J. Child Lang. **15** (1988), 17–37. Printed in Great Britain

Refining transcriptions: the significance of transcriber 'errors'*

CLIFTON PYE, KIM A. WILCOX
AND KATHLEEN A. SIREN
University of Kansas

(*Received* 29 *April* 1987)

ABSTRACT

This work describes the importance of the transcription process in studies of speech and language acquisition. Using data collected from a hearing child of deaf parents, the three authors derived independent transcriptions of the same speech sample and systematically compared their transcripts with each other and with the best estimate of the speaker's actual productions. The resultant transcripts were used to produce two descriptions of this child's phonological system, one based on a liberal estimate and one on a conservative estimate of the potential error in the transcripts. Discussion includes suggestions for deriving percentages of inter-transcriber agreement and the utility of such figures as a metric of transcription difficulty as well as transcriber ability.

INTRODUCTION

Studies of children's language are fundamentally dependent upon the original transcriptions of the language whether done live or from tape-recordings. There is, however, little information available on the transcription errors that are an inherent part of this process. A better appreciation of error rates and types is particularly important when tests of theoretical importance rest on a small number of utterances from a single subject (e.g. Pinker 1984, Ferguson and Farwell 1975). Systematic mistakes at the phonetic level might easily lead to the mistaken attribution of grammatical morpheme or clitic use. Furthermore, errors are themselves an important source of information about subject variability, indistinct articulation, and transcriber biases. Few researchers have the time or resources to investigate the error rates in all of their transcriptions, so studies of

[*] Address for correspondence: Clifton Pye, Department of Linguistics, 420 Blake Hall, University of Kansas, Lawrence, KS 66045, USA.

TABLE 1. *Summary of phonetic transcription methods and reported percent inter-transcriber agreement from selected studies of children's phonology*

Author(s)	Transcription method	Agreement (%)
Elbert & McReynolds (1979)	Two independent transcriptions; disagreements resolved by consensus	Mean = 94 (R = 90–98)
Schwartz, Leonard, Folger & Wilcox (1980)	Number of transcribers not given; disagreements excluded from corpus	At least 95
Smith & Oller (1981)	Two independent transcriptions; disagreements resolved by consensus	Independent = 80 Consensus > 90
Klein (1981)	One transcriber; second individual transcribed random sequences as a check	90–92 for random sequences
Hodson, Chin, Redmond & Simpson (1983)	Two live transcriptions plus two from audiotape; disagreements resolved by consensus	None reported
Smith & Stoel-Gammon (1983)	Two independent transcriptions; disagreements resolved by consensus	None reported
Lynch, Fox & Brookshire (1983)	One transcriber; three others transcribed together and compared with first; disagreements excluded from corpus	None reported
Schiff-Myers & Klein (1985)	Two independent transcriptions; third individual made independent judgements on disagreements	Agreement reached on 94% of words
Stoel-Gammon (1985)	One member of a four member team transcribed independently; then a second team member one of every eight utterances	94
Camarata & Leonard (1986)	Two independent transcriptions	95
Andrews & Fey (1986)	Two independent transcriptions	Mean = 91 (R = 87–94)

transcription errors are needed to provide a necessary caution against unwarranted theoretical claims.

For most of the available studies that report a percentage of transcriber agreement, the number given is typically greater than 85% (see Table 1). Although such figures may allay concerns about the integrity of the resultant transcript, there is little objective foundation for placing confidence in the number. Few studies of children's language, except for some specifically concerned with articulatory performance, provide even cursory information about the level of transcription used in making inter-transcriber comparisons. Most of these studies compare transcriptions at the lexical level, which is already several steps removed from the phonetics. Even fewer studies provide information on the details of the comparison procedure. Hence, agreement rates could vary enormously, depending upon how the comparison was made. Moreover, only two transcriptions are usually compared when assessing accuracy, with items in disagreement often discarded from the sample. The addition of other transcribers should, in general, increase the number of disagreements and thus lower the reported agreement rates while at the same time increasing the overall accuracy of the transcription.

Guidlines for developing composite transcripts are now available (Ingram, in press, Shriberg, Kwiatkowski & Hoffman 1984). These procedures utilize the democratic principle that more common features present in the various transcriptions guarantee accuracy. However, any differences which might exist between individual transcribers are lost in such a process. These unreported transcriber 'errors' could provide additional information about children's speech skills as well as the biases of the individual transcribers.

Error rates also provide information on the reliability of the composite transcription technique. If transcriber differences are significant, there is no reason to believe that a composite transcription is more objective than any one of the individual transcripts. The overall question then, is how to separate the variability introduced through the transcription process from that which is inherent in a child's speech, and further how best to compare these two measures. For reasons of space the present discussion will be limited to the determination of phonetic reliability.

METHOD

Speech sample

All of the data described here are derived from the speech of a single male child who shall be referred to as Becos. This child's history is unique, in that he is the only son of two deaf parents, and has been raised with no other individuals living in the home. As part of a larger project describing Becos' speech and language development, videotape recordings have been collected at regular intervals from the time that Becos was 2:8 to the present. From

that pool of video data, four tapes were selected for the present analysis. Table 2 shows Beco's age, MLU and proportion of unintelligible utterances (total and partial) for each of the four sessions. Prior to session I, Becos' primary exposure to language had consisted of approximately three months of half-day attendance at a community preschool. Near the time of that first recording, Becos began attending preschool five full days per week, and continued doing so through the remainder of the period covered here. As a result, the tape recordings cover a period when Becos' speech and language skills were just beginning to evolve, and so they seemed an appropriate corpus for the assessment of transcriber effects in phonological analysis.

TABLE 2. *Age and transcript data for the four analysed sessions*

Session	Age	MLU	No. of utterances	Unintelligible (%)
I	3:0	1·77	261	23
II	3:2	—	100	0
III	3:3	2·30	273	32
IV	3:5	—	100	0

During each of the four recording sessions, Becos interacted with the same familiar examiner. In Session I and Session III, Becos and the examiner read from a picture book. These two sessions were designed to elicit a variety of lexical items and facilitate language through the repetitive use of selected grammatical forms (Peters 1983, Snow & Goldfield 1983). However, given the high degree of unintelligibility of Becos' speech, it was difficult to identify the lexical or phonological target for many of his utterances in these two sessions. By contrast, Session II and Session IV consisted of two administrations of the *Compton-Hutton Phonological Assessment* (Compton & Hutton 1978). These second two sessions allowed for the transcription of single word productions with known targets.

Transcription

The three authors independently transcribed Becos' first fifty utterances on each of the four videotapes. The first author (CP) had been trained in child language acquisition as well as anthropological linguistics and had done linguistic field work with Native American groups in both Guatemala and Canada. The second author (KW), a speech-language pathologist, had previously conducted research in the area of children's misarticulations, while the third author (KS), a graduate student in speech-language pathology, was experienced in the phonetic transcription of children's speech. It should be noted that although KS had previously worked with each of the

other two authors, her only formal phonetics course had been taken from a third individual. No precautions were taken to reduce the ambient noise level during transcription because we were interested in the types of transcriber differences that can occur in a 'typical' context of child language transcription. Each of the transcribers were free to spend as much or as little time as they needed in order to complete their transcriptions. Further, few restrictions were placed on the phonetic form of the transcription itself. So that for instance, some transcribers chose to use different types and numbers of diacritic markings than others. We used the standard IPA symbols to transcribe Becos' Midwestern dialect of American English.

A composite transcript was then derived for each of the four sessions, by means of a segment-by-segment comparison of the three transcripts (Ingram 1981, in press). Rules for developing the composite transcript included:

(1) When two or more transcribers agreed at the segmental level, that segment was entered into the composite transcription. Furthermore, any diacritic markings which were agreed upon by two or more of the transcribers were also included in the composite.

(2) When all three transcribers disagreed on a particular phonetic unit, a segment was identified which shared the largest number of features with each of the other three segments (e.g. two of the transcriptions [tid], [gid], [did] have a voiced initial consonant, while two have an initial consonant in alveolar position; this results in the selection of [did] for the composite).

(3) When no common or 'middle' segment could be identified, all three segments were included in brackets in the composite (e.g. [tid], [wid], [hid] → [{t, w, h}id]).

In most instances, the process of developing the composite transcript produced a broad phonetic transcription. This is because the frequency with which two or more individual transcribers chose to use the same diacritic marker for the same segment was quite small. Nonetheless, early on it became evident that the presence of diacritic markings in the original transcripts was extremely helpful, and their availability helped to reduce the number of unresolved differences, and hence the number of brackets in the composite.

From the composite transcripts of each session, several different analyses were completed. The first analysis involved a comparison of the original transcriptions and the derived composites. Next, phonemic inventories for each of the four sessions were derived following the method described by Ingram (1981). Finally comparisons between Becos' error patterns and the loci of the transcriber disagreements were undertaken in order to establish the extent to which various transcriber effects were responsible for the patterns observed in Becos' speech.

Throughout the remainder of this work, the terms initial and final

consonants refer to syllable-initial and syllable-final rather than word-initial and word-final positions. This categorization strategy was used in order to expand the analysable corpus without adding unnecessary complexity to the analyses. Methodologically, the CV shape was defined as the basic syllable form so that nearly all of the word-medial consonants in the corpus were described as being syllable-initial. Further, all of the segments present in a given consonant cluster were categorized in the same way. So that in the word [prEzInt] for example, there were three initial ([p],[r],[z]) and two final consonants ([n],[t]).

RESULTS

Frequency of transcriber differences

As hypothesized, the addition of just one transcriber to the phonetic transcription process added considerable variability to the data. Table 3 presents inter-transcriber agreement data for each of the four sessions. All of the percentages presented in this table are segment-by-segment comparisons, with entries corresponding to the number of individual segments which were transcribed identically by two or three transcribers. The number of segments compared in each calculation is given in parentheses. Across the four sessions and three segment groups, all comparisons between two transcribers are in the 53% to 88% range. While for the very same data set, three-judge agreement ranges only from 26% to 79%. Further, these agreement figures are very stable across time, indicating that they were relatively unaffected by elicitation condition or stage of Becos' language development.

Table 4 lists the number of transcriber disagreements by session and word position. The disagreements were evenly distributed across the three transcribers. No individual judge was responsible for a significant proportion of errors, thus confirming the independence of the original transcriptions.

The relatively low percentages of agreement for the two-way comparisons may reflect the amount of unintelligible speech present in the transcripts. Note the difference in agreement levels present between Sessions I and III and Sessions II and IV. The second and fourth sessions consisted of two separate administrations of the *Compton-Hutton Phonological Assessment* and thus contained primarily single word productions where the target was known to the transcribers. Further, many of the Compton-Hutton pictures were unfamiliar to Becos, and/or represented items for which he had no lexical term. Thus, the examiner modelled the majority of Becos' productions throughout these sessions, and in some cases, had to produce the model several times in order to elicit a response. As a result, the relatively high levels of inter-transcriber agreement present in the Compton-Hutton sessions may be due to increased articulatory accuracy on Becos' part, resulting both from simplified single-word contexts and examiner models. They might, however,

TABLE 3. *Two- and three-way inter-transcriber percentages of agreement for the four sessions (Numbers have been rounded to the nearest whole number, with the total number of comparisons given in parenthesis.)*

	Initial consonants	Vowels	Final consonants
	Session I		
n	(108)	(145)	(89)
2 Transcribers			
CP–KW	66	59	63
CP–KS	63	63	53
KW–KS	68	58	55
3 Transcribers	53	47	37
	Session II		
n	(144)	(126)	(110)
2 Transcribers			
CP–KW	80	81	78
CP–KS	83	82	80
KW–KS	81	82	83
3 Transcribers	74	26	72
	Session III		
n	(134)	(135)	(53)
2 Transcribers			
CP–KW	53	68	87
CP–KS	64	64	83
KW–KS	72	72	85
3 Transcribers	41	52	75
	Session IV		
n	(163)	(107)	(104)
2 Transcribers			
CP–KW	84	79	85
CP–KS	83	88	77
KW–KS	83	82	77
3 Transcribers	78	79	69

also be due to the fact that the transcribers were similarly biased by their notion of the speaker's presumed target.

Types of transcriber differences

The differences between transcribers fall into three major categories: additions of segments, omissions of segments, and single-feature alterations within a segment. Table 5 shows the distribution of these differences across the four sessions. The disagreements do not form any discernible pattern except that the largest category of differences were feature changes. There is also a suggestion that the transcribers added segments more often in syllable-final position. This may reflect a bias on the part of the transcribers to complete words that were understood, but not produced intact.

TABLE 4. *Transcriber differences computed from the composite by session and by word position*

Transcriber	CP	KW	KS
Session		Initial consonants	
I	25	18	19
II	13	18	11
III	41	28	21
IV	13	15	16
Total	92	79	67
Session		Vowels	
I	30	35	31
II	13	13	11
III	27	18	24
IV	9	14	5
Total	79	80	71
Session		Final consonants	
I	17	17	27
II	15	10	9
III	4	3	5
IV	8	8	17
Total	44	38	58
Grand total	215	186	197

TABLE 5. *Percentage of occurrence of three types of transcriber differences (The total number of disagreements is shown in parentheses.)*

	Initial consonants	Vowels	Final consonants
Session I	($n = 62$)	($n = 96$)	($n = 61$)
Additions	21	5	49
Omissions	10	10	18
Features	69	84	33
Session II	($n = 42$)	($n = 37$)	($n = 34$)
Additions	12	16	18
Omissions	7	5	12
Features	81	78	70
Session III	($n = 80$)	($n = 69$)	($n = 12$)
Additions	32	9	42
Omissions	14	13	42
Features	54	78	17
Session IV	($n = 44$)	($n = 28$)	($n = 33$)
Additions	11	11	21
Omissions	20	7	21
Features	68	82	58
Total	($n = 228$)	($n = 230$)	($n = 140$)
Additions	21	9	34
Omissions	13	10	19
Features	66	81	46

Additions and omission were not confined to particular segments or transcribers. Only two segments, [t, ə], were added in all four transcripts, while only one segment, [ə], was omitted in all four transcripts. The vowel [ɪ] was added in three transcripts while [t, d, n, s] were omitted in three of the four transcripts. An idea of the significance of these additions and omissions can be gained by comparing them with the frequency rank orders of the segments shown in Table 6. The two vowels [ɪ, ə] occur the most frequently in the transcript, and thus would be expected to have the most additions and omissions. The final consonants [t, n] are also frequent. The frequency of their addition and omission is therefore proportional to their frequency of appearance in the transcript. The initial consonants [t, d, s] suggest a different story. While they occur frequently in the transcripts, they do not occur as frequently as [r, l, b, k, n]. Thus, either some factor (e.g. sonority) reduced the number of our differences with respect to [r, l, b, k, n] or some other factor increased our chance of omitting [t, d, s] in initial position.

An anonymous reviewer for this journal suggested that the phonetic context might have affected our transcription. For example, it is possible that we were more likely to omit consonants in consonant clusters than single consonants in either initial or final position. This could be part of the

TABLE 6. *Frequency rank orders for all segments in composite transcription*

Initial consonants	Rank	Vowels	Rank	Final consonants	Rank
r	1	ɪ	1	k	1
l	2	ə	2	t	2
b	3·5	ʌ	3	s	3
k	3·5	i	5	n	4
n	5	ɛ	5	r	5
t	6·5	æ	5	m	6
d	6·5	a	7	ʧ	7·5
p	8	o	8·5	g	7·5
s	9	ɝ	8·5	d	9·5
f	10	eɪ	10	ŋ	9·5
w	11	e	11	p	11·5
g	12	aɪ	12	f	11·5
h	13	ʊ	13·5	l	13
m	14	aʊ	13·5	ʃ	14
ɾ	15	oʊ	15	θ	15
j	16	ɔ	16·5	b	16
ʃ	17	u	16·5	ʔ	18
z	19			z	18
ʧ	19			β	18
ð	19			v	21
ʍ	21			ʤ	21
θ	22			ð	21
v	23				
ʤ	24				

explanation for the differences over [t, d, s] which were frequently omitted. We re-examined the number of omissions in clusters in initial and final position. Our results are shown in Table 7. This table shows that roughly one-half to two-thirds of our disagreements with omissions occurred in the context of a consonant cluster in both initial and final positions.

Our examination of the contexts of transcriber omissions revealed a complex cluster of effects. Table 7 shows that the occurrence of consonant clusters had a definite effect on transcriber accuracy. The omissions also show a systematic difference between word-initial and word-final positions.

TABLE 7. *Contexts of transcriber omissions*

	Initial consonants	Final consonants
Session I		
Total disagreements	62	61
Total no. of clusters	8	6
Total no. of omissions	4	11
Omissions in clusters	—	4
Omissions of utterance-medial consonants	4	9
Session II		
Total disagreements	45	34
Total no. of clusters	28	13
Total no. of omissions	6	4
Omissions in clusters	5	2
Omissions of utterance-medial consonants	4	2
Session III		
Total disagreements	80	15
Total no. of clusters	7	4
Total no. of omissions	11	8
Omissons in clusters	—	7
Omissions of utterance-medial consonants	5	7
Session IV		
Total disagreements	44	33
Total no. of clusters	44	10
Total no. of omissions	9	8
Omissions in clusters	8	8
Omissions of utterance-medial consonants	1	4
Total		
Total disagreements	231	143
Total no. of clusters	87	33
Total no. of omissions	30	31
Omissions in clusters	13	21
Omissions of utterance-medial consonants	14	22

All six clusters in initial position with /s/ lost the /s/ while four of six clusters with /s/ in the final position retained /s/ and lost a stop consonant. The two final clusters with /ks/ lost the /s/ rather than the /k/. Clusters containing the liquids /l, r/ did not exhibit any particular pattern. In initial position /fr/ was reduced to /r/ once and /f/ once. The cluster /kl/ was reduced to /k/ once and /l/ once in initial position.

Table 7 also shows that the type of language sample had considerable effect on the likelihood of transcriber omissions. For example, the first and third transcripts contain 15 instances of omission by one of the transcribers in initial position. None of these occurred in consonant clusters. The second and fourth transcripts had 13 omissions in consonant clusters out of 15 omissions in all. These results partly reflect the fact that the Compton–Hutton task contains many items with consonant clusters in word-initial position. This increases the likelihood that if there is a disagreement about an initial consonant, it is likely to occur in a cluster. However, another factor was the difference between the single word elicitation in the Compton–Hutton sessions versus the samples of spontaneous speech in the first and third sessions. Transcriber omissions in the spontaneous samples were most likely to occur in the middle of utterances. Here, utterance interpretation seemed to play a major role in consonant omission. If one transcriber interpreted a syllable as an indefinite article while the other two transcribers treated it as the definite article, the result was an instance of consonant omission. There are three cases in the third sample where omission in a consonant cluster resulted from the complete absence of the syllable. We used the number of utterance medial consonant omissions as a crude way of measuring the effect of complex utterance environments on transcription. Some of the factors involved would be: stress, intonation, and word boundary interactions. Becos' typical spontaneous utterance contained one to three initial, unstressed, syllables with short, lax vowels followed by a final, stressed syllable with a long vowel. The number of omissions of utterance-medial consonants in Table 7 shows that this prosodic envelope had a considerable effect on the likelihood of transcriber omissions.

One striking fact in Table 5 is the almost total absence of additions or omissions of diphthongs. Only one diphthong was mistakenly added in the last transcript, [ou]. The explanation probably lies in the diphthongs' low frequencies of occurrence. The fricatives as a class also appear to be underrepresented in the segments that were added or omitted in the transcription. There are instances of [f, s, z, ð, h] being added and [f, s, θ, ð, h] being omitted, but this still leaves [v, z, ʃ, ʧ, ʤ, θ] unexpectedly absent. Part of the reason is that Becos had not acquired voiced fricatives at the time of these sessions (see discussion below). Another part of the explanation may be the low frequencies with which the segments [v, z, ʃ, ʤ, θ] occur. However, this still leaves the absence of [ʧ] unaccounted for.

Turning to the distribution of feature changes, there was also a fairly random distribution of transcription differences. Most of the differences follow the frequency distribution of the segments in the composite transcript. The initial consonants [d, f, θ, ð] show a greater than expected number of feature changes. The segments [d, f] show changes in the features voicing, place and release, while five of seven feature changes to [θ, ð] are in release. The vowels [a, ɔ] also show a greater than expected number of feature changes. [a] had differences in vowel height, place and gliding, while two of three differences in the transcription of [ɔ] turned it into the glide [oʊ]. Finally, the final consonants [θ, ð] had a greater than expected number of feature changes. Seven of thirteen changes were in the feature of release, although [θ] also had three changes in place and one in voicing.

In general, there were few patterns to the differences observed across transcribers. Exceptions to this rule, however, included both notational conventions (e.g. Transcriber 1 chose to transcribe the vowel in 'bake' as [e] while Transcribers 2 and 3 chose [eɪ]) and segment-specific diacritic markings which may have been indicative of either perceptual differences or altered productions (e.g. Transcriber 3 identified a number of segments as [n] while Transcribers 1 and 2 labelled them as [ŋ]).[1]

Subject variability and disagreement rates

Identification of the effects Becos' mispronunciations had on the transcription process was the next analysis goal. As a first step, in this process, it was necessary to derive a presumed target for each production in order to judge the correctness of his attempts. For sessions II and IV, the items on the Compton–Hutton score sheet were taken as the targets. For Sessions I and III target utterances were derived by means of a morpheme-by-morpheme comparison of the three transcribers' glossed versions of the tape. At least two of the three transcribers had to agree on the gloss in order to include the morpheme in the sample. This fact accounts for the relatively small number of segments available for analysis in the first and third sessions below.

Table 8 presents the data on inter-transcriber disagreements arranged by Becos' correct and incorrect productions. In Session I, for example, Becos produced 40 initial consonants correctly. Twelve per cent of these segments had transcriber disagreements. However, there was some transcriber disagreement on each of the five initial consonants that Becos produced

[1] Only 52 of the 598 total transcriber errors appeared to be attributable to simple notational differences. More importantly, these 52 items were concentrated within Becos' vowel productions where 48 of the 186 transcriber vowel differences involved differences in notation.

incorrectly in Session I. From this table, it is apparent that disagreements between the three transcribers are not evenly distributed across the correct and incorrect productions. Instead, the relatively small set of misarticulated segments accounts for nearly half of all of the transcriber disagreements. This finding suggests that transcriber effects may be greatest for misarticulated rather than correctly articulated productions.

TABLE 8. *Percentage of inter-transcriber disagreements on Becos' correct and incorrect productions* (*The total number of segments in each category is shown in parentheses.*)

	Initial consonants (%)	Vowels (%)	Final consonants (%)
Session I			
Correct	12 (40)	27 (45)	14 (29)
Incorrect	100 (5)	75 (8)	0 (1)
Session II			
Correct	16 (112)	30 (105)	22 (65)
Incorrect	42 (52)	60 (5)	44 (41)
Session III			
Correct	14 (37)	31 (36)	19 (26)
Incorrect	33 (3)	67 (3)	0 (1)
Session IV			
Correct	13 (136)	38 (100)	23 (84)
Incorrect	39 (28)	50 (6)	64 (22)
Average			
Correct	14 (325)	32 (286)	20 (204)
Incorrect	44 (88)	63 (22)	49 (65)

Separating subject and transcriber variability

In order to assess the overlap between Becos and the transcribers, the segment-by-segment variability for the two groups was plotted against one another (Figs 1 and 2). Such plots demand many occurrences of each segment in each position to be informative. These requirements were only met in Sessions II and IV in which the Compton–Hutton test was administered. In each graph, the percentage of variability among the transcribers is plotted on the X-axis, while Becos' variability is plotted on the Y-axis. Just as in Table 3, the percentages for the transcribers are calculated per number of segments, and not number of individual opportunities for transcriber error. In each graph, those items clustered in the upper right-hand corner were seldom misarticulated by Becos and met with general consensus among the transcribers. By contrast, those items in the lower left-hand portion of

each figure were frequently misarticulated by Becos and were the source of large numbers of disagreements among the three transcribers.

A review of Figs 1 and 2 reveals a consistent pattern across the four sessions. The large number of items in the upper right-hand corners

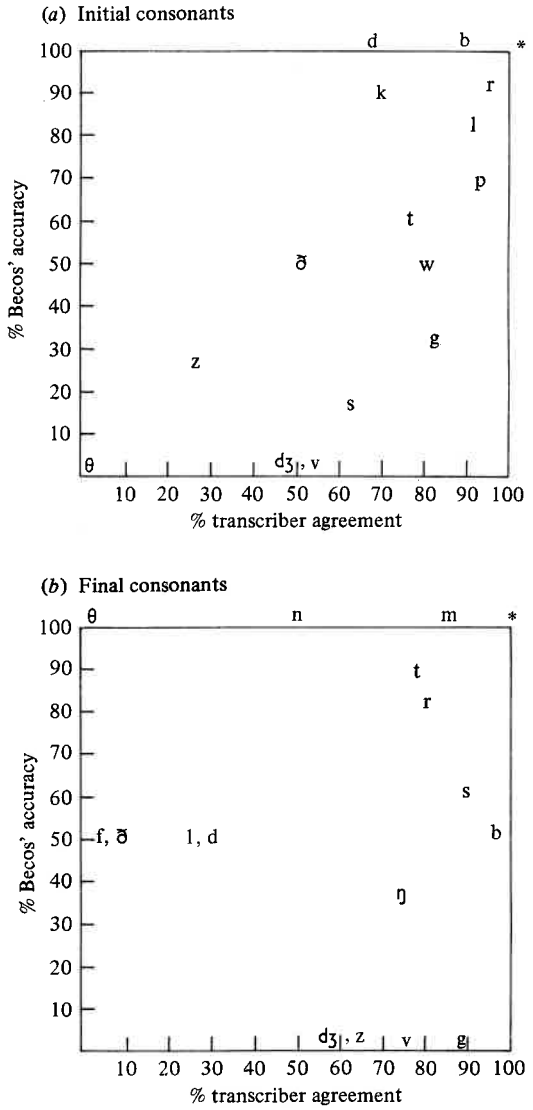

Fig. 1. Becos' % correct production plotted against % of transcriber agreement by (*a*) initial and (*b*) final consonants in Session II. * = m, n, f, ʃ, h, j in (*a*), and p, k, ʃ, ʧ in (*b*).

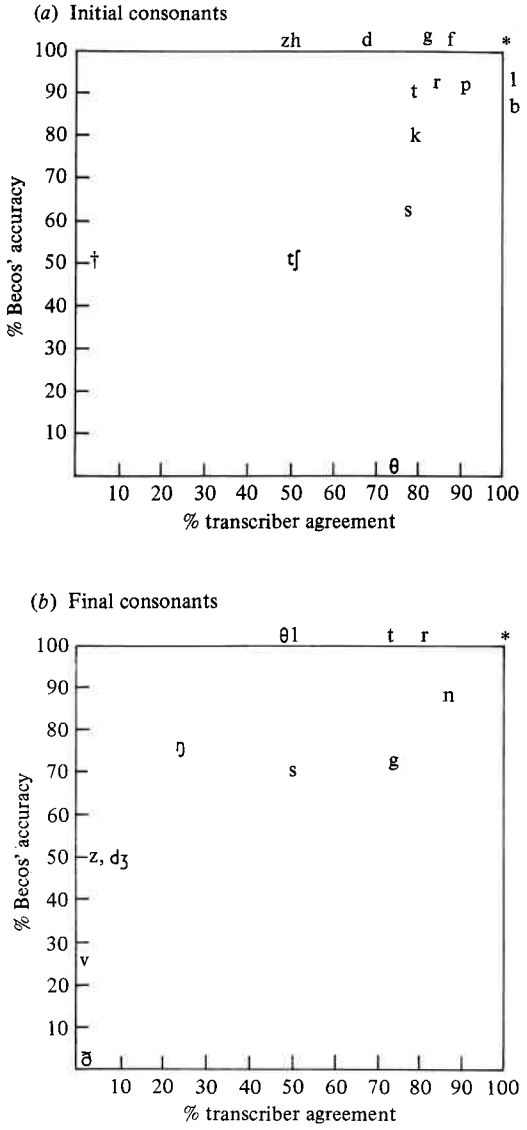

Fig. 2. Becos' % correct production plotted against % of transcriber agreement by (*a*) initial and (*b*) final consonants in Session IV. * = m, n, ʃ, j, w; † = ʤ, v, ð in (*a*); * = *p*, *k*, *b*, *d*, *m*, *f*, ʃ, ʧ in (*b*).

indicates that for most segments, Becos was accurate in his productions and the transcribers agreed upon its phonetic rendering. Far fewer segments are included in the other three quadrants of the graph which include those segments which were incorrectly produced by Becos, were highly variable among the transcribers, or both.

Interpretation of the segments lying along the outside of the graphs is rather straightforward, because those items on the perimeter represent 100% or 0% performance by either Becos or the transcribers. However, interpretation of the various points within the margins of the grid is much more difficult. For example, had the three transcribers done a better job of transcribing Becos' attempts at /ð/ in Fig. 1, the reported accuracy of his productions may have been greatly increased. Instead, there is no way to be certain that errors in the initial transcriptions were not partially responsible for Becos' apparent articulatory difficulties.

Comparison of transcriber performances with phonological analyses

Following Ingram (1981), phonemic inventories were developed from each of the four composite transcriptions. As can be seen in Table 9, Becos' acquisition of final consonants was more severely delayed than was the acquisition of initial ones. Further, the voiced consonants were later developing than the voiceless ones in both word-initial and word-final position. Not depicted in this table, however, is the degree of variability present in Becos' speech. For none of the developing segments in any of the phonologies, was there a stable error pattern. Instead, some items were sometimes correctly used, at other times they were omitted entirely, while at still other times any number of other segments may have been substituted in their place.

Combining the information in Fig. 1 and Fig. 2 with the corresponding portions of Table 9, produces a more conservative estimate of Becos' phonological development at the time of the two *Compton–Hutton* administrations. In Table 10, Becos' phonology is displayed with the presumed transcriber effects removed. In compilation of this table, all segments in the initial phonologies which were at or above the 75% point on both axes of Fig. 1 or Fig. 2 were classified as reliable units for inclusion. Further, those items upon which the three transcribers, but not Becos, met or exceeded the 75% point were classified as acceptable, but slightly less reliable. That second set of segments is marked by asterisks in the table.

Comparisons between Tables 9 and 10 reveal some interesting patterns of change. In general, less confidence should be placed in Becos' fricative development than was assumed initially. This is because 10 of the 17 segments eliminated from the first analysis are fricatives. The other consistent changes between analyses include loss of initial /d/ and final /l/. Given the

TABLE 9. *Becos' phonological development (after Ingram* 1981)

Session																			
	Initial consonants																		
II	b	d		p	t	k*	m	n	f*	(s)	ʃ		h	(ð)		r***	l**	w	j
IV	b	d	g	p	t*	k*	m	n	f	s****	ʃ	ʧ	h	v	ð	z	r	w	j
	Final consonants																		
II	(b)			(p)	t**	k****	m	n		(θ)	s****	(ʃ)	ʃ	(·)		r	(l)		
IV	(b)	d	g*	(p)	t*	k**	m	n	ŋ	f	θ	s*		ʃ	ʧ	z	r	l	

* A sound with one asterisk occurs at twice the criterion frequency, a sound with two asterisks occurs at three times the criterion frequency, and so on.

TABLE 10. *Becos' phonological development following elimination of transcriber effects*

Session														
	Initial consonants													
II	b		p^a	t^a		m	n	f		ʃ	h	r	l	w^a	j
IV	b	g	p	t	k	m	n	f	s^a	ʃ		r	l	w	j
	Final consonants														
II	$(b)^a$			(p)	t	k	m	n		s^a	(ʃ)	ʧ	r		
IV	(b)	d	g	(p)	t	k	m	n	f		ʃ	ʧ	r		

[a] Indicates that the segment met the 75% accuracy criterion for the transcribers only

several precautions taken in deriving this second analysis, it represents a highly conservative estimate of phonological development. Hence the differences observed between Tables 9 and 10 should be an estimate of the maximum differences attributable to transcriber effects.

DISCUSSION

The present work had three major goals. First, assess the frequency and distribution of differences that occur when two or more researchers transcribe a child language sample. Second, analyse the assessment of transcription reliability. Third, address the implications of transcriber variability for the evaluation of children's phonologies.

These results show that the frequency of transcriber differences depends upon a complex interaction of factors, including: (1) the type of speech sample (single words or running conversation); (2) the intelligibility of the child's speech; (3) the frequency of occurrence of individual segments in the child's speech; (4) the number of transcribers; (5) the level of transcript comparison (phonetic, syllabic, morphological, or lexical); and (6) the transcribers' previous experience and training. Contrary to our expectations, the differences between transcribers did not display any consistent patterns.

The frequency of transcriber differences was evenly distributed between the transcribers and across the different segments and phonetic features. Thus, this work shows that the transcription process adds a certain amount of noise to the child's data.

These results have important implications for the process of assessing transcription reliability. Foremost among them is the implication that the consensus method does not assure transcription accuracy. The consensus method can quickly eliminate differences that are due to different transcription practices such as a [e] instead of [eɪ]. However, it relies on a conscious recognition of the factors which may have affected transcriber reliability. Transcribers have to realize when the language sample was unintelligible for them, too fast for them, or when they might have been inattentive in order to agree upon a correct consensus transcription. When the transcribers are not conscious of the factors affecting their transcription, or when they attribute their differences to the wrong factors, the consensus method amounts to an arbitrary selection among different transcriptions.

In addition, previous authors have acknowledged that in group transcribing sessions, one judge can directly influence another judge's perceptions (Shriberg *et al.* 1984). The consensus decision method ignores these social and psychological pressures to select the transcription of the most senior or 'experienced' member of the team. Thus, it is imperative that each judge complete all initial transcription independently.

Investigators often assume that failure to achieve a minimum 85 % level of inter-judge reliability is evidence that one or both transcribers is deficient in their transcription skills; or that, at the very least, the two transcribers are not utilizing the same notational conventions in their transcription. This simple assumption ignores the reality of the transcription process. Firstly, it is much easier to attain a given criterion level when the comparison is made on a lexical rather than a segmental basis simply because more information is available for identifying an intended word than an intended sound. Second, it is much easier to transcribe single words with an adult model than running conversation with few contextual clues. Third, it is easier to transcribe the speech of a child with a fully developed phonological inventory than it is to transcribe the speech of a child who is in the process of acquiring a class of sounds. Finally, the agreement is guaranteed to be higher when two transcriptions rather than three are compared.

In the present study, the backgrounds and previous experience of the transcribers also served as a source of transcriber disagreement. Two of the judges had previous transcription experience as part of their clinical work in speech/language pathology. Hence, their focus reflected that of the profession in that differences between correct and incorrect forms of a segment often serve as the basis for transcription. Thus, positional diacritics (e.g. [n]) are often preferred to non-English phonetic symbols (e.g. [ɲ]) as a

means of organizing future phonetic shaping of a child's speech in therapy. The third judge, on the other hand, was experienced in writing previously undescribed languages. In this work, comparisons to correct English forms are not appropriate. Instead, relationships between surface phonetic forms and underlying phonemic contrasts are of the greatest importance, and hence utilization of non-overlapping phonetic symbols is desirable.

All this suggests that reported percentages of inter-transcriber reliability should not serve merely as an evaluation of transcriber competency. Instead, these numbers are a valuable index of the form and complexity of the child's speech, the homogeneity of the investigators' backgrounds, and the difficulty of the transcription task. Further, in order to make this number maximally useful, a detailed description of the means of developing the reported agreement level is also needed. This description should minimally include any steps taken to reconcile original transcripts.

One clear finding from this work is the benefit of a third transcription. The importance of including more than two judges in the transcription process is borne out by the fact that of the 1,418 total segments available in the four tapes, some consensus transcription, derived by either segment or feature combination, was achieved for all but 9 segments, or approximately 99·5% of the total. Thus, even when there was disagreement in absolute segment identification among transcribers, there was generally always agreement at some level of analysis between at least two of the three judges.

For differences of addition and omission of segments, the presence of three rather than two transcribers allowed for a straightforward decision to include a segment. Such decisions are much more difficult with only two transcribers. In cases of single feature changes, the third transcriber allowed for a reliable mapping of many of the subtle behaviours in Becos' speech. For example, many of these single feature differences involved presence or absence of voicing. Many of the apparent voicing discrepancies across transcripts involve items which one or several of the transcribers marked as intermediate in voicing characteristics (e.g. CP = [f], KW = [f̬], KS = [f̬]. With three narrow transcriptions to work from, reliable transcription of intermediate voicing characteristics (e.g. [f̬]) was possible. In most other consensus transcription formats, such information is often lost.

The results reported above have important implications for the assessment of children's phonologies. Recall that the three transcribers differed on those sounds that Becos was in the process of acquiring. While the variation in these novel sounds is real, it now seems as though the transcribers also contributed to the variability in the transcription of these sounds. Ferguson and Farwell (1975) introduced a method for describing children's sounds which uses every variation in production as an indicator of a broader phone class. Thus, if a child sometimes produces a word with an initial [b] and other times with an initial [v], this is used as evidence that these sounds belong to

a single class of sounds /b–v/. However, according to these results, transcriptions of such variable productions are likely to differ across transcribers as well as across utterances. The effect would be that the phone classes would include some sounds that were the result of variable transcription rather than variable production. One way of partially eliminating such errors from phonological assessments would be to establish a criterion frequency for every sound included in an analysis (as in Ingram 1981). If transcription errors are scattered in a more random fashion than the variable productions the frequency criterion should discriminate between them. The frequency criterion, however, will not eliminate errors that are the result of a bias on the part of the transcriber to write a sound consistently as another sound.

REFERENCES

Andrews, N. & Fey, M. (1986). Analysis of the speech of phonologically impaired children in two sampling conditions. *Language, Speech & Hearing Services in Schools* **17**. 187–98.

Camarata, S. & Leonard, L. (1986). Young children pronounce object words more accurately than action words. *Journal of Child Language* **13**. 51–65.

Compton, A. J. & Hutton, J. S. (1978). *Compton–Hutton Phonological Assessment*. San Francisco CA: Carousel House.

Elbert, M. & McReynolds, L. (1979). Aspects of phonological acquisition during articulation training. *Journal of Speech and Hearing Disorders* **44**. 459–71.

Ferguson, C. A. & Farwell, C. (1975). Words and sounds in early language acquisition: English initial consonants in the first fifty words. *Language* **51**. 419–39.

Hodson, B., Chin, L., Redmond, B. & Simpson, R. (1983). Phonological evaluation and remediation of speech deviations of a child with a repaired cleft palate: a case study. *Journal of Speech and Hearing Disorders* **48**. 93–8.

Ingram, D. (1981). *Procedures for phonological analysis of children's language*. Baltimore MD: University Park Press.

Ingram, D. (in press). *First language acquisition: method and explanation*. Cambridge: C.U.P.

Klein, H. (1981). Productive strategies for the pronunciation of early polysyllabic lexical items. *Journal of Speech and Hearing Disorders* **24**. 389–405.

Lynch, J., Fox, D. & Brookshire, B. (1983). Phonological proficiency of two cleft palate toddlers with school-age follow-up. *Journal of Speech and Hearing Disorders* **48**. 274–85.

Peters, A. (1983). *Units of language acquisition*. Cambridge: C.U.P.

Pinker, S. (1984). *Language development and language learnability*. Cambridge MA: Harvard University Press.

Schiff-Myers, N. & Klein, H. (1985). Some phonological characteristics of the speech of normal-hearing children of deaf parents. *Journal of Speech and Hearing Research* **28**. 466–74.

Schwartz, R., Leonard, L., Folger, M. & Wilcox, M. (1980). Early phonological behavior in normal-speaking and language disordered children: evidence for a synergistic view of linguistic disorders. *Journal of Speech and Hearing Disorders* **45**. 357–77.

Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research* **27**. 456–65.

Smith, B. & Oller, D. (1981). A comparative study of pre-meaningful vocalizations produced by normally developing and Down's Syndrome infants. *Journal of Speech and Hearing Disorders* **46**. 46–51.

Smith, B. & Stoel-Gammon, C. (1983). A longitudinal study of the development of stop

consonant production in normal and Down's Syndrome children. *Journal of Speech and Hearing Disorders* **48**. 114–18.

Snow, C. E. & Goldfield, B. A. (1983). Turn the page please: situation-specific language acquisition. *Journal of Child Language* **10**. 551–69.

Stoel-Gammon, C. (1985). Phonetic inventories, 15–24 months: a longitudinal study. *Journal of Speech and Hearing Research* **28**. 505–12.