most clinical and research data in child phonology reflect the perceptual decisions of persons trained in phonetic transcription. Despite its importance to theory and practice (e.g. 'without good phonetics there can be no good phonology'—Buckingham and Yule, 1987, p. 123), the archival literature on the validity and reliability of phonetic transcription in communicative disorders is composed of fewer than three dozen non-programmatic studies. The methodological yield from these isolated research efforts for improved training, clinical practice, and research has been a continuing source of professional concern. Oller and Eilers (1975, p. 302) recommended that researchers have a 'healthier skepticism' about transcription data; Bailey (1978, p. 141) noted that 'theoretical phonetics and phonology have made great advances in the last decade or so, but the practical field of transcriptional phonetics has not done so'; and Pye, Wilcox and Siren (1988, p. 19) suggested that there currently is 'little objective foundation' for placing confidence in the point-to-point, inter-judge transcription agreement figures of 85% or above that are typically reported in the child phonology literature.

Address correspondence to: Lawrence D. Shriberg, Phonology Project, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53705, USA.

0269-9206/91 $3.00

A historical overview of transcription studies in communicative disorders in America suggests a division into two periods, beginning with Henderson's (1938) call for programmatic research. From the 1930s through approximately 1970, research reflected the emerging discipline's attempt to develop effective ways to rate, score, and transcribe deviant speech. Impelled by efficiency criteria the assessment paradigm evolved to a form of nominal scoring of responses evoked primarily by articulation test stimuli (e.g. Burkowsky, 1967, 1971; Irwin, 1970; Irwin and Kraftchick, 1965; Jordan, 1960; March, Weaver, Morrison and Black, 1958; Milisen, 1954; Sharf, 1968; Sherman and Morrison, 1955; Siegel, 1962; Stitt and Huntington, 1963; Van Demark, 1964; Wright, 1954). A second period, beginning in the early 1970s and putatively concluding with the decade of the 1980s, witnessed renewed interest in phonetic transcription as input for linguistic description procedures developed for young normally developing children and children with intelligibility deficits (e.g. Ball, 1988; Bush, Edwards, Lackau, Stoel, Macken and Peterson, 1973; Costley and Broen, 1976; Crystal, 1985; Fokes, Bond, Ritter and Krackenfels, 1986; Duckworth, Allen, Hardcastle and Ball, 1990; Grunwell, 1982; Johnson and Bush, 1971; Kresheck, Fisher and Rutherford, 1972; MacWhinney and Marengo, 1986a,b; McCauley and Skenes, 1987; Norris, Harden and Bell, 1980; PRDS Project Working Party, 1983; Pye *et al.*, 1988; Riley, Hoffman and Damico, 1986; Schissel and Flournoy, 1978; Shriberg, Hinke and Trost-Steffen, 1987; Shriberg, Kwiatkowski and Hoffmann, 1984; Siren and Wilcox, 1990; Stockman, Woods and Tishman, 1981; Trost, 1981; Van Borsel, 1989; Vieregge and Cucchiarini, 1989). For reasons discussed below, validity and reliability problems in transcription and the availability of microprocessor technology may provide the impetus and means for the onset in the 1990s of a third era characterized by widespread use of acoustic-aided phonetic transcription.

Validity

The validity of phonetic transcription centres around three questions: (a) do perceptual data agree with data from physiologic, kinematic, or acoustic measures (e.g. Amorosa, von Benda, Wagner and Keck, 1985; Hoffman, Stager and Daniloff, 1983; Kornfield, 1974; Riley *et al.*, 1986; Weismer, 1984a,b; Weismer, Dinnsen and Elbert, 1981); (b) do perceptual decisions differ when obtained under various audio–video and linguistic presentation conditions (e.g. Elbert, Shelton and Arndt, 1967; Daniloff, Wilcox and Stephens, 1980; Hoffman and Schuckers, 1978; McNutt, Wicki and Paulsen, 1985; Oller and Eilers, 1975; Ruscello, Lass, Posch and Jones, 1980; Shelton, Johnson and Arndt, 1974; Shriberg, 1972; Stephens and Daniloff, 1977; Witting, 1962); and (c) do data derived from phonetic description match the percepts associated with clinical disorder (e.g. Barker, 1942; Burgi and Matthews, 1960; Mowrer, 1978; Oyer, 1959; Perrin, 1954; Silverman, 1976). Validity issues associated with the second and third questions are seldom posed in contemporary research, but studies concerned with the first question continue to appear in the clinical literature. Specifically, the claim is that acoustic analyses can reveal phonemic contrasts that are not observable at the level of perceptual transcription. In such studies to date, the training, skill level, and intra-judge/inter-judge reliability of the perceptual transcribers have not been well described, nor in many studies have measurement error boundaries for the acoustic data been fully reported (cf. Cole, 1973; Cole, Rudnicky, Zue and Reddy, 1980; Green, Pisoni and Carrel, 1984; Klatt and Stevens, 1973; Liberman, Cooper,

Shankweiler and Studdert-Kennedy, 1968; Millar and Wagner, 1983; Weismer, 1984a,b; Zue and Cole, 1979). What is needed is a well-controlled validity study, using representative speech samples to compare measurement outcomes from (a) skilled perceptual transcription, (b) skilled instrumental analyses, and (c) skilled, instrumentally-aided perceptual transcription (cf. Bauer and Kent, 1985). Until findings from such optimized research designs become available, the validity of unaided perceptual transcription in research and clinical speech pathology remains unattested.

Reliability

The reliability of phonetic transcription estimates the repeatability of judgements generated within a specific transcription system. Thus, both intra-judge and inter-judge agreement assess the degree of similarity in descriptions of speech by persons trained to use the same transcription system. A comprehensive review and synthesis of research on the reliability of phonetic transcription has not been available, presumably due at least in part to the complexity of relevant sources of variance.

Table 1 is an organizational heuristic that divides phonetic transcription reliability research into 16 major sources of variance. The stimuli in each reliability study reflect characteristics of certain *subjects*; data are derived from certain *analyses*; judgements are made on targets imbedded in certain linguistic *contexts*; and the speech targets in each study reflect certain phonetic and phonological *units*. Within each study, one or more of these sources of variance is the focus of investigation.

Table 1. *Sources of variance in phonetic transcription reliability*

Source	*Variable*	*Sample levels*
A. Subjects	1. Intelligibility	High, medium, low
	2. Severity of involvement	Mild, moderate, severe
	3. Type of error	Deletion, substitution, distortion
	4. Clinical significance	Articulation error, acceptable allophone
B. Analyses	5. Transcribers	Background, training
	6. Type of agreement	Intra-judge, inter-judge, consensus
	7. Type of system	Broad, narrow (International Phonetic Alphabet, other)
	8. Agreement criteria	Exact, within-class, other
C. Contexts	9. Sampling mode	Continuous speech, articulation test
	10. Structural, grammatical, and stress forms	Canonical, grammatic, stress
	11. Word position	Initial, medial, final
	12. Target environment	Stimulus context, phonetic context
D. Units	13. Class	Consonants, vowels
	14. Features	Manner, place, voicing, height
	15. Sounds	24 consonants, 17 vowels/ diphthongs
	16. Diacritics	35 symbols for narrow transcription

Table 2. *Sample findings and conclusions categorized by 12 sources of variance in phonetic transcription reliability*

Source	*Variable*	*Study*	*Mean agreement/findings*[a]	*Conclusions*
A. Subjects	1. Intelligibility	Philips and Bzoch (1969)	Inter-judge variability with intelligibility: $r = 0{\cdot}19$	There is little association between inter-judge agreement and intelligibility.
	2. Severity of involvement	Irwin (1970)	Inter-judge: correct sounds = 88%; misarticulated sounds = 66%	Agreement is considerably higher when calculated on correct compared to misarticulated sounds.
	3. Type of error	Philips and Bzoch (1969)	Agreement on error type (five classifications): intra-judge = 50–91%; inter-judge = 6–19%	Judges have low levels of agreement on each of five categories of error classifications.
		Norris, Harden and Bell (1980)	—	Inter-judge agreement on omissions was higher than on substitutions.
	4. Clinical significance	Shriberg, Kwiatkowski and Hoffmann (1984)	Exact retest consensus reliability was 68%; with non-error diacritics removed exact agreement was 76%.	Consensus transcription agreement is higher when based only on sound changes that have clinical significance.
B. Analyses	5. Transcribers	Siegel (1962)	Inter-judge: $r = 0{\cdot}92$	Transcribers can be trained to high agreement levels in making correct/incorrect judgements of sounds in isolated words, but differences on scores assigned by individual transcribers can be considerable.
		Burkowsky (1967, 1971)	Intra-judge = 64%; inter-judge = 36%	'The field of speech pathology does not produce students with proven competence in listening to speech sound production (p. 1).'
		Irwin (1970)	Inter-judge = 87%	'Undergraduate majors in speech pathology were relatively [reliable]' (p. 554).
		Diedrich and Bangert (1976, 1981)	—	Sounds may be judged as correct more often by judges who also are functioning as the child's clinician.

	6.	Type of agreement	Schissel and Flournoy (1978)	—	Intra-judge agreement was higher than inter-judge agreement for both experienced and inexperienced listeners.
	7.	Type of system	Amorosa, von Benda, Wagner, and Keck (1985)	Inter-judge: live = 56%; tape = 72%	Transcription procedures that do not allow for unlimited replays will result in over-diagnosis of phonologic disability because 'all other information on phonetic detail has either been omitted or must be considered unreliable' (p. 286).
			Pye, Wilcox, and Siren (1988)	—	Broad transcriptions were used because 'the frequency with which two or more individual transcribers chose to use the same diacritic marker for the same segment was quite small' (p. 21).
	8.	Agreement criteria	—	—	—
C. Contexts	9.	Sampling mode	Irwin and Krafchick (1965)	—	Identification of misarticulations in words was better than in phrases.
			Pye, Wilcox and Siren (1988)	—	Inter-judge agreement was higher for articulation test responses than for continuous speech.
	10.	Structural, grammatical and stress forms	McCauley and Skenes (1987)	Significantly more unstressed than stressed /r/s scored as correct	(Among other interpretations), listeners may have a more lenient standard for correct /r/ in unstressed compared to stressed contexts.
	11.	Word position	Philips and Bzoch (1969)	Interjudge: word-initial = 80%; word-medial = 78%; word-final = 67%	'sounds in final positions account for a greater portion of the diagreements' (p. 28).
	12.	Target environment	Ruscello, Lass, Posch, Jones (1980)	6–24% judgement shifts on correct, moderate errors, and severe errors on /r/ and /s/	'alterations injudgement ... occur when individuals listen repeatedly to the same stimuli' (p. 5).
			Norris, Harden, and Bell (1980)	—	Syllabic function contributed little to inter-judge agreement.

[a] Blank entries in this column reflect situations in which single agreement figures would not be representative.

with relative control of all other variables reflecting the study's unique design. As suggested in Table 1 by the number of potential interactions among typical independent (1–12) and dependent (13–16) variables, the diversity of interactive factors in any one study makes it difficult to formulate generalizations about the reliability of phonetic transcription.

Table 2 is a sample of the type of findings and conclusions reported in the phonetic transcription literature, with the independent variables in Table 1 providing the organizational framework. Selection of illustrative studies was constrained by topic (primarily normal and disordered child phonology), method (empirical studies, rather than discussions in textbooks), and access (English language publications, local availability). No attempt was made to judge the adequacy of methodology, with most studies using few subjects, samples, and transcribers (with the notable exception of the multi-state study by Diedrich and Bangert, 1976, 1980). Several examples are included to illustrate the range of some of the categories; no appropriate example could be found for level B8, Agreement Criteria.

A conclusion reached after detailed review of the transcription literature is that it is not currently possible to make useful generalizations about transcriber agreement. Even seemingly robust generalizations such as 'broad transcription is always more reliable than narrow transcription' or 'intra-judge reliability is always higher than inter-judge reliability' can be shown to be false for comparisons of certain transcribers, subjects, sampling modes, sounds, error types, and target contexts. Thus, for three goals of transcription research—improving transcription training, improving clinical practice decisions, and improving the reliability of research data—the current literature provides few useful guidelines. The purpose of the present report is to develop generalizations based on close examination of the sources of variance in broad and narrow phonetic transcription.

Method

Subject tapes

A series of transcriber reliability estimates obtained during a research programme in childhood speech disorders provided an opportunity for a retrospective study of transcription reliability. The method was to assemble a database that reflects the diversity of sources of variance described in Table 1. For subject variables, diversity included age, gender, causal origin, intelligibility, severity of involvement, error type, and clinical significance. Table 3 is a description of the demographic, structural, and speech characteristics of eight subject groups on whom reliability data had been obtained during the research program. Subject characteristics for each of the sample groups, data sets A–H, are listed at the foot of Table 3. The typical number of tape-recorded speech samples used for reliability estimates was 10–20% of the total number of subject tapes or speech tokens analysed in each study. The 51 subjects, including five adult subjects with mental retardation, reflect a diverse sample of persons with normally developing, delayed, and deviant speech.

Table 3 also contains descriptive statistics for several structural and speech variables. Structural statistics for each subset include the number of utterances, words, consonants, and vowels/diphthongs (henceforth, vowels) in the continuous speech samples and the articulation test responses. Severity of involvement information consists of descriptive statistics for the Percentage of Consonants Correct (PCC)

Table 3. *Demographic, structural, and speech characteristics of eight transcription reliability samples (see below for brief description of the data sets)*

	Demographics						*Structural Statistics*							*Speech severity indices*								
				Age (years; months)			*No. of utterances*	*No. of words*		*No. of consonants*		*No. of vowels*		*PCC*[c]			*PVC*[d]			*Intelligibility index*[e]		
Data set	*No. of subjects*	*Gender M*	*F*	$\bar{X}$	*SD*	*Range*	*CS*[a]	*CS*	*AT*[b]	*CS*	*AT*	*CS*	*AT*	$\bar{X}$	*SD*	*Range*	$\bar{X}$	*SD*	*Range*	$\bar{X}$	*SD*	*Range*
A	5	4	1	4;9	6;6	3;2–13;3	47	126	131	226	333	144	180	70·1	16·9	54·3–94·4	90·0	2·7	88·2–91·8	86·0	9·0	76·7–97·0
B	4	3	1	6;5	1;5	4;8–7;11	78	231	74	427	200	255	104	71·3	14·9	49·8–91·2	89·7	5·6	81·8–97·6	71·5	23·8	49·1–95·4
C	5	4	1	4;6	0;8	4;0–5;5	181	680	—	1257	—	824	—	77·4	8·7	65·5–86·9	92·6	3·6	86·8–96·5	91·3	7·2	80·9–98·0
D	5	2	3	4;4	0;8	3;2–4;10	57	205	124	356	330	221	170	66·6	6·3	60·6–75·4	93·4	3·5	87·9–96·9	94·1	3·8	89·5–99·3
E	5	4	1	3;11	0;6	3;3–4;6	135	430	12[f]	763	27[f]	342	17[f]	62·0	10·9	49·4–78·1	93·8	1·0	92·6–95·0	86·7	10·9	71·2–98·7
F	6	2	4	4;7	1;0	3;4–5;9	91	297	—	427	—	301	—	70·9	12·7	54·9–93·2	93·0	3·6	87·4–98·1	94·0	4·6	86·0–99·5
G	5	3	2	36	7;7	26–43	353	761	—	1357	—	916	—	76·5	4·5	69·8–80·4	95·4	0·6	94·8–96·1	86·9	21·1	49·3–98·6
H	16	9	7	4;6	1;0	3;1–6;1	89	359	—	611	—	424	—	76·9	7·5	63·0–93·3	95·3	2·5	88·7–98·2	96·1	1·9	92·4–98·8
Total	51	31	20				1031	3089	341	5424	890	3427	471									
Mean				5;3[g]			128·9	386·1	85·3	678	222·5	428	117·8	71·5			92·9			88·3		

Set A: Five children with speech delays of unknown origin; Set B: four children with speech delays with unknown origin being followed in a longitudinal study; Set C: five native American children with suspected speech delays associated with recurrent otitis media; Set D: five children with repaired clefts of the palate; Set E: five children with speech delays of unknown origin; Set F: six children with speech delays of unknown origin; Set G: five adults with mental retardation; Set H: 16 preschool children with normally developing speech.

[a] Continuous speech (Shriberg and Kwiatkowski, 1980); [b] Articulation test (Pendergast, Dickey, Selmar and Soder, 1969); [c] Percentage of Consonants Correct (Shriberg and Kwiatkowski, 1982); [d] Percentage of Vowels Correct (Shriberg, 1986); [e] Shriberg (1986); [f] Data were available for only one subject; [g] All data sets except Set G.

(Shriberg and Kwiatkowski, 1982; Shriberg, Kwiatkowski, Best, Hengst and Terselic-Weber, 1986), the Percentage of Vowels/Diphthongs Correct (PVC) (Shriberg, 1986), which is calculated on vowels and diphthongs using the same rationale as established for the PCC, and Intelligibility Index scores (Shriberg, 1986; Shriberg and Kwiatkowski, 1982).

Transcribers and transcription procedures

During a 7-year period, five transcription teams completed narrow phonetic transcriptions of the children and one group of adults in the eight data sets. The first author initially trained two persons to complete phonetic transcription individually and by consensus (Shriberg *et al.*, 1984). For different studies within the project, pairs of these three persons comprised two consensus transcription teams (Team I, Team II). Eventually, the first author and a research colleague (J. Kwiatkowski) trained two new persons to form a third consensus transcription team (Team III). After approximately 2 years, the third team had the primary responsibility of training a fourth team (Team IV) using a set of procedures designed specifically for selecting and training persons for narrow phonetic transcription by consensus (Shriberg *et al.*, 1987). Finally, for an off-site study, the first author trained two persons in another state to form a fifth consensus transcription team (Team V). Training for all individual transcribers and team transcription included illustrated lectures, self-study on audio-tutorial materials (Shriberg and Kent, 1982), criterion tests to identify areas for additional study, and intensive training sessions to increase technical knowledge, perceptual skills, and familiarity with all transcription and software conventions. Each of the four local transcription teams worked 12–20 hours per week transcribing by consensus, with intra-judge and inter-judge reliability samples obtained as transcription of each data set was completed.

On-site consensus transcription was accomplished in a sound-treated acoustic suite using one of several well-maintained Dictaphone 2550 series playback devices (published bandwidth = 200–5000 Hz; S/N ratio = 40 dB). A group of five experienced transcribers unanimously selected these devices as having the best audio quality for free-field transcription, compared to three other commercially available audiocassette transcribers. The off-site team trained in a quiet room using a Dictaphone 2890 device. All teams followed a detailed set of written instructions and guidelines, including such procedural variables as distances from the playback device, number of allowable replays, and notational conventions. Laboratory notebooks kept during the tenure of each on-site consensus team provided records on all technical and procedural problems, as well as anecdotal comments on specific transcription difficulties.

Analyses

Table 4 is an overview of the 22 reliability samples assembled for the present study. Cell entries indicate the inter-judge and intra-judge reliability data for the nine individual transcribers and the five consensus transcription teams. Using the same letter codes for the studies described in Table 3, reliability estimates from the continuous speech samples are in the top of each cell and estimates from articulation testing (Pendergast, Dickey, Selmar and Soder, 1969) are in the bottom. As shown in Table 4, 12 of the 22 reliability samples are based on the continuous speech and articulation test responses obtained in Set A, including inter-judge consensus esti-

Table 4. *Reliability samples transcribed by individual transcribers and consensus transcription teams[a]. Intra-judge samples for individual and consensus team transcription appear along the diagonal, with available inter-judge samples in the remainder of the cells. Within each cell of the table reliability samples based on continuous speech samples are above; those based on articulation test responses are below.*

	Transcribers													
Transcribers	*1*	*2*	*3*	*Team I (1&2)*	*Team II (2&3)*	*4*	*5*	*Team III (4&5)*	*6*	*7*	*Team IV (6&7)*	*8*	*9*	*Team V (8&9)*
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3	—	H	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Team I (1&2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Team II (2&3)	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Team III (4&5)	—	—	—	A	—	—	—	B,E	—	—	—	—	—	—
	—	—	—	A	—	—	—	B	—	—	—	—	—	—
6	—	—	—	—	—	—	—	—	G	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	—	—	—	—	—	—	—	—	G	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Team IV (6&7)	—	—	—	A	—	—	—	A	—	—	C,D,F	—	—	—
	—	—	—	A	—	—	—	A	—	—	D	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Team V (8&9)	—	—	—	A	—	—	—	A	—	—	A	—	—	—
	—	—	—	A	—	—	—	A	—	—	A	—	—	—

[a] See Table 3 for a key to the eight data sets, A–H.

mates for four consensus transcription teams. Of the remaining 10 samples, eight are intra-judge consensus team samples for sets B–G and two are inter-judge individual transcriber samples for sets G and H. The reliability data obtained from these 22 samples reflect the variety of subject, analysis, context, and unit variables proposed in the organizational model (Table 1).

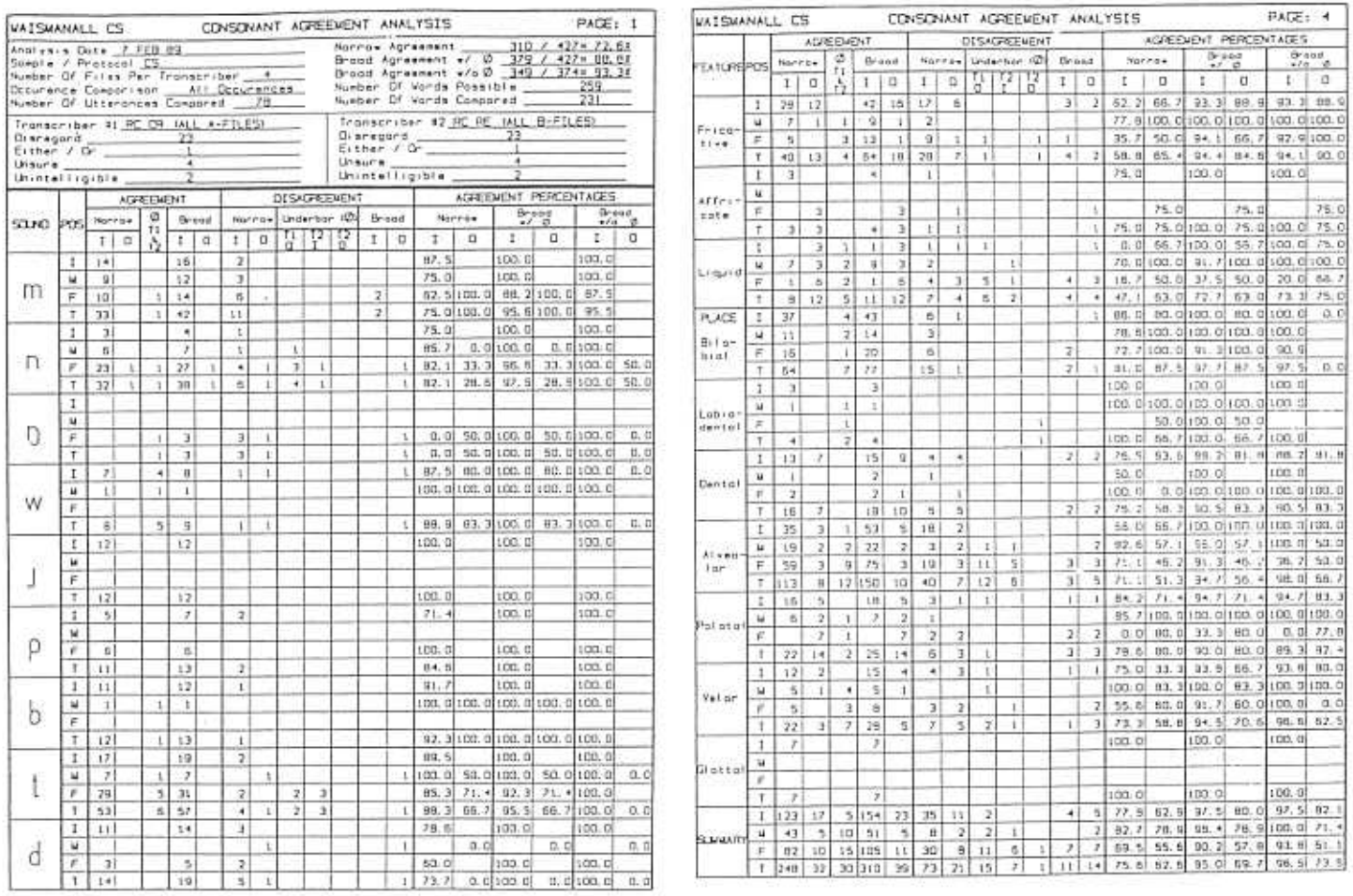
Software and response definitions

An application program was developed to provide detailed quantitative analyses of consonant, vowel, and diacritic-level transcription agreement (Shriberg and Olson, 1987). The program is based on the data structures used in PEPPER (Shriberg, 1986), a series of programs that provided the speech data shown in Table 3. The transcriber agreement software required two exactly similarly glossed transcripts as input. Thus, the agreement and disagreement data were not complicated by the validity issues associated with differences in glossing (Oller and Eilers, 1975; Shriberg *et al.*, 1984, 1987; Witting, 1962). Using the original and second (reliability) tran-

scripts obtained for each study, a research assistant entered into the agreement program all utterances with similar glosses on both transcripts. Glosses for responses to the articulation test words were almost always similar; those that were not (i.e. those containing different intended vowels or different intended affixes) were excluded. A variety of words from the articulation test were reflected across the reliability studies. For a few of the intra-judge and inter-judge reliability samples obtained after the reliability software became available, transcribers were given the original glosses from which to derive their repeated transcriptions.

Figures 1, 2, and 3 are sample pages of output from the consonant, vowel, and diacritic agreement programs, respectively. The first and last pages of output shown in Figures 1 and 2 provide transcriber agreement information at the phoneme (24 consonants, 17 vowels) and feature (class, voicing, manner, place) levels, with percentages cross-tabulated for word-initial, word-medial, and word-final positions. At each of these levels, data are available for both broad and narrow transcription. Separate agreement/disagreement entries are available for occasions when one or both transcribers indicated the speaker said the Intended (I) sound (i.e. the sound expected from the gloss) or an Other (O) sound (i.e. a different phoneme than expected from the gloss). Separate tabulations are also available for comparisons in which the referent (T1) or comparison (T2) transcriber or both transcribers heard a phoneme deletion (∅), which as shown in Figures 1 and 2, is indicated in the underbar column. Row-level entries provide data for *I*nitial, *M*edial, and *F*inal positions and an overall position *T*otal. In addition to these cell-level data, single subject and grouped percentages of agreement are printed for broad and narrow phonetic transcription. It is important to note that all summary-level percentage calculations are appropriately weighted by the number of contributing entries, rather than reflecting averages of the individual percentages (i.e. the data are not the average of the averages).

The sample output shown in Figure 3 provides transcriber agreement information at the level of diacritic use. For the current purposes, 35 of the 45 diacritics (Shriberg and Kent, 1982) available in the speech analysis program (symbols marking juncture and stress were excluded from analysis) are divided into seven categories or classes: nasality, lip, stop release, tongue configuration, tongue position, sound source, and timing/other. Thus, as shown in Figure 3, the Diacritic Agreement output provides agreement data at both the level of individual diacritics and the level of diacritic class. The example shown in Figure 3 is the output page for the Tongue Configuration class, which includes six diacritics marking (in the order shown at the top of Fig. 3) dentalized, palatalized, lateralized, rhotacized, velarized, and derhotacized. One page of output is produced for each of the seven classes of diacritics. Exact agreement within or between transcribers (or teams) requires that both used the same diacritic, as indicated within the bolded diagonal cells. Within-class agreements are shown by off-diagonal tallies indicating that one transcriber (or team) heard one of the diacritics in the class while the other transcriber heard another member of the class. A tally in the cells corresponding to Other (termed *A*ny in the following figures, i.e., use of *any* diacritic) represents an out-of-class agreement, in which transcribers each heard a diacritic, but the comparison transcriber heard a diacritic in one of the other six classes. A tally in the None column indicates that a diacritic was used by only one of the transcribers. Finally, each of these diacritic-level transcription values is available for sounds on which transcribers had the *S*ame main character, a *D*ifferent main character and, as used in the present study, for *A*ll (i.e. both) conditions. The



WAISMANALL CS — CONSONANT AGREEMENT ANALYSIS — PAGE: 1

Analysis Date 7 FEB 89
Sample / Protocol CS
Number Of Files Per Transcriber 4
Occurence Comparison All Occurences
Number Of Utterances Compared 78

Narrow Agreement 310 / 427= 72.6%
Broad Agreement w/ Ø 379 / 427= 88.8%
Broad Agreement w/o Ø 349 / 374= 93.3%
Number Of Words Possible 259
Number Of Words Compared 231

Transcriber #1 RC 09 (ALL A-FILES)
Disregard 23
Either / Or 1
Unsure 4
Unintelligible 2

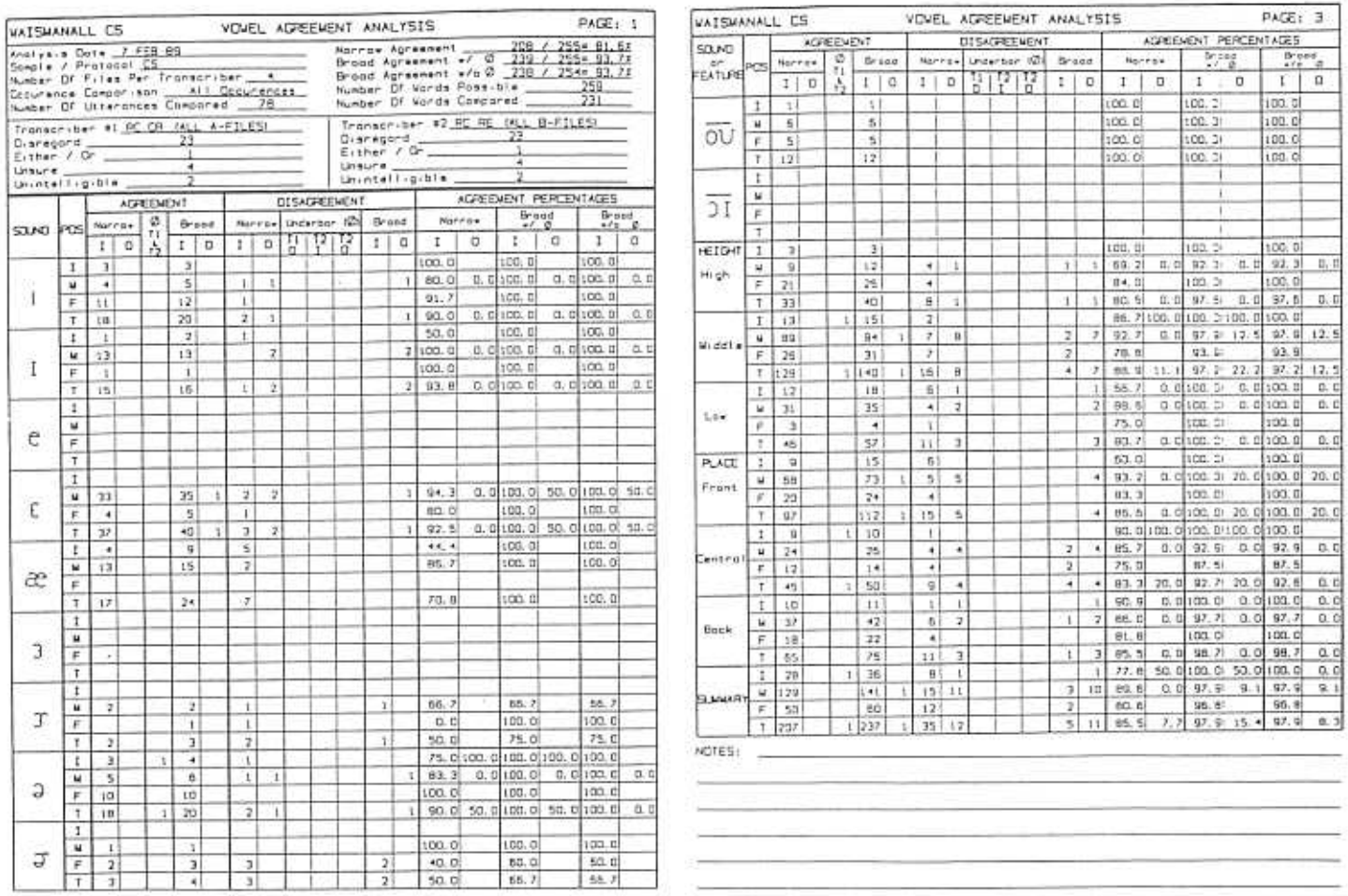
Transcriber #2 RC RE (ALL B-FILES)
Disregard 23
Either / Or 1
Unsure 4
Unintelligible 2

SOUND	POS	AGREEMENT: Narrow T	AGREEMENT: Narrow O	AGREEMENT: Ø T1	AGREEMENT: Ø T2	AGREEMENT: Broad T	AGREEMENT: Broad O	DISAGREEMENT: Narrow T	DISAGREEMENT: Narrow O	DISAGREEMENT: Underbar (Ø) T1	DISAGREEMENT: Underbar (Ø) T2	DISAGREEMENT: Underbar (Ø) T3	DISAGREEMENT: Broad T	DISAGREEMENT: Broad O	AGREEMENT PERCENTAGES: Narrow		Broad w/ Ø		Broad w/o Ø	
		T	O	T1	T2	T	O	T	O	T1	T2	T3	T	O	T	O	T	O	T	O
m	I	14				16		2							87.5		100.0		100.0	
	W	9				12		3							75.0		100.0		100.0	
	F	10		1		14		6	-				2		62.5	100.0	88.2	100.0	87.5	
	T	33		1		42		11	-				2		75.0	100.0	95.6	100.0	95.5	
n	I	3				4		1							75.0		100.0		100.0	
	W	6				7		1		1					85.7	0.0	100.0	0.0	100.0	
	F	23	1	1		27	1	4	1	3	1			1	82.1	33.3	96.6	33.3	100.0	50.0
	T	32	1	1		38	1	5	1	4	1			1	82.1	28.6	97.5	28.6	100.0	50.0
D	I																			
	W																			
	F			1		3		3	1					1	0.0	50.0	100.0	50.0	100.0	0.0
	T			1		3		3	1					1	0.0	50.0	100.0	50.0	100.0	0.0
w	I	7		4		8		1	1					1	87.5	80.0	100.0	80.0	100.0	0.0
	W	1		1		1									100.0	100.0	100.0	100.0	100.0	
	F																			
	T	8		5		9		1	1					1	88.9	83.3	100.0	83.3	100.0	0.0
J	I	12				12									100.0		100.0		100.0	
	W																			
	F																			
	T	12				12									100.0		100.0		100.0	
p	I	5				7		2							71.4		100.0		100.0	
	W																			
	F	6				6									100.0		100.0		100.0	
	T	11				13		2							84.6		100.0		100.0	
b	I	11				12		1							91.7		100.0		100.0	
	W	1		1		1									100.0	100.0	100.0	100.0	100.0	
	F																			
	T	12		1		13		1							92.3	100.0	100.0	100.0	100.0	
t	I	17				19		2							89.5		100.0		100.0	
	W	7		1		7			1					1	100.0	50.0	100.0	50.0	100.0	0.0
	F	29		5		31		2		2	3				85.3	71.4	92.3	71.4	100.0	
	T	53		6		57		4	1	2	3			1	88.3	66.7	95.5	66.7	100.0	0.0
d	I	11				14		3							78.6		100.0		100.0	
	W								1					1		0.0		0.0		0.0
	F	3				5		2							60.0		100.0		100.0	
	T	14				19		5	1					1	73.7	0.0	100.0	0.0	100.0	0.0

WAISMANALL CS — CONSONANT AGREEMENT ANALYSIS — PAGE: 4

FEATURE	POS	AGREEMENT: Narrow T	AGREEMENT: Narrow O	AGREEMENT: Ø T1	AGREEMENT: Ø T2	AGREEMENT: Broad T	AGREEMENT: Broad O	DISAGREEMENT: Narrow T	DISAGREEMENT: Narrow O	DISAGREEMENT: Underbar (Ø) T1	DISAGREEMENT: Underbar (Ø) T2	DISAGREEMENT: Underbar (Ø) T3	DISAGREEMENT: Broad T	DISAGREEMENT: Broad O	AGREEMENT PERCENTAGES: Narrow		Broad w/ Ø		Broad w/o Ø	
		T	O	T1	T2	T	O	T	O	T1	T2	T3	T	O	T	O	T	O	T	O
Fricative	I	39	12			42	15	17	6				3	2	62.2	66.7	93.3	88.9	93.3	88.9
	W	7	1	1		9	1	2							77.8	100.0	100.0	100.0	100.0	100.0
	F	5		3		13	1	9	1	1		1	1		35.7	50.0	94.1	66.7	92.9	100.0
	T	40	13	4		64	18	28	7	1		1	4	2	58.8	65.4	94.4	84.8	94.1	90.0
Affricate	I	3				4		1							75.0		100.0		100.0	
	W																			
	F		3				3		1					1		75.0		75.0		75.0
	T	3	3			4	3	1	1					1	75.0	75.0	100.0	75.0	100.0	75.0
Liquid	I		3	1		1	3	1	1	1				1	0.0	66.7	100.0	66.7	100.0	75.0
	W	7	3	2		8	3	2			1				70.0	100.0	91.7	100.0	100.0	100.0
	F	1	6	2		1	6	4	3	5	1		4	3	16.7	50.0	37.5	50.0	20.0	66.7
	T	8	12	5		11	12	7	4	6	2		4	4	47.1	63.0	72.7	63.0	73.3	75.0
PLACE	I	37		4		43		6	1					1	86.0	80.0	100.0	80.0	100.0	0.0
Bilabial	W	11		2		14		3							78.6	100.0	100.0	100.0	100.0	
	F	15		1		20		6					2		72.7	100.0	91.3	100.0	90.9	
	T	64		7		77		15	1				2	1	81.5	87.5	97.7	87.5	97.5	0.0
Labiodental	I	3				3									100.0		100.0		100.0	
	W	1		1		1									100.0	100.0	100.0	100.0	100.0	
	F			1								1				50.0	100.0	50.0		
	T	4		2		4						1			100.0	66.7	100.0	66.7	100.0	
Dental	I	13	7			15	9	4	4				2	2	76.5	93.6	88.2	81.8	88.2	91.8
	W	1				2		1							50.0		100.0		100.0	
	F	2				2	1		1						100.0	0.0	100.0	100.0	100.0	100.0
	T	16	7			19	10	5	5				2	2	76.2	58.3	90.5	83.3	90.5	83.3
Alveolar	I	35	3	1		51	5	18	2						66.0	66.7	100.0	100.0	100.0	100.0
	W	19	2	2		22	2	3	2	1	1			2	82.6	57.1	95.0	57.1	100.0	50.0
	F	59	3	9		75	3	19	3	11	5		3	3	71.1	46.2	91.3	46.2	98.2	50.0
	T	113	8	12		150	10	40	7	12	6		3	5	71.1	51.3	94.1	56.4	98.0	66.7
Palatal	I	16	5			18	5	3	1	1			1	1	84.2	71.4	94.7	71.4	94.7	83.3
	W	6	2	1		7	2	1							85.7	100.0	100.0	100.0	100.0	100.0
	F		7	1			7	2	2				2	2	0.0	80.0	33.3	80.0	0.0	77.8
	T	22	14	2		25	14	6	3	1			3	3	78.6	80.0	90.0	80.0	89.3	82.4
Velar	I	12	2			15	4	4	3	1			1	1	75.0	33.3	93.8	66.7	93.8	80.0
	W	5	1	4		5	1			1					100.0	83.3	100.0	83.3	100.0	100.0
	F	5		3		8		3	2					2	55.6	80.0	91.7	80.0	100.0	0.0
	T	22	3	7		28	5	7	5	2	1		1	3	73.3	58.8	94.5	70.6	98.6	82.5
Glottal	I	7				7									100.0		100.0		100.0	
	W																			
	F																			
	T	7				7									100.0		100.0		100.0	
SUMMARY	I	123	17	5		154	23	35	11	2			4	5	77.5	62.5	97.5	80.0	97.5	82.1
	W	43	5	10		51	5	8	2	2	1			2	82.7	78.9	99.4	78.9	100.0	71.4
	F	82	10	15		105	11	30	8	11	6	1	7	7	69.5	55.6	90.2	57.8	93.8	51.1
	T	248	32	30		310	39	73	21	15	7	1	11	14	75.6	62.5	95.0	69.7	96.5	73.5

Figure 1. *Sample output from the Consonant Agreement Analysis program. The left and right panels show the formats for the sound-level and feature-level information, respectively. See text for explanation of column headings and abbreviations.*



WAISMANALL CS — VOWEL AGREEMENT ANALYSIS — PAGE: 1

Analysis Date: 7 FEB 89
Sample / Protocol: CS
Number Of Files Per Transcriber: 4
Occurence Comparison: All Occurences
Number Of Utterances Compared: 78

Narrow Agreement: 208 / 255= 81.6%
Broad Agreement w/ Ø: 239 / 255= 93.7%
Broad Agreement w/o Ø: 238 / 254= 93.7%
Number Of Words Possible: 259
Number Of Words Compared: 231

Transcriber #1 PC CR (ALL A-FILES)
Disregard: 23
Either / Or: 1
Unsure: 4
Unintelligible: 2

Transcriber #2 PC RE (ALL B-FILES)
Disregard: 22
Either / Or: 1
Unsure: 4
Unintelligible: 2

SOUND	POS	AGREEMENT: Narrow I	Narrow O	Ø T1	Ø T2	Broad I	Broad O	DISAGREEMENT: Narrow I	Narrow O	Underscr (Ø) T1	Underscr (Ø) T2	Underscr (Ø) T3	Broad I	Broad O	AGREEMENT PERCENTAGES: Narrow I	Narrow O	Broad w/ Ø I	Broad w/ Ø O	Broad w/o Ø I	Broad w/o Ø O
I	I	3				3									100.0		100.0		100.0	
	M	4				5		1	1					1	80.0	0.0	100.0	0.0	100.0	0.0
	F	11				12		1							91.7		100.0		100.0	
	T	18				20		2	1					1	90.0	0.0	100.0	0.0	100.0	0.0
I	I	1				2		1							50.0		100.0		100.0	
	M	13				13			2					2	100.0	0.0	100.0	0.0	100.0	0.0
	F	1				1									100.0		100.0		100.0	
	T	15				16		1	2					2	93.8	0.0	100.0	0.0	100.0	0.0
e	I																			
	M																			
	F																			
	T																			
ɛ	I																			
	M	33				35	1	2	2					1	94.3	0.0	100.0	50.0	100.0	50.0
	F	4				5		1							80.0		100.0		100.0	
	T	37				40	1	3	2					1	92.5	0.0	100.0	50.0	100.0	50.0
æ	I	4				9		5							44.4		100.0		100.0	
	M	13				15		2							86.7		100.0		100.0	
	F																			
	T	17				24		7							70.8		100.0		100.0	
ɔ	I																			
	M																			
	F	-																		
	T																			
ɪ	I																			
	M	2				2		1						1	66.7		66.7		66.7	
	F					1		1							0.0		100.0		100.0	
	T	2				3		2						1	50.0		75.0		75.0	
ə	I	3		1		4		1							75.0	100.0	100.0	100.0	100.0	
	M	5				6		1	1					1	83.3	0.0	100.0	0.0	100.0	0.0
	F	10				10									100.0		100.0		100.0	
	T	18		1		20		2	1					1	90.0	50.0	100.0	50.0	100.0	0.0
ɝ	I																			
	M	1				1									100.0		100.0		100.0	
	F	2				3		3						2	40.0		60.0		50.0	
	T	3				4		3						2	50.0		66.7		66.7	

WAISMANALL CS — VOWEL AGREEMENT ANALYSIS — PAGE: 3

SOUND OR FEATURE	POS	AGREEMENT: Narrow I	Narrow O	Ø T1	Ø T2	Broad I	Broad O	DISAGREEMENT: Narrow I	Narrow O	Underscr (Ø) T1	Underscr (Ø) T2	Underscr (Ø) T3	Broad I	Broad O	AGREEMENT PERCENTAGES: Narrow I	Narrow O	Broad w/ Ø I	Broad w/ Ø O	Broad w/o Ø I	Broad w/o Ø O
OU	I	1				1									100.0		100.0		100.0	
	M	5				5									100.0		100.0		100.0	
	F	5				5									100.0		100.0		100.0	
	T	12				12									100.0		100.0		100.0	
ɔI	I																			
	M																			
	F																			
	T																			
HEIGHT: High	I	3				3									100.0		100.0		100.0	
	M	9				12		4	1					1	69.2	0.0	92.3	0.0	92.3	0.0
	F	21				25		4							84.0		100.0		100.0	
	T	33				40		8	1					1	80.5	0.0	97.5	0.0	97.5	0.0
Middle	I	13		1		15		2							86.7	100.0	100.0	100.0	100.0	
	M	88				94	1	7	8					2	92.7	0.0	97.9	12.5	97.8	12.5
	F	28				31		7						2	78.8		93.9		93.9	
	T	129		1		140	1	16	8					4	88.9	11.1	97.2	22.2	97.2	12.5
Low	I	17				18		6	1					1	66.7	0.0	100.0	0.0	100.0	0.0
	M	31				35		4	2					2	88.6	0.0	100.0	0.0	100.0	0.0
	F	3				4		1							75.0		100.0		100.0	
	T	46				57		11	3					3	80.7	0.0	100.0	0.0	100.0	0.0
PLACE: Front	I	9				15		6							60.0		100.0		100.0	
	M	58				73	1	5	5					4	93.2	0.0	100.0	20.0	100.0	20.0
	F	20				24		4							83.3		100.0		100.0	
	T	87				112	1	15	5					4	85.5	0.0	100.0	20.0	100.0	20.0
Central	I	8		1		10		1							90.0	100.0	100.0	100.0	100.0	
	M	24				25		4	4					2	85.7	0.0	92.9	0.0	92.9	0.0
	F	12				14		4						2	75.0		87.5		87.5	
	T	45		1		50		9	4					4	83.3	20.0	92.7	20.0	92.8	0.0
Back	I	10				11		1	1					1	90.9	0.0	100.0	0.0	100.0	0.0
	M	37				42		5	2					1	88.0	0.0	97.7	0.0	97.7	0.0
	F	18				22		4							81.8		100.0		100.0	
	T	85				75		11	3					1	85.5	0.0	98.7	0.0	98.7	0.0
SUMMARY	I	28		1		36		8	1					1	77.8	50.0	100.0	50.0	100.0	0.0
	M	179				141	1	15	11					3	89.8	0.0	97.9	9.1	97.9	9.1
	F	53				80		12						2	80.8		96.8		96.8	
	T	237		1		237	1	35	12					5	85.5	7.7	97.9	15.4	97.9	8.3

NOTES:

Figure 2. *Sample output from the Vowel Agreement Analysis program. The left and right panels show the formats for the sound-level and feature-level information, respectively. See text for further description.*

DIACRITIC AGREEMENT ANALYSIS

SK/BB TRANSALL PAT — TONGUE CONFIGURATION — PAGE: 5

BB ALLS4

SK ALLS1	Diacritic		˕	˖	ˆ	˷	˜	˘	OTHER	NONE	TOTAL AGREE	TOTAL DISAGREE	PERCENT AGREE
˕		S								2		2	0.0
		O								1		1	0.0
		A								3		3	0.0
˖		S		1							1		100.0
		O											
		A		1							1		100.0
ˆ		S		1	3						3	1	75.0
		O			1						1		100.0
		A		1	4						4	1	80.0
˷		S	1	1		5					5	2	71.4
		O											
		A	1	1		5					5	2	71.4
˜		S											
		O											
		A											
˘		S						7		1	7	1	87.5
		O								6		6	0.0
		A						7		7	7	7	50.0
OTHER		S	1				1						
		O											
		A	1				1						
NONE		S	1	2			1	2					
		O											
		A	1	2			1	2					
TOTAL AGREE		S		1	3	5		7					
		O			1								
		A		1	4	5		7					
TOTAL DISAGREE		S	3	4			2	2					
		O											
		A	3	4			2	2					
PERCENT AGREE		S	0.0	20.0	100.0	100.0	0.0	77.8					
		O			100.0								
		A	0.0	20.0	100.0	100.0	0.0	77.8					

CLASS SUMMARY		CLASS TOTAL AGREE	CLASS TOTAL DISAGREE	CLASS PERCENT AGREE
CLASS TOTAL AGREE	S	16	6	72.7
	O	1	7	12.5
	A	17	13	56.7
CLASS TOTAL DISAGREE	S	11		COMBINED PERCENT AGREE
	O			
	A	11		
CLASS PERCENT AGREE	S	59.3		55.3
	O	100.0		22.2
	A	60.7		58.6

Figure 3. *Sample output from the Diacritic Agreement Analysis program. This sample includes sample transcription agreement data for the use of Tongue Configuration.*

row and column totals reflect figures based on the vertical axis and horizontal axis transcribers, respectively. The Combined Percent Agreement summary provided in the bottom right section reflects weighted percentages from the two comparison transcriptions. Specifically, as agreement figures do not presume a standard comparative stimulus, all percentages of agreement (Exact, Within, Other) were computed using the formula:

$$\text{Agreement} = \frac{\text{Total agreements (Transcriber 1 + Transcriber 2)}}{\frac{\text{Disagreements (Transcriber 1 + Transcriber 2)}}{2} + \text{Total agreements (Transcriber 1 + Transcriber 2)}} \times 100$$

In Figure 3, for example, the overall point-to-point agreement on all the six tongue configuration symbols was 58·6%, including 65·3% agreement when based on the same main character and 22·2% agreement when based on occasions when transcribers differed on the main character as well as the diacritic.

Research questions and statistical approach

The goal of this report is to identify, for use and further study, generalizations about sources of variance in phonetic transcription. The analysis approach was exploratory, with graphic profiles used to discover and support generalizations. Non-parametric inferential statistics were used sparingly to support trends. The analyses proceed upwards using the categories in Table 1, beginning with reliability questions about diacritics and concluding with information about subjects.

Results and discussion

Units

Diacritics

Diacritics per word rates. Transcriber agreement computations for diacritics might be biased by differences in the average number of diacritics used by different transcribers and transcription teams. To provide a metric of diacritic use independent of sample size, Diacritics Per Word (DPW) indices were computed by dividing the number of diacritics included in each sample by the total number of transcribed words in the sample. Figure 4 is a plot of these data, with the sections separated by the dashed line providing individual DPW data for inter-judge and intra-judge (*O*riginal, *R*epeat) reliability samples for seven of the eight data sets (A–G) in two sampling modes (*C*ontinuous *S*peech, *A*rticulation *T*est). The trends in Figure 4 suggest the following two generalizations:

> There are substantial differences in the average number of diacritics per word used by different consensus transcription teams within and between sampling modes and subject groups.
>
> There is fairly stable consistency in the average number of diacritics per word used by the same consensus transcription team doing narrow phonetic transcription on the same speech sample.

The data points for the four transcription teams and one transcriber in Figure 4 yield a range of approximately 0·40 to 2·0 diacritics per word, a 5:1 ratio across the two types of reliability estimates, two sampling modes, and seven subject groups. Most notable were differences among the four consensus teams on Set A, which included their inter-judge agreement for continuous speech and articulation test responses.

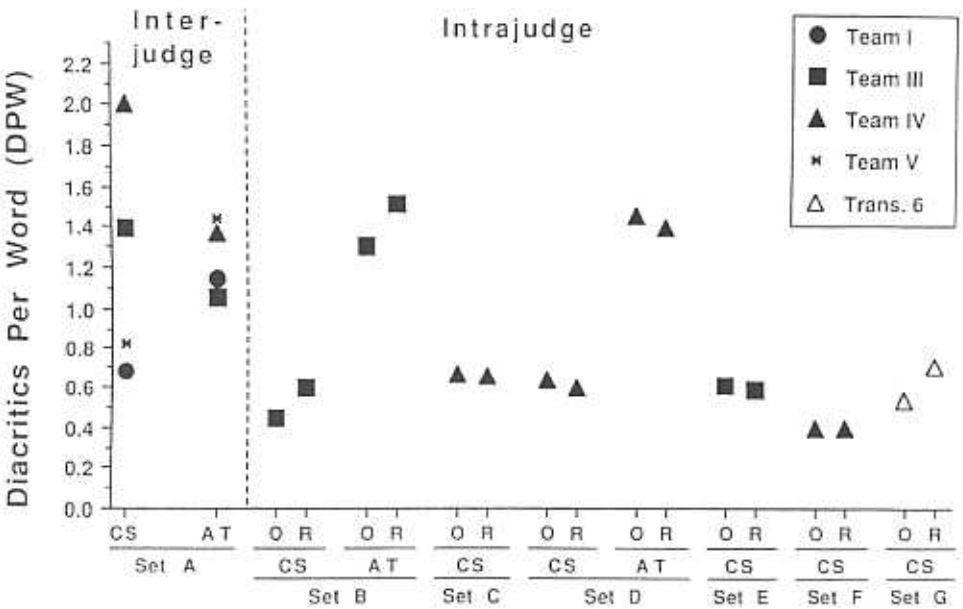


Figure 4. *Diacritics Per Word (DPW) rates for the five consensus transcription teams and one individual transcriber by sampling mode (Continuous Speech, Articulation Test). Intra-judge agreement data points include the Original and Repeat transcriptions.*

Of the eight comparisons of consensus teams' intra-judge consistency of diacritics per word in the same sampling mode (all data points to the right of the dashed line) the range of differences was within 0·30 DPW. As shown in Figure 4, six of the eight intra-judge comparisons were based on continuous speech samples, as well as two based on articulation test responses. Compared to the overall dispersion of inter-judge DPWs, which average nearly three times these differences, the rates of diacritics per word for different consensus teams appear to be stable when they rescore the same speech sample.

These two generalizations about the absolute frequencies of diacritic use play a primary role in the interpretation of the agreement data to follow. Recall that for this study any transcription difference in diacritics is considered a disagreement, even a difference that would not result in a subject obtaining a different articulation severity score. For example, if one transcription team used a diacritic to mark an unreleased stop (which is an acceptable word-final stop allophone) and the other team or a retest transcription did not include this diacritic, the agreement program counted the difference as a disagreement. Thus, base-rate differences in individuals' or teams' use of any of the 35 diacritics result in lower narrow transcription agreement.

Proportional occurrence of diacritics. Figure 5 includes proportional occurrence data for the 35 diacritics. These percentages are averaged over all teams and studies, reflecting the number of times a diacritic was used divided by the total number of diacritics used by each transcriber or transcription team. The averaged percentages are sorted in descending order of proportional occurrence. Labels for the diacritic symbols in Figure 5 occur in corresponding order in the first column in Table 6, to follow.

As a gross division of the proportional occurrence distribution based on the

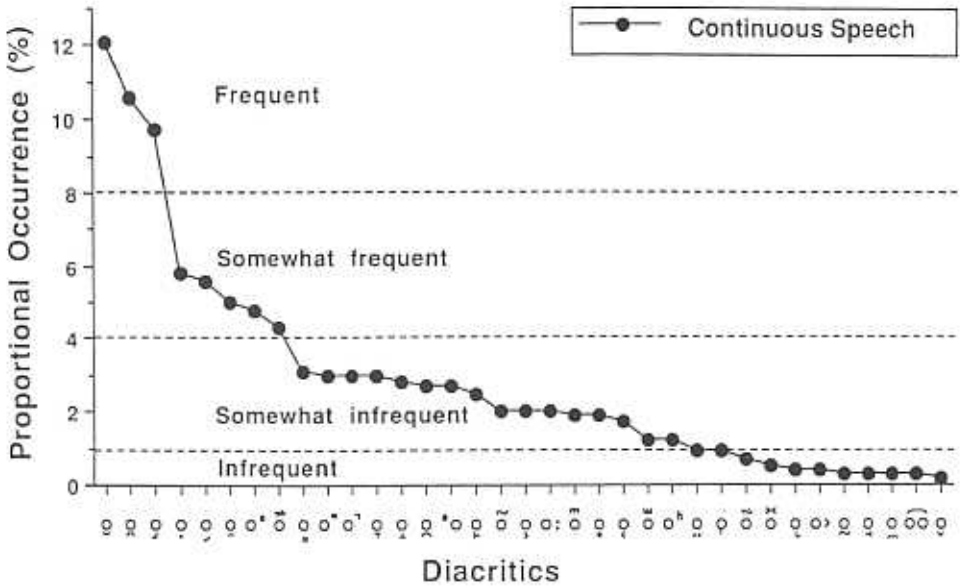


Figure 5. *Proportional occurrence of diacritics across all studies and transcription teams. Labels for the 35 diacritics are provided in Table 6, with the left-to-right order of labels in the figure corresponding to their vertical order in the table.*

magnitudes of the percentages and natural breaks in the trends, the 35 diacritics in Figure 5 are divided into four categories. The subgroup of three tongue configuration diacritics marking derhotacized (consonant /r/ and vowels /ɚ/, /ɝ/), dentalized (primarily fricatives), and palatalized (primarily fricatives) were clearly *frequent* across data sets, transcribers, and sampling mode. The second subgroup of five *somewhat frequent* diacritics included allophone-level differences from several different phonetic feature classes. Sixteen *somewhat infrequent* diacritics also reflected a diversity of diacritic classes. A subgroup of 11 *infrequent* diacritics (nearly one-third of the 35 diacritics) from diverse classes each contributed less than 1% to total diacritics used.

Table 5 includes the results of Spearman rho calculations for all available pairwise, rank-order comparisons of the proportional occurrence of diacritics shown in Figure 5 (including an entry for data set H). The absolute magnitudes of most of these coefficients (11 of the 30 coefficients account for over 50% of variance; 28 of the 30 are statistically significant at the 0·05 level or better) suggest that the proportional occurrence distributions are variably stable within (intra-judge) and between (inter-judge) individual transcribers or teams, subject groups, and sampling modes. The generally parallel trends in each of the panels in Figure 5 support the following generalization:

> The proportional occurrence of individual diacritic symbols in narrow phonetic transcription ranges from low to moderately high depending on consensus transcription teams, subject groups, and sampling modes.

As with the absolute rates of diacritic use, these data on proportional individual diacritic use address a possible constraint on interpretation of the narrow transcription agreement data to follow. Specifically, although the proportional occurrence of diacritics is not perfectly comparable across the transcription teams, subject groups,

Table 5. *Spearman rho coefficients for all available inter-judge and intra-judge comparisons of the proportional occurrence of diacritics*[a]

	A[b]			*B*			*C*	*D*			*E*			*F*	*G*	*H*
Transcriber/ team	*CS*[c]	*AT*	*CS& AT*	*CS*	*AT*	*CS& AT*	*CS*	*CS*	*AT*	*CS& AT*	*CS*	*AT*	*CS& AT*	*CS*	*CS*	*CS*
I			0·40[d]													
II																0·92
III			0·73	0·87	0·89	0·59					0·88	0·47	0·44			
IV			0·66				0·90	0·66	0·79	0·18				0·90		
V			0·73													
I&III	0·74	0·57														
I&IV	0·53	0·64														
I&V	0·42	0·53														
III&IV	0·60	0·66														
III&V	0·32	0·37														
IV&V	0·43	0·62														
Tran 6															0·89	
Tran 6&7															0·58	

[a] All coefficients and *p* values are corrected for ties. [b] Data sets A-H are described in Table 3. [c] CS = Continuous speech; AT-articulation test. [d] Entries lower than 0·35 are non-significant; 0·35–0·50 yields $p < 0·05$; and above 0·50 yields $p < 0·01$.

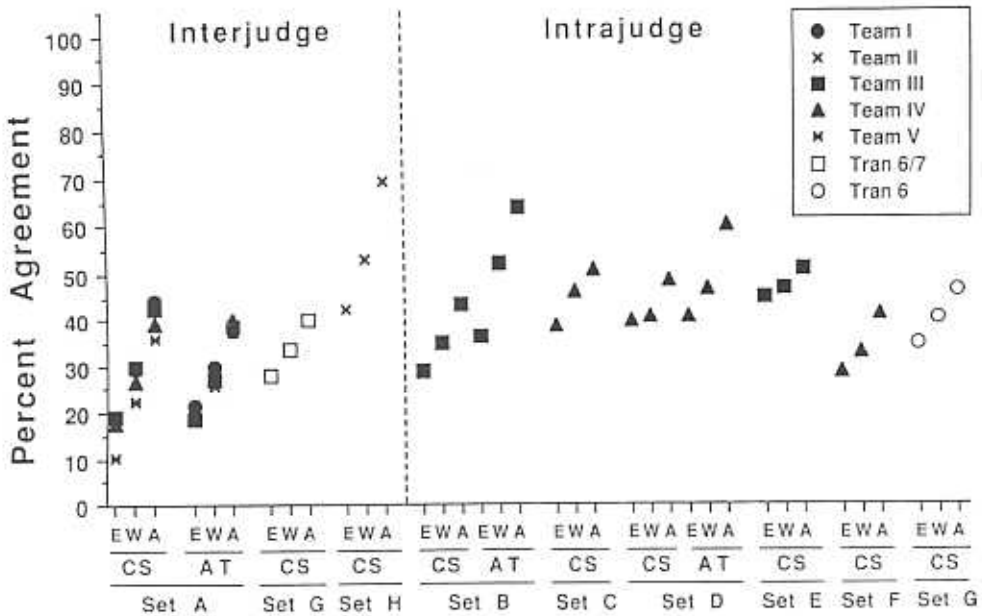


Figure 6. *Diacritic agreement for all consensus transcription teams and two individual transcribers. For each data set (A–H) and sampling mode (Continuous Speech, Articulation Test), agreement is shown for Exact, Within-class, and Any diacritic criteria (see text for explanation).*

and sampling modes, diacritic data will be averaged across these levels in certain analyses.

Diacritic agreement. Figure 6 and Table 6 are summaries of the diacritic agreement data. Figure 6 provides overall diacritic agreement for transcription teams and individual transcribers, with independent variables including data set (A–H) and mode of sampling (*C*ontinuous *S*peech, *A*rticulation *T*est). Dependent variables include point-to-point reliability percentages based on *E*xact agreement (same diacritic), *W*ithin-class agreement (any diacritic within the same diacritic class), and *A*ny diacritic (any diacritic, regardless of diacritic class). The data in Table 6 provide this information at the level of individual diacritics, including exact, within-class, and any diacritic agreement information. The 35 diacritics are arranged in the same descending order of proportional occurrence as in the top left panel in Figure 5. The following two generalizations are based on the trends in Figure 6 and Table 6, as well as some additional analysis to be described:

> Transcription agreement on an individual diacritic is essentially independent of its proportional occurrence in a speech sample.
>
> The average inter-judge and intra-judge percentage of agreement estimates for diacritic transcription are below acceptable reliability boundary levels, even at the least strict agreement criteria.

The first generalization is apparent on visual examination of the data in the rank-ordered proportional occurrence of diacritics (Table 6, second column) in relation to the remaining columns for percentage of agreement. These comparisons indicate only low positive agreement in rank orderings. Spearman rank-order correlations between proportional occurrence and mean agreement percentages in each sampling mode yielded rho values of 0·32 with Articulation Test and 0·29 with Continuous

Table 6. *Individual diacritic agreement data arranged by decreasing proportional occurrence across all datasets. All means and ranges are percentages of inter-judge and intra-judge agreement.*

		Exact agreement					*Within-class agreement*					*Any diacritic agreement*				
		Articulation test		*Continuous speech*		*Both modes*	*Articulation test*		*Continuous speech*		*Both modes*	*Articulation test*		*Continuous speech*		*Both modes*
Diacritic	*Proportional occurence (%)*	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Mean*	*Range*	*Mean*	*Range*	*Mean*	*Mean*	*Range*	*Mean*	*Range*	*Mean*
Derhotacized	12·1	69·8	54·5–96·8	67·4	49·9–83·0	68·1	70·4	54·5–96·8	67·9	52·0–84·5	68·7	73·1	54·5–100	70·1	52·0–87·0	71·0
Dentalized	10·6	41·4	25·3–54·5	55·9	17·6–85·7	51·6	58·2	31·2–88·9	64·2	24·7–85·7	62·4	66·3	46·4–88·9	71·7	44·1–100	70·1
Palatalized	9·7	45·5	33·3–66·7	57·9	32·6–95·0	54·2	78·0	76·0–80·4	70·5	37·2–95·0	72·8	81·3	76·0–90·2	74·5	41·9–95·0	76·6
Glotalized	5·8	0	—	18·7	0–52·2	16·4	0	—	20·2	0–52·2	17·7	10·4	—	25·3	0–55·1	23·5
Weak	5·6	17·8	0–33·3	27·7	18·2–40·0	24·8	17·8	0–33·3	27·7	18·2–40·0	24·8	32·9	16·5–50·0	39·2	27·4–46·7	37·3
Centralized	5·0	16·2	0–28·6	33·0	11·4–60·0	27·9	39·7	16·2–60·0	45·1	28·6–60·0	43·5	46·6	36·9–60·0	50·0	28·6–64·5	49·0
Off glide	4·8	58·4	43·0–70·6	37·4	0–64·5	43·7	63·6	45·1–76·5	37·7	0–64·5	45·4	73·5	58·3–85·3	49·4	25·0–73·7	56·7
Denasalized	4·3	24·7	0–60·0	33·8	0–78·0	31·1	25·5	0–60·0	33·8	0–78·0	31·3	46·6	16·7–80·0	51·0	16·7–87·5	49·7
Frictionalized	3·1	31·5	0–54·5	48·4	0–75·0	43·3	33·8	6·9–54·5	49·9	11·1–75·0	45·1	50·2	18·0–72·7	56·6	28·6–79·2	54·7
Unaspirated	3·0	73·4	66·7–83·3	44·7	0–78·3	55·4	78·3	66·7–84·9	44·2	0–78·3	57·0	78·8	66·7–86·3	50·8	0–78·3	61·3
Unreleased	3·0	50·0	0–100	10·8	0–50·8	22·0	50·0	0–100	11·3	0–54·2	19·9	50·0	0–100	13·8	0–55·9	21·8
Lowered tongue	3·0	33·3	15·7–50·0	35·6	0–100	34·9	40·4	21·2–50·0	48·6	8·7–10·0	46·2	41·3	23·8–50·0	53·8	8·7–100	50·0
Raised tongue	2·8	28·6	0–54·5	43·2	0–83·3	38·8	42·2	33·3–54·5	49·3	11·1–90·0	47·2	42·6	33·3–54·5	50·5	11·1–90·0	48·1
Lateralized	2·7	59·0	0–90·9	59·5	0–90·2	52·7	85·4	71·4–94·0	75·4	0–100	78·8	100	—	77·8	0–100	85·2
Onglide	2·7	46·3	33·1–61·5	32·3	0–61·5	36·5	52·0	39·9–71·8	33·2	0–61·5	38·9	53·7	39·9–76·9	37·7	0–69·2	42·5
Retracted tongue	2·5	35·4	0–56·2	32·9	0–54·5	33·8	72·8	50·0–100	36·8	0–57·7	48·8	73·1	50·0–100	41·6	0–63·2	52·1
Nasalized	2·0	23·0	0–40·0	22·6	0–75·0	22·7	24·1	0–40·0	22·6	0–75·0	23·1	31·9	0–55·8	37·4	0–87·5	35·8
Backed	2·0	6·7	0–20·0	17·8	0–40·0	14·1	8·1	0–20·0	18·6	0–50·0	15·1	51·9	30·0–60·0	29·6	0–60·0	37·0
Lengthened	2·0	13·7	0–28·6	24·2	0–66·7	17·4	24·8	12·4–33·3	30·3	0–66·7	28·6	46·1	28·7–66·7	38·9	0–83·3	41·1
Non-labialized	1·9	9·5	0–18·9	35·7	0–66·7	33·3	17·8	16·7–18·9	38·1	0–66·7	33·0	37·5	25·0–50·0	42·3	0–66·7	41·1
Devoiced	1·9	24·6	24·2–25·0	10·9	0–34·8	14·3	25·5	25·0–25·4	11·6	0–34·8	15·0	40·3	30·7–50·0	25·1	0–50·0	28·9
Fronted	1·7	16·7	0–25·0	39·3	0–75·0	32·5	18·5	5·6–25·0	41·9	0–75·0	34·9	51·2	22·2–75·0	63·9	25·0–83·3	60·1
Labialized	1·2	14·1	0–28·6	22·1	0–66·7	19·7	18·9	0–42·9	26·9	0–100	24·5	22·2	0–42·9	29·3	0–100	27·2
Aspirated	1·2	65·1	28·5–100	8·8	0–52·6	27·5	64·5	26·9–100	13·0	0–63·2	30·2	68·8	28·5–100	31·9	0–63·2	44·2
Breathy	0·9	0	—	3·7	0–22·2	3·2	0	—	7·4	0–27·8	6·3	8·3	—	15·7	0–50·0	14·7
Syllabic	0·9	8·3	—	72·0	0–100	61·4	8·3	—	72·0	0–100	61·4	15·2	—	72·0	0–100	62·5
Nasal emission	0·7	33·0	—	61·7	0–100	56·0	33·3	—	61·7	0–100	56·0	72·2	—	66·7	0–100	67·8
Inverted	0·5	0	—	63·1	0–100	42·1	0	—	63·1	0–100	42·1	25·0	0–50·0	63·1	0–100	50·4
Rhotacized	0·4	21·3	0–50·0	37·5	0–100	30·6	69·7	62·5–77·7	77·5	50·5–100	74·1	91·6	87·4–100	77·5	50·0–100	83·6
Shortened	0·4	50·0	0–100	28·8	0–100	35·9	50·0	0–100	28·8	0–100	35·9	50·0	0–100	51·6	23·1–100	51·1
Velarized	0·3	7·0	0–22·2	10·0	0–40·0	8·9	16·4	0–33·3	25·0	0–100	21·3	28·9	0–55·6	25·0	0–100	26·7
Advanced tongue	0·3	3·4	0–6·7	0	—	1·0	7·8	0–15·6	30·0	0–100	23·7	7·8	0–15·6	31·6	0–100	24·8
Whistled	0·3	38·9	11·1–66·7	0	—	13·0	38·9	11·1–66·7	0	—	13·0	66·7	44·4–88·9	7·1	0–28·6	27·0
Synchronic tie	0·3	70·0	40·0–100	40·0	—	60·0	85·0	70·0–100	40·0	—	70·0	95·0	90·0–100	80·0	—	90·0
Partially voiced	0·2	0	—	33·3	0–100	20·0	10·4	0–20·8	33·3	0–100	24·2	10·4	0–20·8	36·8	0–100	26·3

Speech, neither of which was significant at the 0·05 level. Thus, although the proportional use of individual diacritics ranged from just a few occurrences to over 12% of the total diacritic occurrences, the relative difficulty in making reliable perceptual decisions on diacritics was not predictably associated with the frequency of making those judgements.

For the present concern with the overall reliability of narrow phonetic transcription of diacritics, the second generalization provides a serious challenge to the use of unaided perceptual transcription for clinical and research tasks in communicative disorders. As shown in Figure 6, the average agreement figures are well below 70%, even at the most liberal level of the three response definitions (i.e. with agreement based on the inclusion of any diacritic). Additional comment on this finding is offered at the conclusion of this section.

Sounds

Associations with percentage of occurrence and percentage correct. Before considering the consonant and vowel agreement data it is necessary to examine the potential associations between transcription agreement for sounds and their percentage of occurrence and percentage correct articulation. A total of 36 Spearman rho coefficients were computed, comparing for each sample the rank ordering of consonant and vowel agreement with the rank orderings of both consonant and vowel occurrence (18 coefficients) and consonants and vowels correct (18 coefficients). Table 7 is a summary of these data. The magnitudes of the coefficients for all comparisons were generally low, with only 3 of the 36 coefficients (8%) significant at the 0·05 level, which is essentially the expected occurrence by chance.

Figure 7 provides sound-level information for the three variables, Percentage Correct, Percentage of Occurrence, and Percentage of Agreement, collapsed across studies and sorted by decreasing percentage of transcription agreement. As shown in the lowest trends in Figure 7, which indicate the average percentage of occurrence of each consonant and vowel, respectively, occurrence percentages do not covary

Table 7. *Spearman rho coefficients[a] for broad transcription agreement with percentage occurrence and percentage correct*

	Percentage occurrence		*Percentage correct*	
Agreement	*Consonants*	*Vowels*	*Consonants*	*Vowels*
Inter-judge				
Set A	−0·40	−0·31	0·32	0·40
Set G	−0·68*	0·13	−0·05	0·39
Set H	−0·28	−0·31	0·26	−0·25
Intra-judge				
Set B	−0·50*	0·24	0·39	0·11
Set C	−0·18	−0·12	0·34	−0·13
Set D	0·32	−0·25	0·47*	−0·44
Set E	−0·28	−0·47	0·16	0·30
Set F	−0·49	−0·14	0·20	0·36
Set G	−0·34	−0·03	0·39	0·18

[a] All coefficients and p values are corrected for ties. *$p < 0·05$.

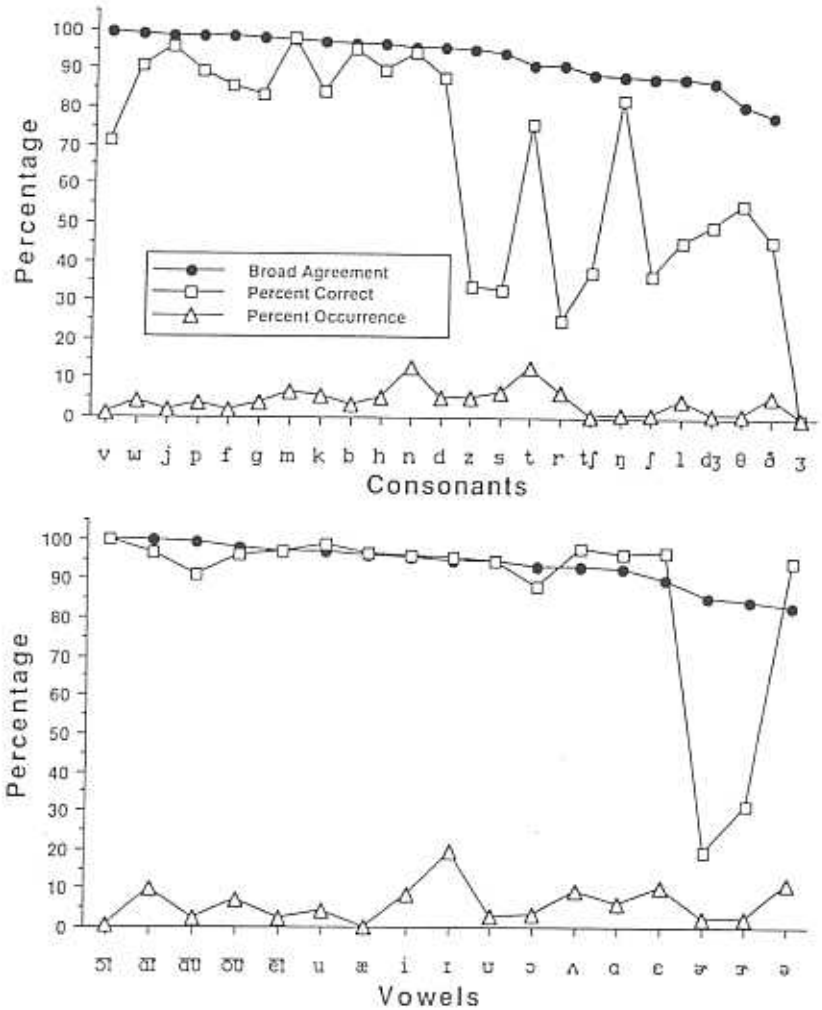


Figure 7. *Broad transcription agreement for consonants (top panel) and vowels (bottom panel) in relation to each sounds' percentage of occurrence and percentage correct.*

with transcription agreement percentages. For the consonant data, however, there is a discernible split between the first and second 12 sounds. Except for /v/, the 12 sounds with the highest broad transcription agreement are also articulated at approximately 85% or above correct. In contrast, except for /t/ and /ŋ/, a second group of 12 sounds have considerably lower percentages of correct articulation and also have the 12 lowest average transcription agreement percentages. These data in Table 7 and Figure 7 suggest the following generalization:

> The reliability of broad transcription of vowels in a sample is essentially independent of their rank-order of occurrence and percentage correct. For consonants, transcription agreement is independent of rank-order of occurrence and lower, but within an acceptable range for the 12 most frequently misarticulated sounds.

Consonant and vowel agreement. Inter-judge and intra-judge reliability data for the 24 consonants and 17 vowels are presented in Tables 8 and 9, respectively; summary

Table 8. *Consonant agreement for all data sets; all numbers are percentages*

		Inter-judge											
		Set A										*Set G*	*Set H*
		Team I		*Team III*		*Team IV*		*Team V*		*Mean*		*Tran 6&7*	*Team II*
Sound	*Level*	*CS*[a]	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
/m/	B[b]	98·5	100·0	98·5	100·0	97·0	100·0	100·0	100·0	98·5	100·0	97·0	100·0
	N	77·2	97·0	82·9	90·2	87·7	97·0	66·1	96·3	78·5	95·1	91·0	96·6
/n/	B	91·9	88·9	91·9	94·0	89·6	94·1	96·3	92·5	92·4	92·4	93·1	100·0
	N	76·0	63·4	76·0	56·5	80·2	63·7	81·8	60·4	78·5	61·0	85·0	85·7
/ŋ/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0
	N	100·0	75·0	100·0	91·7	100·0	91·7	100·0	91·7	100·0	87·5	100·0	100·0
/w/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	94·3	100·0
	N	68·8	88·9	60·4	88·9	45·9	66·7	66·7	88·9	60·4	83·4	84·6	70·4
/j/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0
	N	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	97·5	66·7
/p/	B	93·3	100·0	93·3	100·0	86·7	100·0	100·0	100·0	93·3	100·0	98·1	100·0
	N	82·2	63·3	82·2	63·3	70·0	70·0	61·1	63·3	73·9	65·0	90·6	50·0
/b/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	93·5	100·0
	N	85·3	70·4	95·2	80·4	95·2	81·5	94·9	81·3	92·7	78·4	89·7	56·5
/t/	B	91·7	93·8	92·6	95·6	92·6	93·4	76·9	90·1	88·4	93·2	95·3	100·0
	N	54·2	66·7	53·7	76·4	38·1	67·8	30·2	71·0	44·1	70·5	81·0	56·3
/d/	B	90·7	93·3	90·7	73·9	89·7	82·2	91·7	91·7	90·7	85·3	90·0	100·0
	N	57·9	33·3	57·9	30·6	33·4	52·8	43·6	33·3	48·2	37·5	69·1	52·6
/k/	B	93·1	98·1	83·8	99·0	87·7	98·0	89·2	99·0	88·4	98·5	98·6	100·0
	N	78·5	86·0	61·6	88·2	58·3	80·3	55·8	86·2	63·6	85·2	82·1	60·0

/g/	B	100·0	100·0	100·0	93·9	100·0	97·0	100·0	99·0	100·0	97·0	97·7	100·0
	N	66·7	75·9	66·7	67·9	66·7	73·8	66·7	75·2	66·7	73·2	93·0	56·3
/f/	B	100·0	95·2	72·2	95·2	88·9	95·2	83·3	85·7	86·1	92·9	100·0	100·0
	N	83·3	84·1	27·8	83·0	77·8	79·4	66·7	76·2	63·9	80·7	90·9	81·8
/v/	B	100·0	88·9	100·0	77·8	100·0	77·8	—	100·0	100·0	86·1	100·0	100·0
	N	100·0	66·7	83·4	66·7	83·4	22·2	—	44·5	88·9	50·0	76·9	88·9
/θ/	B	33·3	88·4	75·0	91·1	75·0	87·8	0·0	100·0	45·8	91·8	100·0	100·0
	N	33·3	85·7	75·0	89·7	75·0	84·9	0·0	100·0	45·8	90·1	83·3	50·0
/ð/	B	33·3	0·0	66·7	0·0	50·0	0·0	50·0	0·0	50·0	0·0	63·2	85·0
	N	16·7	0·0	50·0	0·0	0·0	0·0	0·0	0·0	16·7	0·0	52·9	77·5
/s/	B	91·7	100·0	91·7	100·0	91·7	100·0	75·0	100·0	87·5	100·0	98·6	97·5
	N	80·0	35·5	80·0	38·1	76·7	36·5	50·0	28·6	71·7	34·7	47·2	75·0
/z/	B	83·3	92·6	83·3	96·3	75·0	92·6	100·0	92·6	85·4	93·5	97·1	100·0
	N	48·3	33·3	47·0	42·9	38·7	37·5	31·7	38·7	41·4	38·1	32·4	77·8
/ʃ/	B	100·0	93·3	100·0	82·5	100·0	89·2	100·0	73·3	100·0	84·6	100·0	100·0
	N	33·3	33·3	33·3	41·7	0·0	61·7	0·0	46·7	16·7	45·8	57·1	55·6
/ʒ/	B	—	—	—	50·00	—	50·00	—	0·00	—	33·33	—	—
	N	—	—	—	—	—	—	—	—	—	—	—	—
/h/	B	97·2	100·0	97·2	100·0	94·5	100·0	100·0	100·0	97·2	100·0	100·0	95·5
	N	91·7	100·0	91·7	100·0	83·3	100·0	100·0	100·0	91·7	100·0	94·1	90·5
/tʃ/	B	100·0	100·0	100·0	100·0	100·0	100·0	—	100·0	100·0	100·0	100·0	100·0
	N	100·0	—	100·0	—	100·0	—	—	—	100·0	—	66·7	100·0
/dʒ/	B	—	100·0	—	100·0	—	100·0	—	100·0	—	100·0	100·0	100·0
	N	—	—	—	—	—	—	—	—	—	—	50·0	60·0
/l/	B	100·0	91·3	78·3	82·0	86·7	87·6	91·7	92·4	89·2	88·3	90·2	90·2
	N	50·0	51·2	33·3	49·8	33·3	40·8	50·0	33·9	41·7	43·9	66·7	50·0
/r/	B	91·7	75·9	95·8	81·4	95·8	85·7	100·0	80·6	95·8	80·9	92·2	100·0
	N	40·5	44·1	30·1	41·4	38·1	34·7	38·9	52·8	36·9	43·3	58·7	70·0

Table 8 (continued)

		Intra-judge								
		Set B		*Set C*	*Set D*		*Set E*		*Set F*	*Set G*
Sound	*Level*	*CS*	*AT*	*CS*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
/m/	B[b]	95·6	100·0	100·0	100·0	100·0	95·7	—	94·6	98·0
	N	75·0	55·6	86·1	95·2	92·3	84·4	—	86·5	91·0
/n/	B	97·5	100·0	99·5	94·6	100·0	99·1	100·0	96·2	91·1
	N	82·1	88·9	85·2	83·8	79·2	88·7	0·0	80·4	83·8
/ŋ/	B	100·0	100·0	50·0	66·7	80·0	87·5	—	100·0	90·0
	N	0·0	0·0	50·0	66·7	60·0	75·0	—	80·0	85·7
/w/	B	100·0	100·0	97·8	100·0	100·0	100·0	—	100·0	100·0
	N	88·9	66·7	79·1	78·6	85·7	87·0	—	66·7	88·5
/j/	B	100·0	100·0	88·2	100·0	100·0	100·0	—	100·0	100·0
	N	100·0	100·0	81·3	100·0	100·0	100·0	—	91·7	95·0
/p/	B	100·0	100·0	100·0	100·0	91·7	100·0	—	100·0	96·2
	N	84·6	88·9	87·1	72·7	83·3	93·1	—	91·3	90·6
/b/	B	100·0	92·9	87·2	90·9	100·0	100·0	100·0	100·0	100·0
	N	92·3	92·9	90·9	90·9	75·0	92·6	100·0	100·0	89·7
/t/	B	95·5	95·0	90·1	82·1	88·2	92·1	100·0	87·2	89·0
	N	88·3	84·2	60·4	52·9	64·7	61·3	66·7	67·6	73·5
/d/	B	100·0	100·0	96·2	100·0	100·0	100·0	100·0	95·0	89·1
	N	73·7	100·0	82·9	62·5	85·7	82·8	100·0	83·3	70·9
/k/	B	90·9	100·0	100·0	100·0	97·0	100·0	—	100·0	95·7
	N	75·0	90·0	87·0	80·0	57·6	87·1	—	70·0	82·1

/g/	B	100·0	100·0	98·3	92·9	88·9	100·0	—	100·0	95·3
	N	90·9	100·0	78·0	75·0	55·6	93·8	—	90·0	88·4
/f/	B	100·0	88·9	100·0	100·0	88·9	100·0	—	100·0	100·0
	N	100·0	55·6	93·3	83·3	88·9	83·3	—	100·0	95·5
/v/	B	100·0	100·0	100·0	100·0	0·0	100·0	—	100·0	94·1
	N	—	100·0	100·0	100·0	0·0	90·0	—	100·0	76·9
/θ/	B	100·0	100·0	100·0	—	0·0	100·0	—	0·0	100·0
	N	100·0	100·0	90·0	—	0·0	—	—	0·0	83·3
/ð/	B	89·5	—	80·0	100·0	—	66·7	—	80·0	89·5
	N	73·7	—	77·8	100·0	—	33·3	—	80·0	76·5
/s/	B	97·0	87·5	84·5	94·1	90·9	100·0	100·0	92·9	97·2
	N	38·7	28·6	47·4	60·0	36·4	86·7	—	60·0	62·7
/z/	B	100·0	100·0	85·7	100·0	100·0	96·2	—	100·0	94·3
	N	0·0	50·0	57·1	88·9	56·0	91·7	—	75·0	44·1
/ʃ/	B	66·7	100·0	85·7	66·7	87·5	—	100·0	100·0	85·7
	N	33·0	33·3	71·4	50·0	50·0	—	—	100·0	57·1
/ʒ/	B	—	—	—	—	—	—	—	—	—
	N	—	—	—	—	—	—	—	—	—
/h/	B	100·0	100·0	100·0	100·0	100·0	87·5	—	90·9	100·0
	N	100·0	100·0	100·0	91·7	100·0	84·6	—	90·0	97·1
/tʃ/	B	100·0	75·0	100·0	0·0	100·0	100·0	—	100·0	100·0
	N	75·0	25·0	100·0	0·0	60·0	80·0	—	100·0	66·7
/dʒ/	B	—	100·0	100·0	33·3	83·3	75·0	100·0	100·0	100·0
	N	—	50·0	75·0	0·0	33·3	75·0	—	100·0	100·0
/l/	B	63·6	90·0	93·7	90·9	88·9	100·0	—	85·7	87·2
	N	25·0	12·5	87·9	70·0	66·7	100·0	—	63·6	71·0
/r/	B	81·8	90·1	87·8	85·7	70·0	92·9	100·0	92·3	90·7
	N	66·7	81·8	73·2	80·0	0·0	36·4	50·0	54·5	54·3

[a] CS = Continuous speech; AT = Articulation test. [b] B = Broad transcription; N = Narrow transcription.

Table 9. *Vowel agreement for all data sets. All numbers are percentages*

		Inter-judge											
		Set A										*Set G*	*Set H*
		Team I		*Team III*		*Team IV*		*Team V*		*Mean*		*Tran 6&7*	*Team II*
Sound	*Level*	*CS*[a]	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
/i/	B[b]	85·9	75·0	88·0	81·1	88·0	85·8	76·7	86·9	84·6	82·2	100·0	100·0
	N	60·6	71·7	63·4	76·9	35·5	81·7	53·3	86·9	53·2	79·3	94·1	88·2
/I/	B	66·9	73·1	84·6	74·5	85·6	78·7	88·4	93·1	81·4	79·8	95·7	100·0
	N	54·4	62·6	74·3	68·1	63·2	72·3	63·0	80·6	63·7	70·9	86·0	97·4
/ɛ/	B	37·5	93·3	71·3	95·5	65·0	95·5	0·0	100·0	43·4	96·1	99·0	97·8
	N	20·8	72·5	58·8	66·5	52·5	72·4	0·0	55·9	33·0	66·8	89·7	86·7
/æ/	B	91·7	91·6	95·8	95·2	95·8	95·2	100·0	97·9	95·8	95·0	98·8	100·0
	N	66·7	52·9	66·0	61·4	29·3	54·1	66·7	66·1	57·2	58·6	82·7	77·3
/ɚ/	B	66·7	—	66·7	—	0·0	—	100·0	—	58·3	—	90·0	100·0
	N	55·6	—	55·6	—	0·0	—	100·0	—	52·8	—	70·0	22·2
/ɔ/	B	62·1	74·1	80·1	88·4	83·7	86·5	76·7	82·4	75·6	82·9	97·2	100·0
	N	62·1	73·2	80·1	80·4	83·7	82·2	76·7	80·2	75·6	79·0	91·5	100·0
/ɔ˞/	B	—	85·8	—	96·7	—	96·7	—	95·8	—	93·8	100·0	100·0
	N	—	55·0	—	60·3	—	41·5	—	71·8	—	57·2	77·8	57·1

/ʌ/	B	91·1	95·5	93·0	95·5	93·0	95·5	100·0	100·0	94·3	96·7	100·0	100·0			
	N	60·0	88·9	47·9	88·6	69·0	83·8	27·8	92·3	51·2	88·4	96·3	93·9			
/u/	B	100·0	100·0	100·0	95·8	100·0	95·8	100·0	91·7	100·0	95·8	100·0	100·0			
	N	91·7	91·5	88·9	77·7	79·4	85·6	93·3	74·4	82·3	82·3	63·0	92·9			
/o/	B	66·7	100·0	83·3	100·0	50·0	100·0	100·0	100·0	75·0	100·0	100·0	100·0			
	N	33·3	83·3	33·3	83·3	0·0	53·3	0·0	86·7	16·7	76·7	87·5	100·0			
/ɔ/	B	72·2	94·4	91·7	88·6	91·7	95·2	88·9	93·3	86·1	92·9	94·1	100·0			
	N	61·1	75·5	75·0	76·3	75·0	76·3	44·4	60·0	63·9	72·1	88·2	88·9			
/ɑ/	B	94·5	93·9	94·2	96·7	94·2	96·7	100·0	90·0	95·7	94·3	97·4	100·0			
	N	72·2	49·4	70·9	60·6	59·8	63·4	77·8	46·7	70·2	55·0	89·7	95·8			
/ɑɪ/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	99·2	100·0			
	N	76·2	78·3	75·8	85·0	70·4	85·0	60·0	75·0	70·6	80·8	76·3	90·6			
/ɑʊ/	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	93·8	100·0			
	N	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	93·8	80·0			
/eɪ/	B	100·0	86·9	100·0	87·6	100·0	91·3	100·0	92·8	100·0	89·6	95·2	100·0			
	N	83·3	65·0	83·3	69·0	83·3	62·4	50·0	53·8	75·0	62·6	90·5	90·9			
/oʊ/	B	93·9	100·0	97·0	100·0	97·0	100·0	100·0	100·0	97·0	100·0	100·0	100·0			
	N	84·3	100·0	87·0	100·0	77·3	100·0	73·3	100·0	80·5	100·0	82·7	77·8			
/ɔɪ/	B	—	—	—	—	—	—	—	—	—	—	100·0	—			
	N	—	—	—	—	—	—	—	—	—	—	100·7	—			

Table 9 *(continued)*

		Intra-judge								
		Set B		*Set C*	*Set D*		*Set E*		*Set F*	*Set G*
Sound	*Level*	*CS*	*AT*	*CS*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
/i/	B	100·0	100·0	87·7	100·0	100·0	96·6	100·0	95·2	96·2
	N	90·0	100·0	78·9	94·1	83·3	89·7	50·0	94·7	82·4
/I/	B	100·0	100·0	91·6	97·6	81·0	95·1	100·0	94·3	98·3
	N	93·8	85·7	82·4	78·0	70·0	87·4	100·0	88·7	80·7
/ɛ/	B	100·0	100·0	87·5	95·2	100·0	90·7	100·0	100·0	97·9
	N	92·5	85·7	70·8	81·0	84·6	76·7	50·0	79·2	78·4
/æ/	B	100·0	100·0	81·7	95·2	96·3	97·6	100·0	96·8	98·8
	N	70·8	64·3	46·7	57·1	77·8	85·7	0·0	90·3	69·1
/ɝ/	B	75·0	—	64·7	100·0	—	—	—	100·0	90·0
	N	50·0	—	35·3	66·7	—	—	—	50·0	65·0
/ə/	B	100·0	100·0	85·4	88·2	81·8	77·8	—	91·7	98·6
	N	90·0	100·0	82·9	78·6	80·0	68·3	—	77·3	87·3
/ɚ/	B	66·7	100·0	94·7	—	100·0	83·3	—	66·7	88·9
	N	50·0	83·3	84·2	—	100·0	16·7	—	33·3	44·4

/ʌ/	B	95·8	100·0	76·9	93·8	100·0	100·0	—	80·8	98·8
	N	91·7	100·0	69·2	93·8	84·6	93·1	—	69·2	86·6
/u/	B	95·2	75·0	89·3	100·0	100·0	100·0	100·0	100·0	88·9
	N	71·4	75·0	71·4	66·7	87·5	90·6	10·0	55·6	55·6
/ʊ/	B	100·0	100·0	90·0	100·0	100·0	100·0	—	87·5	100·0
	N	75·0	100·0	80·0	85·7	100·0	75·0	—	75·0	75·0
/ɔ/	B	100·0	100·0	90·5	100·0	75·0	100·0	100·0	75·0	97·1
	N	100·0	60·0	52·4	83·3	75·0	100·0	100·0	62·5	82·4
/ɑ/	b	100·0	100·0	88·9	66·7	81·8	95·8	—	94·4	97·4
	N	100·0	100·0	77·8	58·3	81·8	95·8	—	88·9	89·7
/a͡ɪ/	B	100·0	100·0	100·0	100·0	100·0	100·0	—	100·0	100·0
	N	90·0	100·0	84·9	87·5	100·0	90·9	—	76·0	72·0
/a͡ʊ/	B	100·0	100·0	100·0	100·0	100·0	100·0	—	100·0	100·0
	N	66·7	100·0	61·5	80·0	100·0	75·0	—	100·0	75·0
/e͡ɪ/	B	100·0	100·0	89·7	91·7	100·0	100·0	100·0	100·0	97·6
	N	83·3	83·3	82·1	83·3	81·8	94·1	100·0	94·4	83·3
/o͡ʊ/	B	100·0	100·0	90·9	100·0	100·0	96·7	—	100·0	100·0
	N	100·0	100·0	69·1	78·6	100·0	76·7	—	91·7	80·0
/ɔ͡ɪ/	B	—	—	100·0	—	—	100·0	—	—	100·0
	N	—	—	100·0	—	—	100·0	—	—	100·0

[a] CS = Continuous speech; AT = Articulation test. [b] B = Broad transcription; N = Narrow transcription

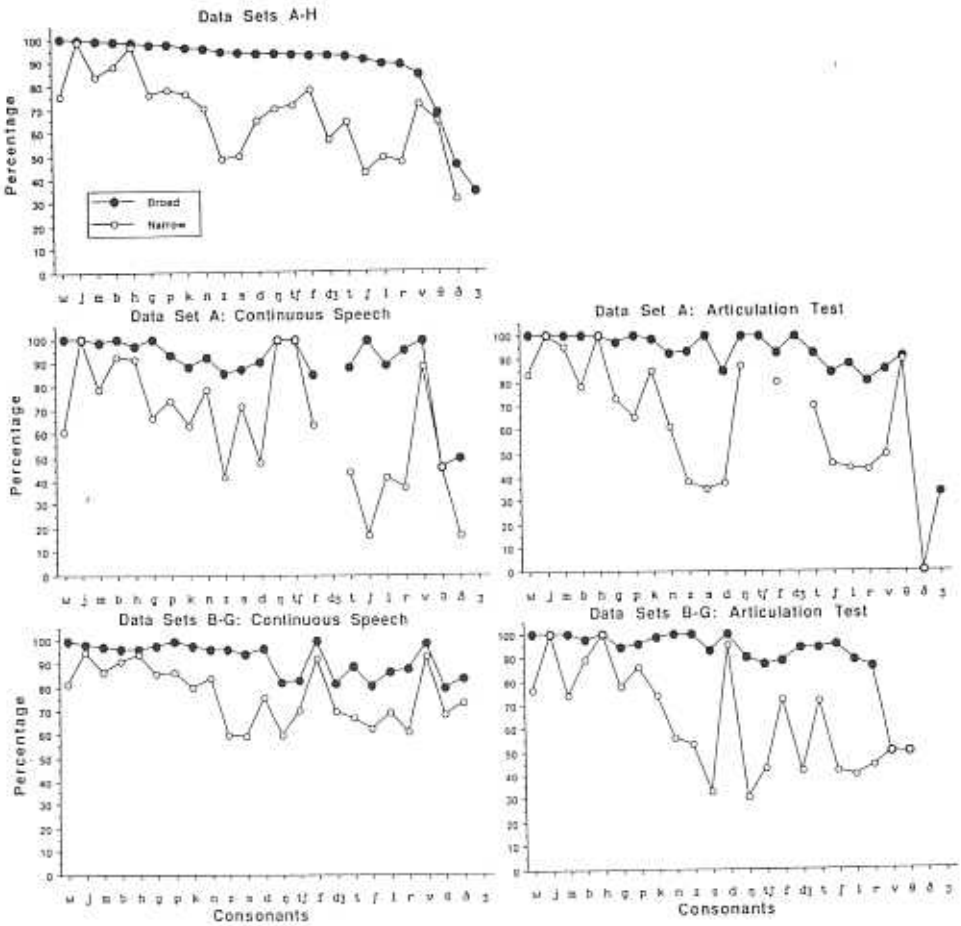


Figure 8. *Transcription agreement for consonants. The top panel includes averaged agreement across data sets (A–H), transcription teams, sampling modes (continuous speech, articulation test), and type of reliability (inter-judge, intra-judge). The middle panels include inter-judge agreement for data set A in continuous speech (left panel) and articulation test (right panel) modes. The bottom panels include intra-judge agreement for data sets B–G in continuous speech (left panel) and articulation test (right panel) modes.*

trends are provided in Figures 8 and 9. The figures are arranged to display segments by decreasing transcription agreement, whereas for archival reference the tables preserve the statistical data by sequencing segments within manner class. The data from these figures and tables suggest the following generalization:

> Average transcription agreement percentages for each of 41 sounds are within acceptable levels for broad transcription, but generally below acceptable ranges for narrow phonetic transcription.

The only phoneme-level transcription agreement data available in the literature for comparison to the rank-ordered consonant data in Table 8 and Figure 8 are data for 97 children 4–6 years old, reported by Norris *et al.* (1980). Although the methodology for computing agreement differed considerably from the present study, a Spearman rank-order coefficient was computed using the consonant agreement

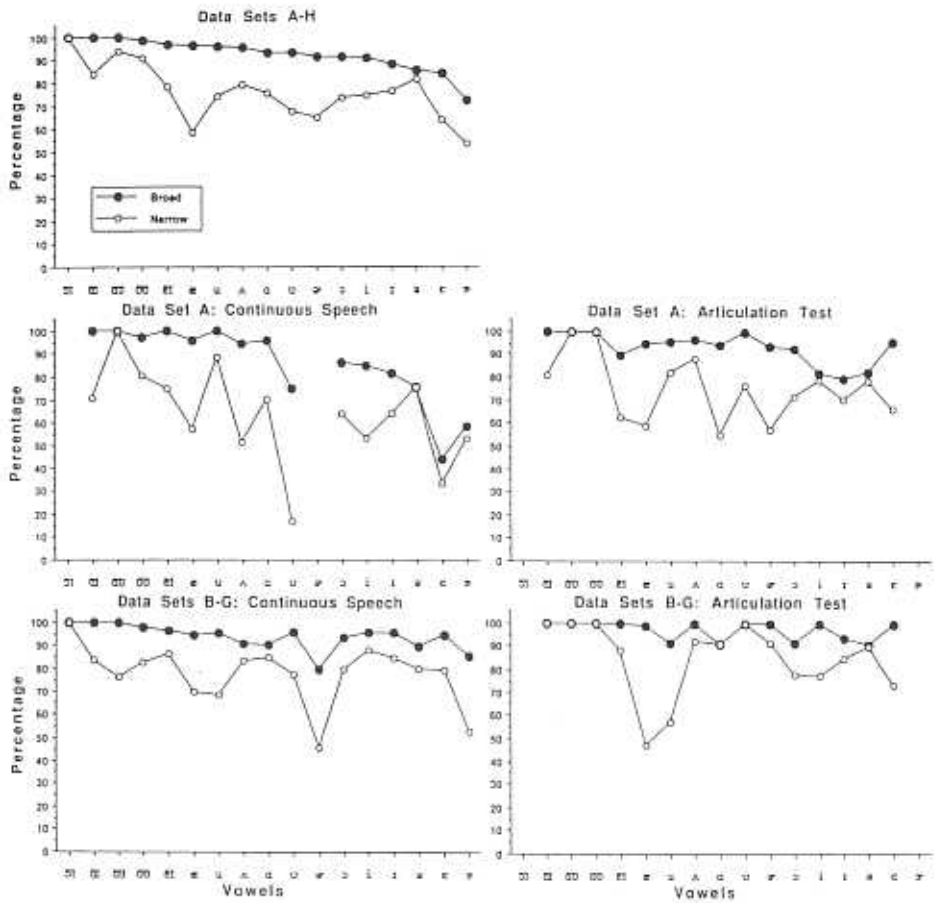


Figure 9. *Transcription agreement for vowels. The format for this figure is similar to that described in the legend for Fig. 8.*

data in Norris *et al.* (Table 2) and the present average agreement data in Table 8. The obtained coefficient of 0·29 was not significant.

Turning to the magnitude of agreement for consonants and vowels, 21 of the 24 consonant sounds (88%) (Figure 8, top panel) and 16 of the 17 vowels (91%) (Figure 9, top panel) had acceptable agreement (at least 85% perfect agreement, cf. Irwin, 1970; Pye *et al.*, 1988) when transcribed using broad phonetic transcription. Using the same criteria, only 3 of the 24 consonants (13%) and 9 of the 17 vowels (53%) had acceptable agreement when transcribed using narrow phonetic transcription. As shown in the other panels in Figures 8 and 9, and in the more detailed data in Tables 8 and 9, agreement percentages for individual sounds varied considerably when broken out by the several independent variables.

Taken together with the previous findings for diacritics, these data indicate that transcribers generally agree at the level of phoneme transcription but not at the level of acceptable and unacceptable allophone transcription. The most direct explanation for the low levels of narrow phonetic transcription agreement is that children's speech productions frequently contain confusing acoustic cues relative to the phoneme and allophone boundaries expected by the ambient community. Weismer's

(1984a,b) reviews of related issues and a number of subsequent studies indicate that such ambiguous and sometimes chimeric acoustic events defy reliable assignment to nominal categories (e.g. Riley *et al.*, 1986). Considering that subjects in the parent study are exactly those children whose delayed phonologic development make them likely to be producing a high rate of such behaviours (Roberts, Burchinal and Footo, 1990), these reliability data are viewed as realistic reflections of the perceptual limits of narrow phonetic transcription.

The acoustic ambiguity explanation for the low levels of narrow transcription agreement is testable using the type of validity studies proposed at the outset of this paper. Specifically, this interpretation would predict that intra-judge and inter-judge agreements would vary significantly as a function of the relative ambiguity of acoustic cues. Considering the context of the present data sets it is unlikely that significant sources of explanation for the lowered agreement figures can be found in alternative transcriber factors, such as limitations in their training, experience, hearing acuity,

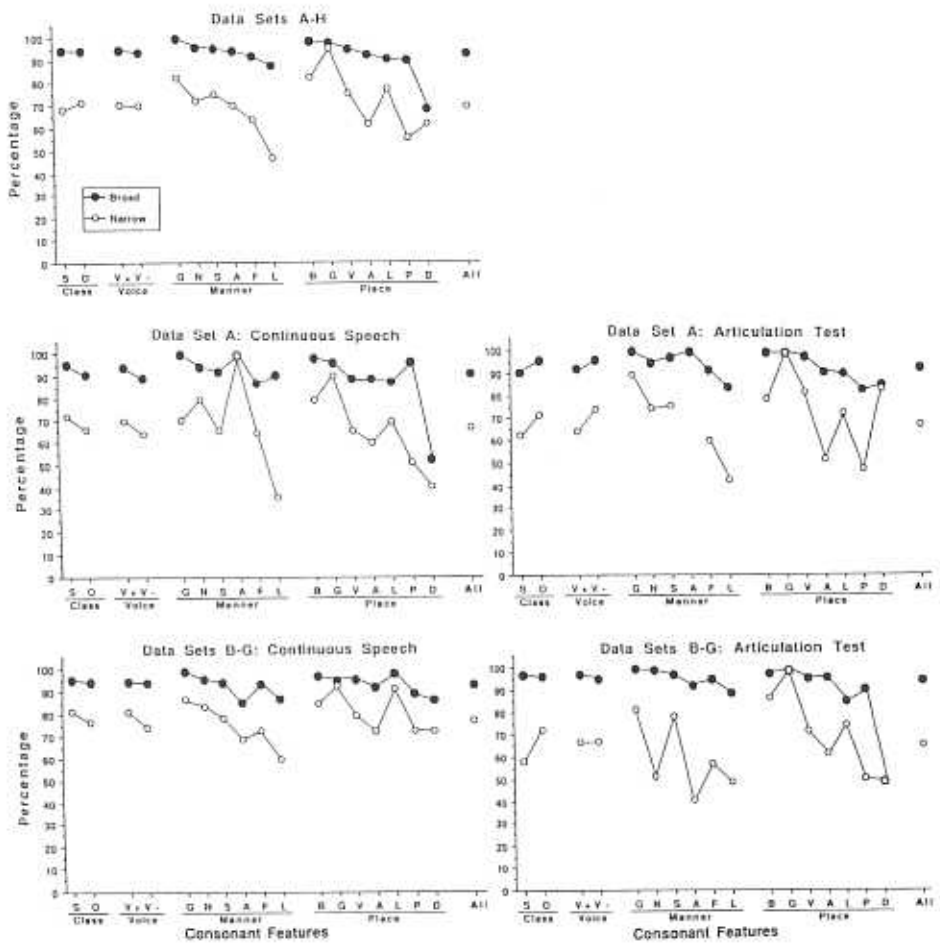


Figure 10. *Transcription agreement for consonant features including class (Sonorant, Obstruent), voice (V+ = Voiced, V− = Voiceless), manner (Glide, Nasal, Stop, Affricate, Fricative, Liquid), and place (Bilabial, Glottal, Velar, Alveolar, Labiodental, Palatal, Dental). The format for this figure is similar to that described in the legend for Fig. 8.*

memory, effort, number of allowable repetitions, playback environment, or other factors that might be reasonably invoked elsewhere when low reliability estimates are obtained.

Features and classes

The final unit-level sources of variability in Table 1 are the constructs of phonetic features (place, manner, voicing) and the superordinate construct of phonetic classes (sonorants, obstruents). Figures 10 and 11 and Tables 10 and 11 include the transcription agreement findings for consonants and vowels, respectively. These data suggest the following generalization:

> Average transcription agreement at the level of phonetic features and classes is within acceptable levels for broad transcription and generally below acceptable levels for narrow phonetic transcription.

These feature- and class-level data simply reflect the previous sound-level findings. Sound-level differences shown in Figs 8 and 9 and Tables 8 and 9 are apparently

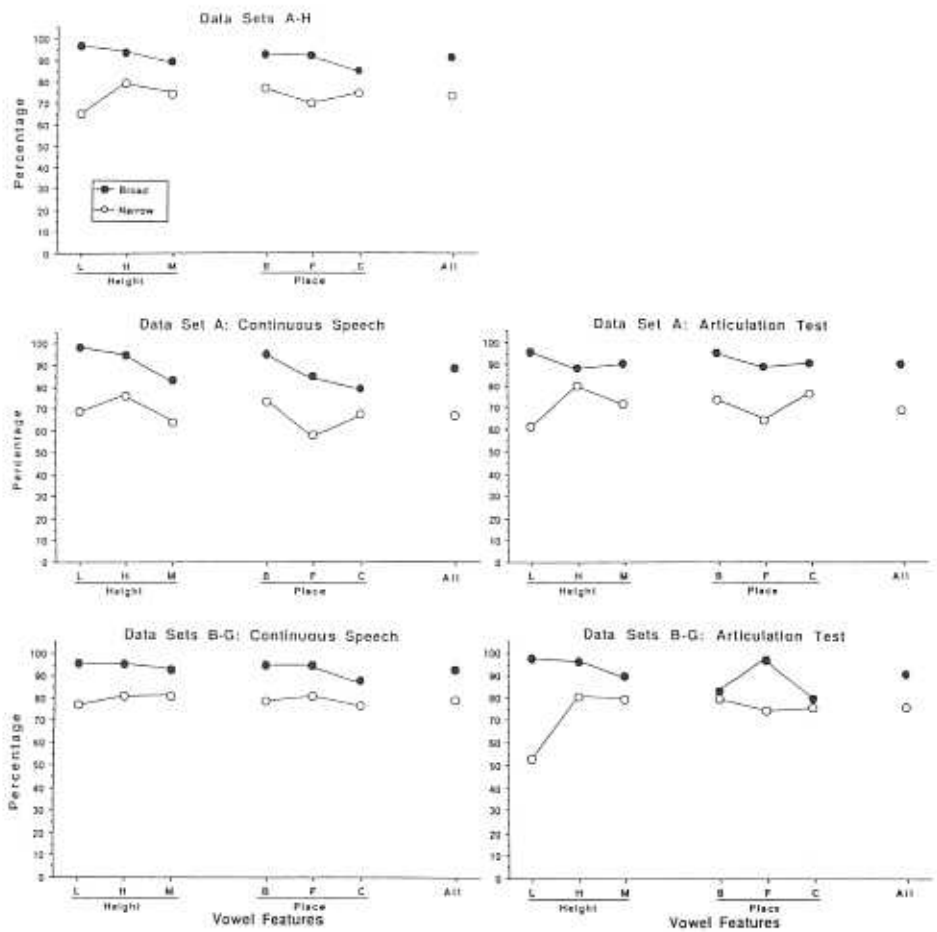


Figure 11. *Transcription agreement for vowel features including height (Low, High, Middle) and place (Back, Front, Central). The format for this figure is similar to that described in the legend for Fig. 8.*

Table 10. *Consonant feature agreement for all data sets; all numbers are percentages*

		Inter-judge											
		Set A										*Set G*	*Set H*
		Team I		*Team III*		*Team IV*		*Team V*		*Mean*		*Tran 6&7*	*Team II*
Feature	*Level*	*CS*[a]	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
Class													
Sonorant	B[b]	95·7	89·4	93·9	89·8	93·7	92·1	97·3	91·8	95·2	90·8	94·5	99·5
	N	71·8	64·0	70·5	60·9	72·5	60·3	72·4	63·1	71·8	62·1	83·7	79·2
Obstruent	B	91·9	96·5	91·3	95·6	91·0	95·6	90·4	95·5	91·1	95·8	95·9	97·8
	N	69·7	69·8	70·0	73·2	62·6	70·5	61·4	72·0	66·0	71·4	72·3	68·3
Voice													
Voiced	B	94·0	92·0	93·5	91·4	92·6	93·0	97·0	93·4	94·3	92·5	93·6	98·1
	N	69·2	63·7	71·8	63·8	69·1	63·0	71·1	65·3	70·3	61·0	79·6	74·9
Voiceless	B	92·0	96·8	90·1	96·6	91·0	96·3	85·3	95·5	89·6	96·3	97·6	98·7
	N	73·0	73·6	67·2	76·6	61·3	72·8	54·1	74·4	63·9	74·4	73·2	67·2
Manner													
Nasal	B	94·5	93·1	94·5	96·4	92·7	96·4	97·5	95·7	94·8	95·4	94·8	100·0
	N	78·0	74·7	79·9	71·5	83·9	77·1	77·6	76·4	79·9	74·9	87·6	89·4
Glide	B	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	100·0	96·8	100·0
	N	75·0	93·3	70·6	93·3	58·9	80·0	77·8	93·3	70·6	90·0	90·2	69·7
Stop	B	94·7	98·1	92·7	97·1	92·8	97·1	91·3	97·5	92·9	97·4	95·6	100·0
	N	69.6	74·0	70·5	77·7	60·7	75·6	63·5	77·8	66·1	76·3	82·7	55·9

Fricative	B	86·7	93·2	88·7	92·4	87·5	92·4	88·4	91·4	87·8	92·3	96·4	95·7
	N	68·8	60·6	68·0	63·5	65·8	58·8	57·4	59·2	65·0	60·6	57·5	79·2
Affricate	B	100·0	100·0	100·0	100·0	100·0	100·0	—	100·0	100·0	100·0	100·0	100·0
	N	100·0	—	100·0	—	100·0	—	—	—	100·0	—	60·0	75·0
Liquid	B	94·7	83·6	85·7	81·6	90·0	86·4	95·8	86·5	91·6	84·5	91·3	98·3
	N	42·6	47·4	27·4	45·0	33·8	37·8	41·7	42·6	36·4	43·2	62·3	66·7
Place													
Bilabial	B	98·8	100·0	98·8	100·0	97·7	100·0	100·0	100·0	98·8	100·0	96·2	100·0
	N	77·0	75·5	83·7	79·5	80·1	81·4	80·5	80·9	80·3	79·3	89·3	71·6
Labiodental	B	100·0	93·9	77·8	91·1	94·4	90·8	83·3	89·6	88·9	91·4	100·0	100·0
	N	91·7	79·5	41·7	78·9	83·3	65·8	66·7	68·5	70·8	73·2	85·7	85·0
Dental	B	33·3	78·4	72·2	85·2	58·4	85·2	50·0	95·8	53·5	86·2	72·0	85·7
	N	25·0	74·4	61·1	86·1	29·2	81·9	50·0	95·2	41·3	84·4	60·9	76·2
Alveolar	B	91·7	92·6	90·0	91·4	88·8	92·2	89·8	92·8	90·1	92·2	94·7	99·3
	N	66·3	54·1	64·3	54·0	57·2	52·4	55·9	48·9	61·0	52·3	69·1	70·0
Palatal	B	95·3	84·3	97·6	84·1	97·6	88·0	100·0	80·1	97·6	84·1	96·1	100·0
	N	57.7	47·1	51·7	45·4	54·8	45·2	43·3	54·6	51.9	48·1	74·5	68·5
Velar	B	94·3	98·6	86·1	97·9	89·3	97·9	91·1	98·5	90·2	98·3	98·4	100·0
	N	72·3	82·8	65·4	83·8	64·9	79·8	62·7	84·1	66·3	82·7	87·2	63·0
Glottal	B	97·2	100·0	97·3	100·0	94·5	100·0	100·0	100·0	97·2	100·0	100·0	95·5
	N	91·7	100·0	91·7	100·0	83·3	100·0	100·0	100·0	91·7	100·0	94·1	90·5
Summary	B	93·4	94·1	92·3	93·6	92·1	94·4	93·1	94·2	92·7	94·1	95·3	98·3
	N	70·5	67·9	70·2	69·2	66·6	67·1	65·6	69·2	68·2	68·3	76·8	71·9

Table 10 *(continued)*

		Intra-judge								
		Set B		*Set C*	*Set D*		*Set E*		*Set F*	*Set G*
Feature	*Level*	*CS*	*AT*	*CS*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
Class	B									
Sonorant	B	93·4	96·2	95·7	95·1	93·2	97·7	100·0	95·2	93·8
	N	75·0	66·7	82·5	84·8	75·0	84·7	33·3	78·7	83·1
Obstruent	B	95·9	96·1	92·9	93·3	92·9	96·5	100·0	93·8	94·1
	N	76·0	75·6	73·3	75·2	61·4	81·1	80·0	79·3	75·4
Voice										
Voiced	B	94·6	97·1	94·7	94·3	94·3	97·0	100·0	95·3	93·7
	N	76.1	72·3	80·5	83·0	68·2	85·5	60·0	81·6	80·3
Voiceless	B	95·5	94·9	92·9	93·5	91·4	95·7	100·0	93·0	94·3
	N	75·0	74·0	71·0	72·9	61·2	77·3	66·7	75·0	75·9
Manner										
Nasal	B	96·6	100·0	98·8	95·1	97·7	97·4	100·0	95·8	93·5
	N	75·6	75·0	84·9	86·9	81·0	86·8	0·0	82·8	86·5
Glide	B	100·0	100·0	95·2	100·0	100·0	100·0	—	100·0	100·0
	N	95·2	75·0	79·7	87·0	88·9	90·6	—	81·0	91·3
Stop	B	96·6	97·3	94·6	93·3	94·9	97·3	100·0	94·9	92·5
	N	84·8	90·4	76·8	72·2	67·3	79·3	80·0	80·2	79·7

Fricative	B	94·4	95·7	89·3	97·2	90·7	95·6	100·0	91·3	96·5
	N	58·8	58·7	66·2	83·6	56·0	85·9	—	76·2	69·1
Affricate	B	100·0	87·5	100·0	25·0	90·9	90·9	100·0	100·0	100·0
	N	75·0	37·5	84·6	0·0	45·5	77·8	—	100·0	80·0
Liquid	B	72·7	90·5	90·3	88·9	78·9	95·8	100·0	88·9	89·1
	N	47·1	52·6	79·3	73·3	46·2	41·7	50·0	59·1	61·0
Place										
Bilabial	B	97·7	97·1	96·4	98·5	98·1	98·4	100·0	97·6	98·3
	N	81·0	80·0	83·4	83·8	82·7	88·7	100·0	87·8	90·2
Labiodental	B	100·0	92·9	100·0	100·0	80·0	100·0	—	100·0	97·4
	N	100·0	71·4	92·5	90·0	80·0	87·5	—	100·0	88·6
Dental	B	90·5	100·0	90·0	100·0	0·0	80·0	—	72·7	92·3
	N	76·2	100·0	84·2	100·0	0·0	33·3	—	72·7	78·3
Alveolar	B	94·7	96·1	92·9	91·9	95·2	97·3	100·0	92·2	91·6
	N	71·1	66·2	71·7	71·9	61·5	80·3	60·0	72·6	71·7
Palatal	B	90·0	91·3	89·1	78·3	84·4	94·1	100·0	96·9	94·3
	N	78·6	60·9	75·4	70·0	44·0	69·0	50·0	79·3	72·4
Velar	B	94·6	100·0	97·9	94·6	93·6	98·4	—	100·0	95·1
	N	73·3	88·5	82·1	77·1	57·4	87·3	—	77·1	84·6
Glottal	B	0·0	100·0	100·0	100·0	100·0	87·5	—	90·9	100·0
	N	100·0	100·0	100·0	91·7	100·0	84·6	—	90·0	97·1
Summary	B	95·0	96·1	94·1	94·0	93·0	96·9	100·0	94·4	94·0
	N	75·6	73·0	77·1	78·8	64·9	82·4	62·5	79·0	78·4

[a] CS = Continuous speech; AT = Articulation test. [b] B = Broad transcription; N = Narrow transcription.

Table 11. *Vowel feature agreement for all data sets; all numbers are percentages*

		Inter-judge											
		Set A										*Set G*	*Set H*
		Team I		*Team III*		*Team IV*		*Team V*		*Mean*		*Tran 6&7*	*Team II*
Feature	*Level*	*CS*[a]	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
Height													
High	B[b]	94·6	86·8	96·0	88·5	96·0	90·6	88·8	94·9	88·7	100·0	100·0	
	N	79·7	81·0	79·5	77·0	63·7	82·9	81·3	79·7	76·1	80·2	83·3	89·6
Middle	B	73·7	87·4	87·0	90·4	85·3	91·7	88·5	94·6	83·6	91·0	97·4	99·6
	N	60·3	71·8	71·4	74·3	67·1	70·9	57·2	70·9	64·0	72·0	88·2	88·7
Low	B	96·8	93·6	97·6	96·3	97·6	96·3	100·0	95·8	98·0	95·5	98·4	100·0
	N	73·6	56·3	73·2	65·0	60·7	62·1	66·7	62·1	68·6	61·4	81·5	87·5
Place													
Front	B	77·4	85·8	88·9	88·6	88·3	90·7	88·4	95·0	85·8	90·1	97·7	99·5
	N	57·9	63·9	68·8	67·2	50·8	65·3	57·3	64·3	58·7	65·2	87·8	90·5
Central	B	72·8	87·0	82·6	93·8	82·7	92·9	84·1	93·4	80·6	91·8	97·8	100·0
	N	61·5	75·0	70·4	79·0	73·2	72·1	68·2	84·0	68·3	77·5	90·7	84·8
Back	B	92·9	97·2	96·5	96·3	96·0	97·2	98·4	94·1	95·9	96·2	98·2	100·0
	N	76·8	74·7	77·8	76·5	70·7	77·4	69·8	68·0	73·8	74·2	81·0	88·6
Summary	B	84·4	88·9	91·8	91·7	90·8	92·8	93·4	94·2	90·1	91·9	97·9	99·8
	N	68·1	69·0	73·6	72·3	64·3	70·1	65·2	69·7	67·8	70·3	85·9	88·6

		Intra-judge								
		Set B		*Set C*	*Set D*		*Set E*		*Set F*	*Set G*
Feature	*Level*	*CS*	*AT*	*CS*	*CS*	*AT*	*CS*	*AT*	*CS*	*CS*
Height										
High	B	97·6	90·0	88·2	100·0	100·0	98·4	100·0	96·7	93·8
	N	80·5	90·0	76·5	84·6	85·7	90·2	66·7	82·1	73·1
Middle	B	97·3	100·0	86·6	95·7	91·8	91·8	80·0	92·2	97·9
	N	88·9	88·1	74·9	81·0	81·1	81·0	70·0	80·3	80·8
Low	B	100·0	100·0	90·5	89·1	93·0	98·5	100·0	97·5	99·2
	N	80·7	77·3	68·8	65·2	81·4	88·7	0·0	86·4	74·0
Place										
Front	B	100·0	100·0	88·3	96·4	94·4	95·3	100·0	96·6	97·9
	N	86·6	80·4	73·5	77·7	78·4	85·9	66·7	89·0	78·2
Central	B	92·7	100·0	81·1	92·3	92·3	86·3	50·0	86·0	97·3
	N	83·3	95·5	73·5	83·3	84·0	73·7	50·0	69·1	82·4
Back	B	98·7	95·7	91·8	93·4	89·7	97·1	66·7	95·4	98·2
	N	85·5	87·0	73·8	75·4	87·2	86·8	66·7	79·3	75·9
Summary	B	97·9	98·9	87·7	94·8	92·9	94·5	85·7	94·2	97·9
	N	85·5	85·7	73·6	78·0	81·6	84·3	64·3	82·2	78·2

[a] CS = Continuous speech; AT = Articulation test. [b] B = Broad transcription; N = Narrow transcription.

eliminated at the level of consonant classes (sonorants, obstruents) and voiced–voiceless features, but some differences are additive at manner and place levels. Particularly for the narrow phonetic transcription data, feature percentages reflecting similar manners and places of production range from acceptable to unacceptable levels across studies. Consonant glides, nasals, and stops are generally, but not always, associated with higher agreement levels than affricates, fricatives, and liquids. These feature-level data were compared to feature-level agreement data reported by Norris *et al.* (1980) and Philips and Bzoch (1969). Due to fundamental differences in methodology, no clear decisions could be made about the agreement in percentage trends across the three studies.

Contexts

Target environment

The analysis software did not have the capability to compute transcriber agreement at the level of the environments of sounds, such as consonants occurring in singleton versus two-element and three-element clusters or consonants in relation to the height of adjacent vowels. Pye *et al.* (1988) reported that approximately one-half to two-thirds of disagreements involving deletions occurred in the context of consonant clusters in both initial and final positions. Their analyses further indicated some interactions between the word position of the cluster and the member of the cluster deleted, with strongest effects for /s/ clusters. Such information is also important when examining differences in transcription agreement associated with mode of sample. Specifically, certain articulation tests deliberately include more words with clusters than occur in a comparable number of words in a continuous speech sample (Shriberg, 1986; Shriberg and Kwiatkowski, 1980).

Word position

Figure 12 provides broad and narrow transcription agreement data for word position, with target consonants divided into word-initial, word-medial, and word-final position. The software defined word-medial as all non-initial or non-final consonants. Thus, medial consonants included all interior members of two- and three-element clusters, as well as intervocalic consonants. The trends in Figure 12 suggest the following generalization:

> Of the three word positions, word-initial consonants are generally transcribed most reliably, with word-final consonants typically associated with the lowest reliability.

Notice that, for narrow phonetic transcription, the data in Figure 12 indicate that final position is always lowest in transcription agreement. Similar positional effects have been reported by Philips and Bzoch (1969) and Pye *et al.* (1988). A likely explanation for the lowered agreement on final sounds is that more errors occur in that position (Edwards and Shriberg, 1983). Moreover, whereas initial fricative errors are more likely to be phonemic substitutions, final fricative errors are more likely to be distortions (but see later discussion of error type in relation to transcriber agreement).

Structural, grammatical, and stress forms

The software did not permit inspection of the agreement data in relation to such potentially interactive word characteristics as structural (canonical) form, grammatical form, and lexical stress or syllabic stress, each being a variable subsumed under

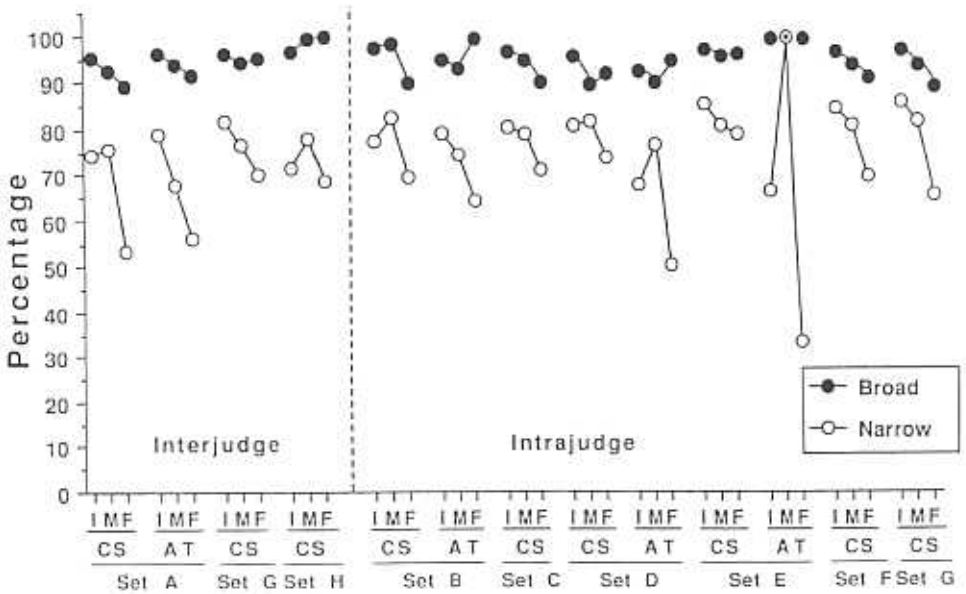


Figure 12. *Transcription agreement for word position (Initial, Medial, Final).*

category 10 in Table 1. For example, articulation responses (citation forms) are realized primarily as stressed nouns, with canonical forms generally more complex than the simple forms associated with function words that occur in continuous speech (Shriberg and Kwiatkowski, 1980; Morrison and Shriberg, (in press)). Such variables might be associated with transcription agreement data. As noted earlier, the striking contrast between the range of consensus team's DPWs in continuous speech compared to articulation test responses (see Figure 4, Sets A, B, and D) may be associated with reliable differences in the types and frequencies of forms and stress patterns found in the two sampling modes.

Of particular importance for subsequent investigation are potential associations between the stress level of a sonorant and transcribers' perceptual boundaries for acceptable articulation. Whereas Hoffman, Schuckers and Ratusnik (1977), using live transcription, found unstressed vocalic /ɚ/ to be the most often correctly articulated allophone, Curtis and Hardy (1959), using recorded analysis, found unstressed vocalic /ɚ/ to be the least often correctly articulated /r/ allophone. McCauley and Skenes (1987) and Shriberg (1972) have described the auditory differences in primary versus non-primary stress that may moderate such findings. Specifically, higher correct scores on unstressed vocalic /ɚ/ may reflect situations in which lower intensity levels associated with more lenient criteria for an acceptable rhotic quality. As suggested earlier, such differences associated with the acoustic characteristics of the presentation media are considered to be in the domain of validity.

Sampling mode

The final environment source of variance in Table 1 is the sampling mode, which in this study contrasts transcription agreement for continuous speech samples with agreement based on responses to articulation test stimuli. Data bearing on sampling mode have been kept separate in prior tables and figures, to allow for specific comparisons within all other variables. As described above, the data in Figure 4

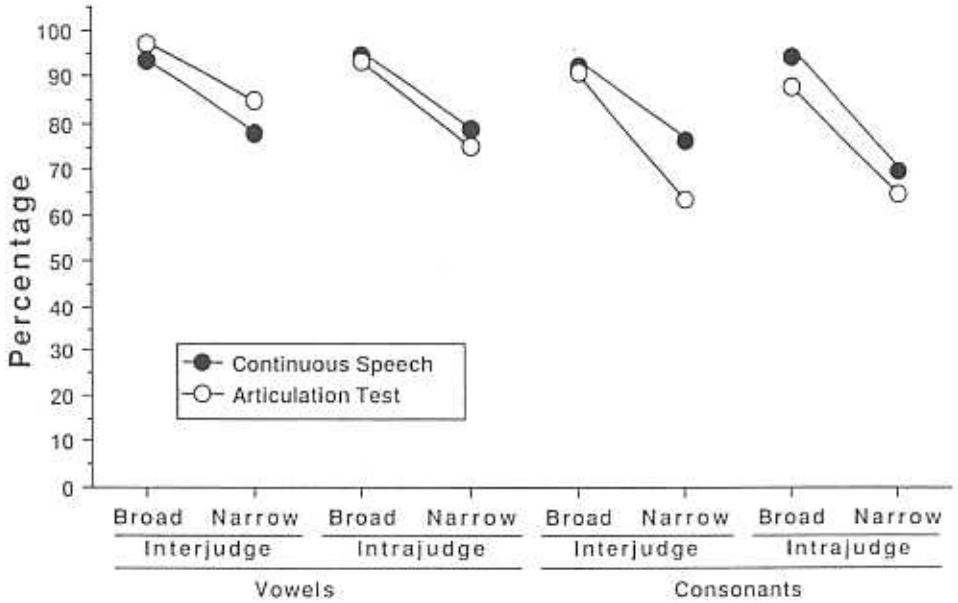


Figure 13. *Percentage agreement for broad and narrow phonetic transcription in two sampling modes.*

suggest that diacritics may be used more often when transcribing whole words from articulation tests. For the overall descriptive analysis shown in Figure 13, however, these lower-level sources of variance (study, transcription team, sounds) were collapsed. The trends in Figure 13 support the following generalization:

> Transcription agreement based on continuous speech samples is somewhat higher than agreement based on articulation test responses.

In six of the eight comparisons shown in Figure 13, average agreement for the continuous speech samples is approximately 1–13 percentage points higher than the averaged data for agreement based on articulation test responses. The two reversals in these trends occur on the inter-judge agreement for vowels, in which agreement based on articulation responses is approximately 4 and 6 percentage points higher for broad and narrow transcription, respectively. That these general trends are reliable is supported by inspection of the individual consonant and vowel agreement data in Tables 8 and 9. The six trends and the two reversals are consistent with observations elsewhere that relate the number and type of misarticulated sounds to structural characteristics of each type of sampling mode. Detailed exploration of such factors is reported in Morrison and Shriberg (in press), including the observation that children's over-articulation of citation forms may create spurious allophones for narrow phonetic transcription.

Analyses

Type, systems and agreement criteria

The data presented in the tables and figures to this point provide considerable information on transcription agreement figures associated with type of agreement

(inter-judge, intra-judge), type of system (broad, narrow) and agreement criteria within systems (exact, within-class, any diacritic). The following generalizations are based on the data used to construct Figure 13:

> The two traditional types of transcription agreement, inter-judge and intra-judge, have essentially similar average percentages of agreement, ranging from the mid-60s to the mid-high-90s.
>
> The two systems of phonetic transcription, broad (93%) and narrow (74%), differ in average transcription agreement by approximately 20 points.
>
> The three types of transcription agreement criteria for diacritics, exact (33%), within-class (40%), and any diacritic (48%), differ in average transcription agreement (uncorrected for chance agreement) by a range of approximately 15 points.

Each of these three generalizations is based on data averaged over the eight subject samples, nine transcription teams/individuals, two sampling modes, and each of the other two variables addressed in this section.

The generalization that inter-judge and intra-judge agreement have essentially similar percentages of agreement is surprising considering the more typical finding that intra-judge agreement is associated with higher transcription agreement. The current study differs from most other reports of intra-judge reliability, however, because in most of the present cases retest agreement is based on responses from consensus teams. From these data it is not possible to infer whether the similarity in mean percentages is due to inter-judge agreement being higher or intra-judge agreement being lower than figures typically reported in the literature.

The finding of a difference of approximately 20 percentage points between broad and narrow phonetic transcription is consistent with the spread of differences found in other studies. Amorosa *et al.* (1985) have the most comparable data; in two studies of narrow phonetic transcription they report average inter-judge agreements of 70% and 74%.

The third generalization, describing low levels of diacritic agreement, however strict the criterion, suggests an important perspective on transcribers' difficulty with diacritics. Dating back to Henderson (1938), agreement on two-way decisions (correct versus incorrect) has been shown to be higher than on five-way scoring (correct, deletion, substitution, distortion, addition), presumably due to the increased complexity of the decision process and the lack of systematic response definitions for substitutions, additions, and distortions (Irwin, 1970; Irwin and Krafchick, 1965; Norris *et al.*, 1980; Philips and Bzoch, 1969). Even when agreement in the present study is based on the lenient agreement criterion of *any* diacritic to mark a distortion, it averages only 48%. Thus, the major source of the problem seems not to be in puzzling over which diacritic to use, but whether to use any diacritic at all.

Transcribers

The tables and figures have retained the agreement data at the level of individual transcribers and consensus transcription teams. Inspection of this information suggests trends for certain of the five teams to have higher inter-judge agreement with all other teams; however, no firm trends emerged. Lowered agreement figures for the one off-site consensus team may have been associated with their reduced amount of initial and continuing training by the first author, with differences in their recording/playback equipment and environment, or with differences in other aspects of the transcription process. Neither these transcription data nor the anecdotal

records kept by consensus transcription teams suggests patterns among transcribers warranting a generalization from this study. Pye *et al.* (1988) also noted their inability to attribute patterns to the differences observed among the transcriptions of their three individual transcribers.

One possible pattern noted anecdotally, but difficult to document quantitatively, is the possibility of 'observer drift' in response definitions. The concept of observer drift in the behavioural sciences refers to a change that occurs in the response criteria learned by observers as they proceed from original calibration to a later period following considerable experience with the task. In the present case we noted a tendency for teams eventually to develop either a more or less stringent perceptual standard for certain frequently occurring targets, such as derhotacized /r/ and dentalized /s/. Other potential differences among individual and consensus transcribers, including personality and prior work experiences that might predict high competence, are discussed and illustrated in Shriberg *et al.* (1987). Based on a job sample task to select and train prospective transcribers for research transcription, findings suggested that competent phonetic transcription appears to require a sound technical grasp of articulatory phonetics, well-developed auditory–perceptual skills, and the temperament to persist at a difficult task (i.e. a positive attitude).

Subjects

Clinical significance

The software was programmed to count any point-to-point difference in transcription as a disagreement, even those differences in the presence/absence or occurrence of diacritics that add descriptive richness but are not associated with articulation errors (e.g. an unreleased final stop). Data presented on the proportional occurrence of diacritics (Figure 5) indicate that those diacritics marking non-error allophones were as frequently used as those marking articulation errors. For example, note (Figure 5 and Table 6) that dentalized and palatalized diacritics are among the three *frequent* and six *somewhat frequent* diacritics, respectively. The software guidelines for the speech analysis program format dentalized fricatives as articulation errors, but palatalized fricatives and dentalized stops as acceptable allophones. Together with the agreement data indicating a lack of association between the clinical significance of a diacritic and its average percentage of agreement, these data support the following generalization:

> Transcriber agreement is not associated with the clinical significance of diacritical description of speech.

Experience suggests that the consequence of one's transcription plays an important role in transcription effort and potential bias. Transcribers may work harder to transcribe those diacritics that affect a subject's severity score than they work on diacritics that add descriptive information but do not affect subjects' articulation scores. The present data do not support these interpretations of the findings in Shriberg *et al.* (1984), which used different agreement procedures. Perhaps a more direct test of this source of variance would use measures that reflect transcriber effort, such as number of replays, or the elapsed times associated with each of the two types of transcriber decisions.

Type of error

To this point the data separated by subject's aetiologic group (i.e. data sets A–H) suggest that transcriber agreement is not associated with aetiology. The final three subject variables—type of articulation error, severity of involvement, and intelligibility—are generally considered the most direct sources of variance in transcriber reliability. Accordingly, the remaining analyses attempt to discern which variables are most strongly associated with transcriber agreement and specific sources for the lowered agreement in narrow, compared to broad, phonetic transcription.

To provide for close inspection of the role of articulation error type, deletion, substitution, and distortion percentages were calculated in two ways for 48 of the 51 individual subjects whose complete continuous speech samples were available from the parent study. One set of percentages on each error type, termed the *absolute error type* percentages, reflected the magnitude of each of the three error types. These six sets of percentages (three for consonants and three for vowels), used the total number of consonant or vowel segments in each subject's sample as the denominator. The other set of percentages, termed the *relative error type* percentages, used the total number of *errors* in each child's speech sample as the denominator for each error type percentage. Thus, whereas the three absolute error type percentages reflect the proportion of each error type in the speech sample, the relative error type percentages reflect the percentage each error type contributed to the total number of errors in the sample.

Figure 14 includes these error type data, with panels for absolute error type percentages and relative error type percentages correlated separately for consonant agreement and vowel agreement. These busy figures retain considerable data for the inspection by the interested reader. In each panel the top trend is the averaged inter-judge/intra-judge agreement for narrow transcription of consonants and vowels, respectively. The Spearman rho coefficients adjacent to each of the error type trends reflect the correlations of each error type with transcription agreement. Thus, the size of the negative correlations indexes the degree to which subjects' rank-ordered average transcription agreement is associated with their rank-ordered percentages for each error type.

Five of the 12 (42%) coefficients in Figure 14 were statistically significant at the 0·05 level, but they range in magnitude from low to moderate. The trends for consonants in particular indicate that most children's error types were divided among the three categories, with distortions the most frequent error type for both consonants and vowels. For some children, as shown most clearly in the relative error type panels, nearly 100% of their errors consist of distortions on consonants or vowels. However, contrary to expectation given the previous data on narrow versus broad transcription and the lowered reliability of diacritics, distortion errors were not the error type most negatively associated with transcription agreement. Rather, the pattern of correlation coefficients and subject-level trends in Figure 14 provides no clear interpretation of the role of subject error type in transcriber agreement. For consonants, the absolute and relative percentages of deletion errors are apparently most associated with agreement, but the magnitudes of the negative coefficients account for only approximately 16–21% of the variance. For vowels, the magnitudes of association are even lower. Thus, although these unpartialled analyses indicate low to moderate associations between error type and transcriber agreement, they fail to provide useful insight towards a model of the primary sources of variance in transcriber agreement.

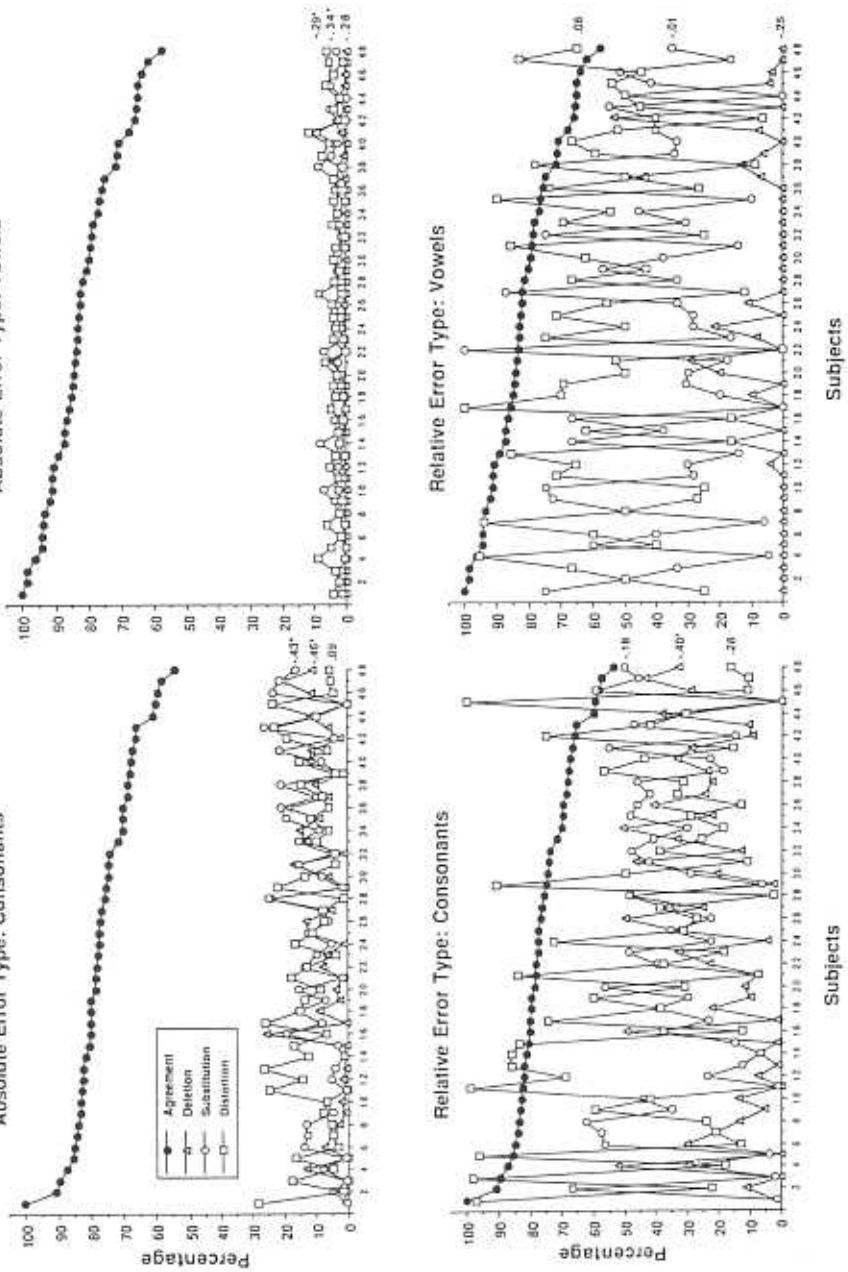


Figure 14. *Transcription agreement in relation to 35 subjects' absolute and relative articulation error types: deletions, substitutions, and distortions. The left panel provides data for consonants; the right panel for vowels. The values next to each of the error type trends are Spearman rho coefficients. See text for explanation of these apparently busy graphs.*

Additional analyses were undertaken on the possibility that the category *distortions* might be insensitive to individual effects within the 35 diacritics marking distortions. To pursue the hypothesis, the 48 subjects' distortion errors in the original speech samples were sorted and tallied using the seven diacritic classes described previously (see Figure 3 for a diacritic class example, Tongue Configuration). Spearman rho coefficients comparing the rank-ordered tallies within each diacritic class with subjects' rank-ordered consonant and vowel transcription agreement were calculated. Of the 14 coefficients (seven diacritic classes with consonant and vowel transcription agreement), four (29%) were significant at the 0·05 level, each reflecting an association of diacritics with transcription agreement for vowels: Lip = −0·35 ($p < 0·032$); Tongue Position = −0·51 ($p < 0·002$); Sound Source = −0·33 ($p < 0·044$); and Other = −0·53 ($p < 0·002$). The magnitude of these correlations does suggest that certain types of vowel distortion errors are particularly difficult to transcribe reliably. Similar findings have been reported in the literature, with subtle changes in vowels among the most difficult perceptual task for persons being trained in phonetic transcription (Shriberg and Kent, 1982).

The overall trends and correlation coefficients in Figure 14 and the additional analyses at the diacritic class level suggest the following generalization:

> Neither the absolute nor relative percentages of each of the primary error types—deletions, substitutions, or distortions—are highly associated with transcription agreement.

The wording of this generalization attempts to express a balanced view of the findings in relation to the *a priori* assumption that distortion errors are the most difficult to transcribe reliably. Although the primary and additional analyses did include some statistically significant coefficients associating distortions (especially on vowels) with transcriber agreement, the magnitudes of the coefficients were relatively modest. Moreover, each of the other two error types, deletions and substitutions, was at least moderately associated with transcriber agreement. Thus, although the source of the 20 percentage point average difference between broad and narrow phonetic transcription computationally implicates distortions, the subject-level data do not provide strong quantitative support for a unique difficulty associated with the use of diacritics to mark distortions.

Severity of involvement and intelligibility

The final two subject-level variables address possible associations between the degree of speech involvement and level of transcription agreement. Whereas the previous error-type analyses assessed the contribution of each of the error types to transcription agreement, the present analyses provide a test of summative associations between articulation errors and transcription agreement. The two summative measures are the Percentage of Consonants Correct (PCC) and the Intelligibility Index. Figure 15 is a display of the relevant information for each variable, using the same 48 speech samples and data formatting procedures described previously for error type. Three of the four correlation coefficients are significant at the 0·05 level, with magnitudes again only in the moderate range (approximately 14–18% of accounted variance). The trends and correlation coefficients shown in Figure 15 support the following generalization:

> Transcriber agreement on consonants and vowels has a low to moderately positive association with subjects' severity of involvement, as indexed by percentage of consonants correct and intelligibility.

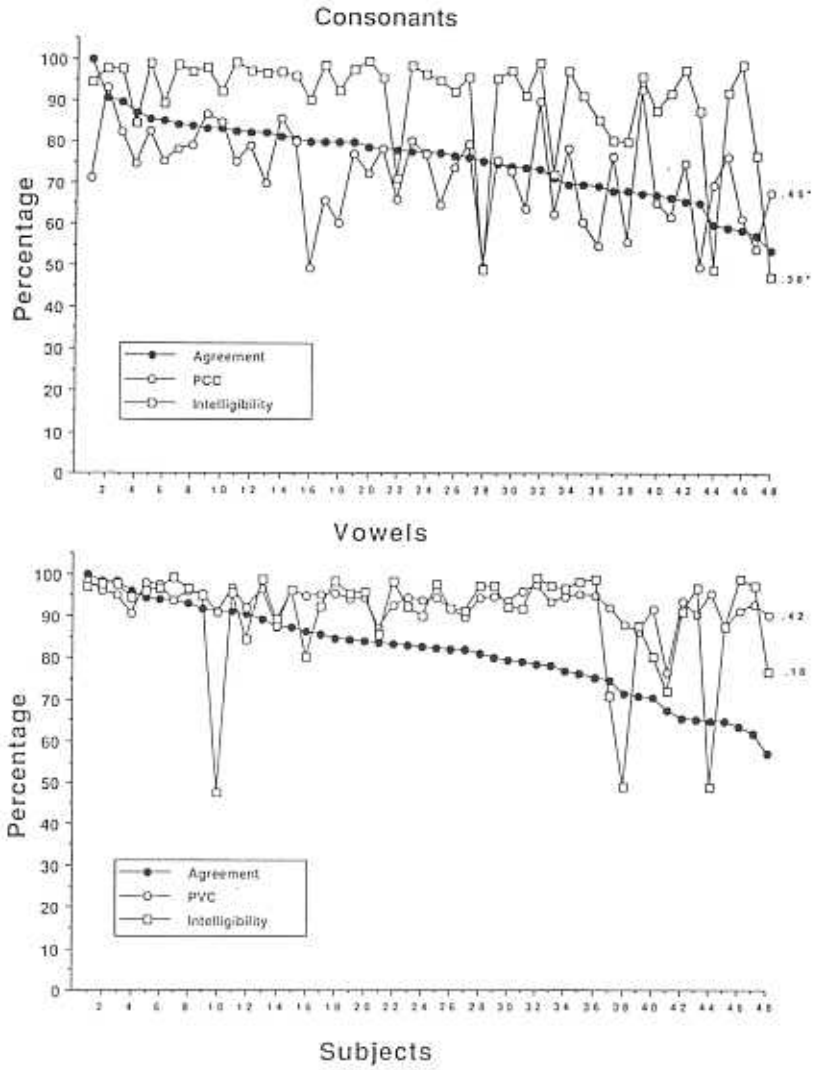


Figure 15. *Transcription agreement in relation to subjects' severity of involvement (Percentage of Consonants Correct; Percentage of Vowels Correct) and intelligibility. The top panel provides data for consonants; the bottom panel for vowels. The values next to two of the trends are Spearman rho coefficients (see text).*

Researchers have observed that, when computed with conventional point-to-point agreement statistics (for alternative agreement statistics see Diedrich and Bangert, 1976; Kearns, 1990; Kearns and Simmons, 1988; Schliesser, Stevens and Bruce, 1973), transcription agreement is lower for transcripts containing many incorrect sounds. Presumably, lowered agreement is functionally associated with the increased complexity of perceptual and cognitive resources needed to process potentially non-acceptable phoneme targets. The present test of this observation does support the generalization, but the strength of the association is again modest. One possible methodological limitation on the magnitudes of the obtained correlations is the relatively restricted ranges of severity in the present study, compared to other

studies based on subjects with normal speech as well as speech errors. Another possible methodological limitation is that severity percentages are based on the entire original speech samples, rather than on the samples excerpted for the transcriber reliability studies. Also, compared to present subjects and transcription systems, studies reporting stronger correlations between transcriber agreement and severity of involvement have used correct/incorrect scoring systems with children having primarily residual distortions (e.g. Diedrich and Bangert, 1980).

Conclusions

The 16 generalizations about the reliability of broad and narrow phonetic transcription are summarized in Table 12, using the same organizational framework followed throughout this report. These conclusions have been derived from graphic descriptions and univariate, nonparametric analyses. Hopefully, they will stimulate controlled studies, using prospective designs that allow multivariate methods.

The primary methodological suggestion from this retrospective study is that, even as carried out by well-trained research personnel, narrow phonetic transcription may be unreliable for certain of the purposes for which it currently is used in communicative disorders. Such restrictions on the limits of perceptual transcription underlie the need for an acoustic-aided technology for segmental and suprasegmental (e.g. Shriberg, Kwiatkowski and Rasmussen, 1990a,b) transcription. The long-term potential for a speech recognition technology able to complete certain clinical and research transcription tasks is also not unrealistic. Current limitations in acoustics instrumentation for general use include the lack of a standardized set of acoustic cues and analysis procedures to identify relevant phonetic parameters, slow processing speeds and limitations in the storage capacity of non-dedicated systems, and the costs of complete signal-processing systems. However, with increasing impetus for technology transfer now evident at national and international levels, considerable progress towards effective, efficient, and generally affordable technologies should be expected during the 1990s.

Until a period when validated acoustic-aided transcription is available for routine clinical use, researchers, university instructors, and clinicians must make a greater effort than currently observed to maximize and document the accuracy of their phonetic transcription. In turn, editors and journal readers must demand sufficient information about the methods and reliability of transcription. Based on the reliability findings from some levels of transcription in the present study, it is clear that certain questions involving phonetic transcription must be approached with extreme caution. Rather than allowing the answer to questions to hinge on the validity and reliability of a few transcribed tokens, multiple sources of evidence should be presented to support each claim.

Acknowledgements

Sincere thanks are extended to each of the following colleagues for their dedicated contributions at different stages of this work: Maria Cavicchio, Barbara Ekelman, Rebecca Hinke, Kit Hoffmann, Joan Kwiatkowski, Jane Loncke, Barbara Pierce, Carmen Rasmussen, Catherine Trost-Steffen, and Carol Widder. Gratitude is also expressed to Annette Ortiz for effective processing of all the data and to Dorothy

Rorick for competent editorial assistance. Finally, we thank *Clinical Linguistics and Phonetics* reviewer George D. Allen for his insightful editorial suggestions. This work was supported by grants from the United States Department of Education, Research in Education of the Handicapped Program, G008400633 and the National Institutes of Health, National Institute of Deafness and Other Communicative Disorders, NS-26246.

Table 12. *Summary of generalizations about sources of variance for further research*

Source	*Variable*	*Generalization*
A. Subjects	1. Intelligibility 2. Severity of involvement	Transcriber agreement on consonants and vowels has a low to moderately positive association with subjects' severity of involvement, as indexed by percentage of consonants correct and intelligibility.
	3. Type of error	Neither the absolute nor relative percentages of each of the primary error types—deletions, substitutions, or distortions—are highly associated with transcription agreement.
	4. Clinical significance	Transcriber agreement is not associated with the clinical significance of diacritical description of speech.
B. Analyses	5. Transcribers	—
	6. Type of agreement	The two types of transcription agreement, inter-judge and intra-judge, have essentially similar average percentages of agreement, ranging from the mid-60s to the mid–high-90s.
	7. Types of systems	The two systems of phonetic transcription, broad (93%) and narrow (74%), differ in average transcription agreement by approximately 20 points.
	8. Agreement criteria	The three types of transcription agreement criteria for diacritics, exact (33%), within-class (40%), and any diacritic (48%), differ in average transcription agreement (uncorrected for chance agreement) by a range of approximately 15 points.
C. Contexts	9. Sampling mode	Transcription agreement based on continuous speech samples is somewhat higher than agreement based on articulation test responses.
	10. Structural, grammatical, and stress forms	—
	11. Word position	Of the three word positions, word-initial consonants are generally transcribed most reliably, with word-final consonants typically associated with the lowest reliability.
	12. Target environment	—

Table 12 *Continued*

Source	*Variable*	*Generalization*
D. Units	13. Class 14. Features	Average transcription agreement at the level of phonetic features and classes is within acceptable levels for broad transcription and generally below acceptable levels for narrow phonetic transcription.
	15. Sounds	The reliability of broad transcription of vowels in a sample is essentially independent of their rank-order of occurrence and percentage correct. For consonants, transcription agreement is independent of rank-order of occurrence and lower, but within an acceptable range, for the 12 most frequently misarticulated sounds. Average transcription agreement percentages for each of 41 sounds are within acceptable levels for broad transcription, but generally below acceptable ranges for narrow phonetic transcription.
	16. Diacritics	There are substantial differences in the average number of diacritics per word used by different consensus transcription teams within and between sampling modes and subject groups. There is fairly stable consistency in the average number of diacritics per word used by the same consensus transcription team doing narrow phonetic transcription on the same speech sample. The proportional occurrence of individual diacritic symbols in narrow phonetic transcription ranges from low to moderately high depending on consensus transcription teams, subject groups, and sampling modes. Transcription agreement on an individual diacritic is essentially independent of its proportional occurrence in a speech sample. The average inter-judge and intra-judge percentage of agreement estimates for diacritic transcript are below acceptable reliability boundary levels, even at the least strict agreement criteria.

References

AMOROSA, H., VON BENDA, U., WAGNER, E. and KECK, A. (1985) Transcribing phonetic detail in the speech of unintelligible children: a comparison of procedures. *British Journal of Disorders of Communication*, **20**, 281–287.

BAILEY, C. J. N. (1978) Suggestions for improving the transcription of English phonetic segments. *Journal of Phonetics*, **6**, 141–149.

BALL, M. J. (1988) *Theoretical Linguistics and Disordered Language*. (San Diego, CA: College Hill Press).

BARKER, J. O. (1942) A numerical measure of articulation. *Journal of Speech and Hearing Disorders*, **7**, 343–356.

BAUER, H. R. and KENT, R. D. (1985) Vocalizations of one-year olds. *Journal of Child Language*, **12**, 491–526.

BUCKINGHAM, H. W. and YULE, G. (1987) Phonemic false evaluation: theoretical and clinical aspects. *Clinical Linguistics and Phonetics*, **1**, 113–125.

BURGI, E. J. and MATTHEWS, J. (1960) Effects of listener sophistication upon global ratings of speech behavior. *Journal of Speech and Hearing Research*, **3**, 348–353.

BURKOWSKY, M. R. (1967) A study of the perception of adjacent fricative consonants. *Phonetica*, **17**, 38–45.

BURKOWSKY, M. R. (1971) A question of perceptual competence. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association, Chicago (November).

BUSH, C. N., EDWARDS, M. L., LACKAU, J. M., STOEL, C. M., MACKEN, M. A. and PETERSON, J. D. (1973) On specifying a system for transcribing consonants in child language: a working paper with examples from American English and Mexican Spanish. Unpublished paper, Child Language Project, Stanford University.

COLE, R. A. (1973) Listening for mispronunciations: a measure of what we hear during speech. *Perception and Psychophysics*, **1**, 153–156.

COLE, R. A., RUDNICKY, A. I., ZUE, V. W. and REDDY, D. R. (1980) Speech as patterns on paper. In R. A. Cole (Ed.) *Perception and Production of Fluent Speech*, pp. 3–50. (Hillsdale, NJ: Lawrence Erlbaum).

COSTLEY, M. S. and BROEN, P. A. (1976) The nature of listener disagreement on judging misarticulations. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association, Houston (November).

CRYSTAL, D. (1985) Things to remember when transcribing speech. *Child Language Teaching and Therapy*, **2**, 235–239.

CURTIS, J. and HARDY, J. (1959) A phonetic study of misarticulation of /r/. *Journal of Speech and Hearing Research*, **2**, 244–257.

DANILOFF, R., WILCOX, K. and STEPHENS, M. (1980) An acoustic-articulatory description of children's defective /s/ productions. *Journal of Communication Disorders*, **13**, 347–363.

DIEDRICH, W. M. and BANGERT, J. (1976) Training speech clinicians in recording and analysis of articulatory behavior. Washington, DC: US Office of Education Grant No. OEG-0–70–1689 and OEG-0–71–1689.

DIEDRICH, W. M. and BANGERT, J. (1980) *Articulation Learning*. (Houston: College Hill Press).

DUCKWORTH, M., ALLEN, G., HARDCASTLE, W. and BALL, M. (1990) Extensions to the International Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics and Phonetics*, **4**, 273–280.

EDWARDS, M. L. and SHRIBERG, L. D. (1983) *Phonology: Applications in Communicative Disorders*. (San Diego: College Hill Press).

ELBERT, M., SHELTON, R. and ARNDT, W. (1967) A task for evaluation of articulation change: I. Development of methodology. *Journal of Speech and Hearing Research*, **10**, 281–289.

FOKES, J., BOND, Z., RITTER, P. and KRACKENFELS, D. (1986) Phonetics instruction and protocol analysis. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association, Detroit (November).

GREEN, B. G., PISONI, D. B. and CARREL, T. D. (1984) Recognition of speech spectrograms. *Journal of the Acoustic Society of America*, **76**, 32–43.

GRUNWELL, P. (1982) *Clinical Phonology*. (Rockville, MD: Aspen Publications).

HENDERSON, F. M. (1938) Accuracy in testing the articulation of speech sounds. *Journal of Educational Research*, **31**, 348–356.

HOFFMANN, P. R. SCHUCKERS, G. H. and RATUSNIK, D. L. (1977) Contextual-coarticulatory inconsistency of /r/ misarticulation. *Journal of Speech and Hearing Research*, **20**, 631–643.

HOFFMAN, P. R. and SCHUCKERS, G. H. (1978) Audio-recording effects upon judgement reliability of children's /r/ misarticulation. *Perceptual and Motor Skills*, **47**, 451–456.

HOFFMAN, P. R., STAGER, S. and DANILOFF, R. G. (1983) Perception and production of misarticulated /r/. *Journal of Speech and Hearing Disorders*, **48**, 210–215.

IRWIN, R. B. (1970) Consistency of judgements of articulatory productions. *Journal of Speech and Hearing Research*, **13**, 548–555.

IRWIN, R. B. and KRAFCHICK, I. P. (1965) An audio-visual test for evaluating the ability to recognize phonetic errors. *Journal of Speech and Hearing Research*, **8**, 281–290.

JOHNSON, C. and BUSH, C. N. (1971) A note on transcribing the speech of young children. *Papers and Reports on Child Language Development*, **3**, 95–100.

JORDAN, E. P. (1960) Articulation test measures and listener ratings of articulation defectiveness. *Journal of Speech and Hearing Research*, **3**, 303–319.

KEARNS, K. P. (1990) Reliability of procedures and measures. In L. B. Olswang, C. K. Thompson, S. F. Warren and N. J. Minghetti (Eds), *Treatment Efficacy Research in Communication Disorders*, pp. 79–90. (Rockville, MD: American Speech–Language–Hearing Foundation).

KEARNS, K. P. and SIMMONS, N. N. (1988) Interobserver reliability and perceptual ratings: more than meets the ear. *Journal of Speech and Hearing Research*, **30**, 131–136.

KLATT, D. H. and STEVENS, K. N. (1973) On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment. *IEEE Transactions on Audio and Electroacoustics*, **AU21**, 3.

KORNFIELD, J. R. (1974) A new twist on an old observation: kids know more than they say. In A. Bruck, R. A. Fox and M. W. LaGaly (Eds), *Papers from the Parssession on Natural Phonology*, pp. 210–219. Chicago: Chicago Linguistic Society.

KRESHECK, J., FISHER, H. and RUTHERFORD, D. (1972) A study of /r/ phones in the speech of three-year-old children. *Folia Phoniatria*, **24**, 301–312.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P. and STUDDERT-KENNEDY, M. (1968) Why are speech spectrograms hard to read? *American Annals of the Deaf*, **113**, 127–133.

MACWHINNEY, B. and MARENGO, K. (1986a) MULTIBET 1.0: A proposal for an ASCII translation and a set of names for extended IPA notation, *Transcript Analysis*, 3 (No. 1, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania), pp. 83–96.

MACWHINNEY, B. and MARENGO, K. (1986b) UNIBET 1.0: A proposal for a single-character translation of IPA for English, *Transcript Analysis*, 3 (No. 1, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania), pp. 97–98.

MARCH, N., WEAVER, C. H., MORRISON, S. and BLACK, J. W. (1958) Observed and predicted estimates of reliability of aspects of a speech articulation rating scale. *Speech Monographs*, **25**(1), 296–304.

MCCAULEY, R. J. and SKENES, L. L. (1987) Contrastive stress, phonetic context, and misarticulation of /r/ in young speakers. *Journal of Speech and Hearing Research*, **30**, 114–121.

MCNUTT, J. C., WICKI, L. and PAULSEN, J. (1985) Sensitivity and variability of judgements of phoneme errors under four modes of audio-visual presentation. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association, Washington, DC (November).

MILISEN, R. (1954) A rationale for articulation disorders. *Journal of Speech and Hearing Disorders*, **4**, 5–17.

MILLAR, J. B. and WAGNER, M. (1983) The automatic analysis of acoustic variance in speech. *Language and Speech*, **26**, 145–158.

MORRISON, J. A. and SHRIBERG, L. D. (in press) Articulation testing versus conversational speech sampling, *Journal of Speech and Hearing Research*.

MOWRER, D. (1978) Effect of lisping on audience evaluation of male speakers. *Journal of Speech and Hearing Disorders*, **43**, 140–148.

NORRIS, M., HARDEN, J. R. and BELL, D. M. (1980) Listener agreement on articulation errors of four- and five-year old children. *Journal of Speech and Hearing Disorders*, **45**, 378–389.

OLLER, D. K. and EILERS, R. E. (1975) Phonetic expectation and transcription validity. *Phonetica*, **31**, 288–304.

OYER, H. J. (1959) Speech error recognition ability. *Journal of Speech and Hearing Disorders*, **24**, 391–394.

PENDERGAST, K., DICKEY, S., SELMAR, J. and SODER, A. (1969) *The Photo Articulation Test*. Danville, IL: Interstate.

PERRIN, E. H. (1954) The rating of defective speech by trained and untrained observers. *Journal of Speech and Hearing Disorders*, **19**, 48–51.

PHILIPS, B. J. W. and BZOCH, K. R. (1969) Reliability of judgements of articulation of cleft palate speakers. *Cleft Palate Journal*, **6**, 24–34.

PRDS PROJECT WORKING PARTY, FINAL REPORT (1983) *The Phonetic Representation of Disordered Speech*. (London: King's Fund).

PYE, C., WILCOX, K. A. and SIREN, K. A. (1988) Refining transcriptions: the significance of transcriber 'errors'. *Journal of Child Language*, **15**, 17–37.

RILEY, K., HOFFMAN, P. R. and DAMICO, S. K. (1986) The effects of conflicting cues on the perception of misarticulations. *Journal of Phonetics*, **13**, 481–487.

ROBERTS, J. E., BURCHINAL, M. and FOOTO, M. M. (1990) Phonological processes decline from 2½ to 8 years. *Journal of Communication Disorders*, **23**, 205–217.

RUSCELLO, D. M., LASS, N. J., POSCH, V. and JONES, C. L. (1980) The verbal transformation effect as studied in judgements of misarticulations. Paper presented at the Annual Convention of the American Speech–Language–Hearing Association, Detroit (November).

SCHISSEL, R. J. and FLOURNOY, J. E. (1978) An investigation of the variability of judgements of experienced and inexperienced listeners in their use of a screening test of articulation. *Journal of Communication Disorders*, **11**, 459–468.

SCHLIESSER, H. F., STEVENS, C. A. and BRUCE, C. E. (1973) Interobserver reliability of direct magnitude estimation of articulatory defectiveness. *Perceptual and Motor Skills*, **36**, 63–66.

SHARF, D. J. (1968) Distinctiveness of 'defective' fricative sounds. *Language and Speech*, **11**(1), 38–45.

SHELTON, R. L., JOHNSON, A. and ARNDT, W. B. (1974) Variability in judgements of articulation when observers listen repeatedly to the same phone. *Perceptual and Motor Skills*, **39**, 327–332.

SHERMAN, D. and MORRISON, S. (1955) Reliability of individual ratings of severity of defective articulation. *Journal of Speech and Hearing Disorders*, **20**, 352–358.

SHRIBERG, L. D. (1972) Articulation judgements: some perceptual considerations. *Journal of Speech and Hearing Research*, **15**, 876–882.

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

SHRIBERG, L. D., HINKE, R. and TROST-STEFFEN, C. (1987) A procedure to select and train persons for narrow phonetic transcriptions by consensus. *Clinical Linguistics and Phonetics*, **1**, 171–190.

SHRIBERG, L. D. and KENT, R. D. (1982) *Clinical Phonetics*. (New York: Macmillan).

SHRIBERG, L. D. and KWIATKOWSKI, J. (1980) *Natural Process Analysis (NPA): A procedure for phonological analysis of continuous speech samples*. (New York: Macmillan).

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

SHRIBERG, L. D., KWIATKOWSKI, J., BEST, S., HENGST, J. and 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. and HOFFMANN, K. (1984) A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research*, **27**, 456–465.

SHRIBERG, L. D., KWIATKOWSKI, J. and RASMUSSEN, C. (1990a) *The Prosody-Voice Screening Profile*. (Tucson, AZ: Communication Skill Builders).

SHRIBERG, L. D., KWIATKOWSKI, J. and RASMUSSEN, C. (1990b) *The Prosody-Voice Screening Profile: Computer Version*. (Tucson, AZ: Communication Skill Builders).

SHRIBERG, L. D. and OLSON, D. (1987) *PEPAGREE: Programs to compute transcriber agreement*. Madison, WI: Waisman Center on Mental Retardation and Human Development.

SIEGEL, G. (1962) Experienced and inexperienced articulation examiners. *Journal of Speech and Hearing Disorders*, **27**, 28–35.

SILVERMAN, E. (1976) Listeners' impression of speakers with lateral lisps. *Journal of Speech and Hearing Disorders*, **41**, 547–552.

SIREN, K. A. and WILCOX, K. A. (1990) The utility of phonetic versus orthographic transcription methods. *Child Language Teaching and Therapy*, **6**, 127–146.

STEPHENS, M. I. and DANILOFF, R. (1977) A methodological study of factors affecting the judgement of misarticulated /s/. *Journal of Communication Disorders*, **10**, 207–220.

STITT, C. L. and HUNTINGTON, D. A. (1963) Reliability of judgements of articulation proficiency. *Journal of Speech and Hearing Research*, **6**, 49–56.

STOCKMAN, E. J., WOODS, D. R. and TISHMAN, A. (1981) Listener agreement on phonetic segments in early infant vocalizations. *Journal of Psycholinguistic Research*, **19**, 593–617.

TROST, J. (1981) Articulatory additions to the classical description of the speech of persons with cleft palate. *Cleft Palate Journal*, **18**, 193–203.

VAN BORSEL, J. (1989) The reliability of phonetic transcriptions: a practical note. *Child Language Teaching and Therapy*, **5**, 327–333.

VAN DEMARK, D. R. (1964) Misarticulations and listener judgements of the speech of individuals with cleft palates. *Cleft Palate Journal*, **1**, 232–245.

VIEREGGE, W. H. and CUCCHIARINI, C. (1989) Agreement procedures in phonetic segmental transcriptions. In M. E. H. Schouten and P. Th. Van Reenen (Eds), *New Methods in Dialectology*, pp. 37–43 (Dordrecht: Foris Publications).

WEISMER, G. (1984a) Acoustic analysis strategies for the refinement of phonological analysis. In M. Elbert, D. Dinnsen and G. Weismer (Eds), *Phonological Theory and the Misarticulating Child* (ASHA Monographs. No. 22, pp. 30–52). (Rockville, MD: American Speech–Language–Hearing Association).

WEISMER, G. (1984b) Acoustic descriptions of dysarthric speech: perceptual correlates and physiological inferences. *Seminars in Speech and Language*, **5**, 293–314.

WEISMER, G., DINNSEN, D. and ELBERT, M. (1981) A study of the voicing distinction associated with omitted, word-final stops. *Journal of Speech and Hearing Disorders*, **46**, 320–327.

WITTING, C. (1962) On the auditory phonetics of connected speech: errors and attitudes in listening. *Word*, **18**, 221–248.

WRIGHT, H. N. (1954) Reliability of evaluations during basic articulation and stimulation testing. *Journal of Speech and Hearing Disorders Monograph*, **No. 4**, pp. 20–27.

ZUE, V. W. and COLE, R. A. (1979) Experiments on spectrogram reading. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 116–119.