A Weighted Reliability Measure for Phonetic Transcription

D. Kimbrough Oller
Heather L. Ramsdell
The University of Memphis, Memphis, TN

Purpose: The purpose of the present work is to describe and illustrate the utility of a new tool for assessment of transcription agreement. Traditional measures have not characterized overall transcription agreement with sufficient resolution, specifically because they have often treated all phonetic differences between segments in transcriptions as equivalent, thus constituting an *unweighted* approach to agreement assessment. The measure the authors have developed calculates a *weighted* transcription agreement value based on principles derived from widely accepted tenets of phonological theory.
Method: To investigate the utility of the new measure, 8 coders transcribed samples of speech and infant vocalizations. Comparing the transcriptions through a computer-based implementation of the new weighted and the traditional unweighted measures, they investigated the scaling properties of both.
Results: The results illustrate better scaling with the weighted measure, in particular because the weighted measure is not subject to the floor effects that occur with the traditional measure when applied to samples that are difficult to transcribe. Furthermore, the new weighted measure shows orderly relations in degree of agreement across coded samples of early canonical-stage babbling, early meaningful speech in English, and 3 adult languages.
Conclusions: The authors conclude that the weighted measure may provide improved foundations for research on phonetic transcription and for monitoring of transcription reliability.

KEY WORDS: phonetic transcription, weighted agreement measure, phonological coding, transcription agreement, transcription reliability

Reliability of Observation in Phonetic Transcription

Phonetic transcription through the International Phonetic Alphabet (IPA) is widely utilized to study vocalizations in a variety of realms including linguistic fieldwork, child phonology, dialectology, evaluations of disordered speech, and infant vocalizations. The descriptions produced by such transcription play central roles in both scientific and clinical work, because they result in characterization of phonemic inventories and allophonic variations within languages, as well as in specification of the presumed nature of speech errors in disorders of communication and child phonology. Tracking of improvement in therapy is similarly often dependent on the accuracy of phonetic transcription and on the implications that transcription entails regarding relations between disordered speech and correctly articulated speech. The accuracy of

Disclosure
The first author is one of the primary developers of the LIPP phonetic transcription and phonological analysis software used in this research and was previously a stockholder in Intelligent Hearing Systems, which produces the software. Although he does not receive royalties or other payment from sales of the software, he does consulting work for a fee with individuals attempting to apply the software in their work.

1092-4388/06/4906-1391

transcription of poorly intelligible speech has fundamental consequences both in determining appropriate goals for intervention and in monitoring change. In the study of infant vocalizations, phonetic transcription is often aimed at determining an inventory of adultlike phonetic elements or syllables commanded by infants, and thus often provides the primary methodology to specify infant vocal capabilities and to predict likely emergent disorders of phonological capability.

However, the reliability[1] of phonetic transcription is often subject to serious question, especially when the speech or speechlike materials to be transcribed are quite distinct from those found in mature well-formed speech. Major discrepancies from well-formed speech are found in samples derived from infants (Koopmans-van Beinum & van der Stelt, 1986; Oller, 1995; Stoel-Gammon, 1992; van der Stelt, 1993), young children (Davis & MacNeilage, 1995; Stoel-Gammon & Cooper, 1984; Stoel-Gammon & Herrington, 1990; Vihman, 1986), and speakers with severe disorders, whether as a result of hearing impairment (Bakkum & Plomp, 1995; Hudgins & Numbers, 1942; Oller & Eilers, 1981; Davis, Morrison, von Hapsburg, & Warner Czyz, 2005; Stoel-Gammon & Otomo, 1986) or production disorders associated with motoric or phonological difficulties (Kent, Weismer, Kent, & Rosenbek, 1989; Pollock & Hall, 1991; Shriberg, Aram, & Kwiatkowski, 1997). In such cases, vocalizations may fail to meet requirements of well-formedness that are assumed by transcription in IPA (Oller, 1992). Even syllabification may be difficult to discern in such cases, and consequently transcription can yield serious disagreements among coders (Oller & Steffens, 1994; Shriberg & Lof, 1991).

In this article, the terms *well formed* and *canonical* are used interchangeably, and in both cases the terms are primarily used to refer to characteristics of the minimal rhythmic unit of vocalization, the syllable. The term *canonical syllable* has been defined in prior work of the first author to refer to any syllable that meets the primary requirements of well-formedness for the most common types of syllables in natural spoken languages. Such syllables must possess at least one vowel-like element to form a syllable nucleus, at least one supraglottally articulated consonant-like element to form a syllable margin, and a rapid uninterrupted formant transition between the nucleus and the margin. A discussion of the formal requirements of canonicity, along with proposed acoustic and articulatory specifications, as well as cross-linguistic observations about occurrence of syllable types in languages, is provided in Oller (2000).

The cited work discusses preliminary evidence that complex utterances consisting of noncanonical syllables (especially syllables violating the rapid transition requirement) tend to be difficult to transcribe in IPA, a topic that has not yet been systematically investigated. Our own experience has been much influenced by the fact that complex utterances consisting of noncanonical syllables occur often in very early human development prior to the middle of the 1st year of life. Usually the canonical stage of vocal development begins at about the middle of the 1st year, when canonical syllable production (also called "canonical babbling") by infants becomes prominent, often in well-formed sequences of reduplicated syllables such as [baba] or [dada] (Koopmans-van Beinum & van der Stelt, 1986; Oller, 1980; Stark, 1980). Prior stages of vocal development are termed *precanonical,* and often include many complex infant utterances composed of noncanonical syllables that have been reported anecdotally and in the preliminary evidence discussed in Oller (2000) as difficult to transcribe.

Because samples of vocalization may vary dramatically in degree of well-formedness, it should perhaps not be surprising that levels of reported reliability are remarkably variable in phonetic transcription studies. Shriberg and Lof (1991) conducted both the most extensive review and the most extensive empirical study to our knowledge on transcription reliability across samples from typical speakers and individuals with disorders of speech. They found transcription reliability ranging from under 20% to nearly 100%, depending on the circumstances of sampling and reliability assessment (Shriberg & Lof, 1991).

With such a range of reliability values, it would seem that expectations about level of reliability need to be adjusted for the participant populations, circumstances, and scientific or clinical questions involved in each particular study. Furthermore, the reliability assessment procedure that is appropriate often varies based on the investigative questions at hand. One important distinction is between investigations that focus on particular characteristics or features of vocalizations (e.g., voicing and nasality in consonants) and investigations that focus on the entire range of vocal characteristics that may influence intelligibility or well-formedness. When only a few features are studied, reliability assessment is appropriately focused on only those features, and each transcribed event can often reasonably be judged as reliable based on a binary judgment—whether the transcribers agree on voicing and nasality, for example. Thus, in such cases, a *featurally restricted* measure of reliability is entirely appropriate. However, when a study focuses on the whole range of speech features (e.g., in studies of intelligibility or well-formedness of vocalization), reliability assessment based on binary judgments is much less informative.

[1] Throughout the article, we use the terms *transcription reliability* and *transcription agreement* interchangeably, although the latter is technically a subtype of the former. With either usage, we intend the more restrictive latter meaning. See Cucchiarini (1996) for the definitions.

In such cases, researchers and clinicians are not so much interested in whether two transcriptions agree, but in the degree to which they agree across the range of all vocalization features combined for assessment in an overall measure of reliability. Featurally restricted reliability assessment can be thought of as a special case of the more general methodology of overall assessment. The present work is intended to contribute to improving overall assessment procedure, and to shed light on the appropriate process of phonetic transcription reliability assessment in general.

Whether one is concerned with overall or featurally restricted reliability assessment, there would appear to exist no general consensus regarding acceptable reliability levels in transcription. Louko and Edwards (2001) suggested that anything greater than 75% is acceptable reliability in phonology. Stoel-Gammon's (2001) review concluded that acceptable reliability measures for transcription of children's speech were between 60% and 80%. Irwin and colleagues (Irwin & Chen, 1941; Irwin & Curry, 1941) asserted that 85% reliability is needed among coders when transcribing vowel-like elements in the cry sounds of babies. The lack of consensus on acceptable levels of reliability exists in part because, as noted previously, there are large differences among samples of vocalizations in their degree of well-formedness and, consequently, large differences in what can be expected in the way of reliability across different sample types. There is much more to be said about acceptable levels of transcription reliability; this issue will be revisited in the Discussion section. There is, however, a logically prior concern, which is how best to assess reliability in the first place. Only if the chosen measure is well designed for its purpose can the process of determining appropriate levels of reliability be sensibly pursued.

Weaknesses of Overall Reliability Assessment Under the Traditional Approach

Variability both in existing recommended standards for transcriber agreement and in outcomes of agreement evaluation suggests that much remains to be understood about how phonetic transcription reliability should be assessed. This investigation is designed to address weaknesses in certain traditional assessment procedures for overall agreement in phonetic transcription. Prior measures of overall transcription agreement have tended to be incomplete in both protocol for alignment of coded segments between transcribers and in weighting of phonetic relations.

The issue of alignment was emphasized by Cucchiarini (1996), who noted that no clear approach to alignment of segments from one coder to another has been provided in past research. This problem is usually trivial in cases in which the speech to be transcribed is fully canonical and intelligible. In such cases, transcribers tend to include segments in the same transcription slots (or "segment columns"), and there is no difficulty in determining which segment from one transcriber should be aligned with a particular segment from the other transcriber. However, with transcriptions of infant sounds or very disordered speech samples, in which well-formedness may be rare or absent, many elements prove to be phonetically ambiguous, and it may be impossible to determine which segments correspond to one another across transcriptions. Additionally, these transcriptions may differ in how many segments are specified for an utterance. As a result, it is often unclear which segments in the different transcriptions should be aligned for the purposes of agreement analysis. These are typical problems in overall reliability assessment for vocalization samples. Cucchiarni made a case for development of a theoretically justifiable procedure for alignment in phonetic transcription reliability assessment.

The most widely used measure of transcription reliability in prior research has been percentage agreement, an approach that has not been well adapted for assessing overall transcription agreement across samples of vocalization, owing to the fact that the procedure has generally included a requirement of absolute match between transcribed segments (see, e.g., Stockman, Woods, & Tishman, 1981). Under the requirement of absolute match, all transcription disagreements are equally weighted. If two transcriptions of the same utterance differ in the number of segments coded, the failure of one transcriber to include a segment coded by the other transcriber produces a reduction in agreement score by the same proportion as a single feature difference on a segment that both transcribers include in their codes. Furthermore, under the absolute match criterion, the natural hierarchy of phonetics is disregarded and all segmental discrepancies are treated as if they were equivalent, a procedure that prevents characterization of the range of differences that can occur between transcribed segments. For example, a [p] and a [b] differ by one feature (voicing), whereas a [p] and a [z] differ by three (voicing, manner, and place), yet with the absolute match approach, disagreements involving these alternations are counted as equivalent. The absolute match criterion, with its lack of weighting for different disagreement types, thus sets up a fundamental scaling problem in which resolution in differentiating degrees of reliability is unreasonably low. It can be predicted that data on agreement produced this way will be skewed by floor effects if the transcription task is difficult, and that the highest scores in reliability that are attainable will tend to be too low even in optimal circumstances of sampling, whenever fine transcription is used.

Another way to look at limitations of the traditional percentage agreement approach is that it treats phonetic units as if they belonged to *flat lists,* where all segments are of the same value. In fact, of course, virtually all theoretical phonetic and phonological approaches treat phonetic units as hierarchical systems rather than as flat lists (Chomsky & Halle, 1968; Goldsmith, 1995; Jakobson, Fant, & Halle, 1952; Ladefoged, 2001; Trubetzkoy, 1939). There is a long history of recognition (and a great deal of recent emphasis on the fact) that the global syllabic structure of utterances in phonological theory involves a higher level of organization than the individual segments embedded within the syllabic structure (Clements & Keyser, 1983; Firth, 1957; Goldsmith, 1995; McCarthy, 1985). A distinction, then, needs to be drawn in agreement assessment between a global syllabic structure level, specifying consonant and vowel slots, and a segment level, corresponding to the particular consonants and vowels that fill the slots. Under this hierarchical approach, global structural disagreements, which can be thought of as deletions or adjunctions in one transcription with respect to another, should result in larger reductions in reliability scores than disagreements involving only some feature of a particular segment included in both transcriptions. This approach is also consistent with empirical information indicating that deletions and adjunctions are more damaging to intelligibility in disordered speech than are segment substitutions (see Hudgins & Numbers, 1942, and Steffens, 1994).

At the segmental level, differences among types of disagreements in phonetic features should also be taken into account. A well-designed agreement assessment would distinguish among featural disagreements that involve differing numbers of features, and would assign higher weightings to disagreements involving more featural differences, all other factors being equal. Furthermore, in some cases it may be desirable to draw distinctions in weightings among disagreements involving marked as opposed to unmarked segments (Prince & Smolensky, 1993; Trubetzkoy, 1939), adjusting the measure appropriately for transcription disagreements that correspond to substitutions known to occur earlier or later in development or for disagreements corresponding to elements known to occur with greater or lesser likelihood in languages.

The need for a weighted approach to assessment of phonetic transcription has also been highlighted by Cucchiarini (1996). Along with Cucchiarini, we argue for development of a structured alternative to the flat list approach to assessing reliability, an alternative involving a *weighted proportion agreement,* taking well-documented hierarchies of phonetic features into account.

The need for a weighted approach has been both implicitly and explicitly recognized by a number of researchers who have in fact introduced elements of weighting into reliability assessment in child phonology, infant vocalizations, and speech disorders research (Davis & MacNeilage, 1995; Ingram, 2002; Oller & Steffens, 1994; Vihman, Macken, Miller, Simmons, & Miller, 1985). Furthermore, no one, as far as we know, has objected to the use of a weighted approach when an overall measure of reliability is desired. The problem is that a well-specified approach to weighting, with explicit theoretical motivation and implementation guidelines, is needed. The goal of the present work is to specify as thoroughly as possible an appropriate weighting approach and alignment procedure.

A Weighted Measure Based on Widely Accepted Tenets of Phonological Theory

To stand the test of time, the agreement assessment measure we envision must be based on phonological theory and should ultimately be supported directly by empirical information regarding phonetic perception and how individual acoustic events can be interpreted as consisting of differing phonetic elements either by different listeners (intertranscriber disagreement) or by the same listener on different occasions (intratranscriber disagreement). Considerable prior research in auditory perception provides data on effects of phonetic expectation and various contextual factors on variability in phonetic perception (Oller & Eilers, 1975; Stoel-Gammon, 2001; Warren & Warren, 1966), and other research provides "confusion matrices" indicating the probability with which a variety of phonetic elements tend to be perceived in response to given phonetic stimuli (the classical work on this topic is Miller & Nicely, 1955). Such research provides empirical support for the idea that commonly recognized principles of phonological relations among elements based on phonological features provide a useful predictive structure regarding likely perceptual disagreements among listeners who are presented with a particular acoustic signal.

Our preference is to develop a reliability measure based on the most general principles of relations among phonological elements, the principles that are widely accepted across the many theoretical approaches to phonology (see review in Gussenhoven & Jacobs, 1998). The assumption underlying our choice is that a measure thus formulated will likely conform to empirical data on perception and provide a useful scale of degrees of disagreement among listeners. As our reasoning goes, reductions in agreement scores, when disagreements occur, should be weighted in accord with these principles of phonological relations. An ideal computer-based assessment needs to be sufficiently flexible to allow for easy user-based adjustments in calculation procedure for agreement scores, so new results on agreement can be readily compared at an appropriate level of detail with prior data analyzed with different weightings, or with prior data that may have been the product of other procedural differences. Also, the ideal tool for agreement assessment will facilitate comparisons across ages and language backgrounds, as well as conditions of possible disorder.

In the present study, we describe and illustrate the utility of a new weighted measure that is consistent with basic phonological theory in analyzing agreement for transcription of an infant vocalization sample. Agreement values for the infant sample obtained under the new measure can be referenced to levels documented in prior research by others and to levels based on transcription of samples from more mature speakers also studied in the present effort. If the measure works well, we expect that transcription of infant vocalizations will produce relatively low agreement values, whereas transcription of mature speech will result in much higher values. While the empirical data evaluated here are derived from individuals without disorders of communication, it is our assumption that the basic principles of weighting should be appropriately applicable across a wide range of samples that are difficult to transcribe.

In this work, the new analysis measure for transcription agreement is implemented in LIPP (Logical International Phonetic Programs; Oller & Delgado, 1999), which provides a transcription and analysis environment facilitating alignment and weighting of disagreements across coders. Theoretical justifications for the procedures and weightings are outlined. To assess the scaling properties of the measure empirically, we consider the following questions and hypotheses:

> *Question 1:* How does the new weighted transcription agreement measure compare in terms of scaling properties with the traditional unweighted measure of percentage agreement for transcription on a broad range of infant vocalizations, both canonical and precanonical? Hypothesis A: The weighted measure will show relatively even distribution of agreement scores across utterances within the infant sample, while the unweighted measure will tend to suffer from floor effects.
>
> *Question 2:* How do levels of agreement obtained for transcription of these infant vocalizations compare with levels of agreement obtained for transcription of citation form utterances from other, more mature, speakers? Hypothesis B: The weighted measure will show consistent agreement scores (across all pairings of transcribers with a standard transcriber) with the following order from lowest to highest agreement values: infant sample < toddler sample < adult samples from languages other than English < adult sample from English.

Method

Transcription Samples

The samples were transcribed from four speakers: (a) an infant at the beginning of the canonical stage of vocalization, (b) a typically developing 2-year-old English learner, (c) a native Korean-speaking adult, (d) a native Ukrainian-speaking adult, and (e) a native American-English-speaking adult. Each sample was recorded on high-fidelity audio equipment, digitized, and imported into LIPP, which allows for transcription of IPA via the traditional keyboard, along with user-designed analysis based on featural characterizations of segments.

The 30 utterances for the infant sample were drawn from a single infant in a longitudinal study. The recordings used occurred prior to and shortly after the onset of canonical babbling in the infant (in this case, 8 and 9 months). The sample was chosen to include both canonical and precanonical vocalizations so as to represent a wide array of sounds, from squeals and vowel-like elements to marginal syllables (syllables with protracted or otherwise ill-formed formant transitions) and canonical babbling (for a description of these types, see Oller, 2000). The canonical babbling ratio (the ratio of canonical syllables to all syllables) for this sample based on the first author's coding was .42. The sample was thus highly canonical, but still included many utterances that would be judged difficult to transcribe reliably by virtually any standard. The infant was quite vocal on the days in which the sample was obtained, with considerable interaction between the parent and infant. Vocalizations were clipped out of the samples so that only the infant's voice could be heard by transcribers. The acoustic quality of the digitized utterances was typical of infant vocalization samples acquired for research, showing moderately good signal quality through 10 kHz.

Twenty words each were recorded in citation form from each of the additional four speakers. It is important to keep in mind that citation form utterances may contrast with the style of vocalization found in the infant case, in which the sample was obtained naturalistically in interaction, and in which individual utterances, spontaneously produced, were extracted from the sample. The citation form utterances may have been articulated more carefully than the infant utterances, because in each case there was a lexical target form specified in the task for the toddler and adults. For the infant utterances, no such target form was provided, of course, because the infant could not have been expected at this stage of development to have imitated the forms. The English words upon which the citation form samples were based (see the Appendix) were chosen to represent a typical array of word lengths and syllable types in

English. The Ukrainian and Korean samples were based on translations of the English words.

The 2-year-old participated in a game with the second author, in which the English words were elicited via repetition, and as with the infant, the digitized recording was later clipped out so that transcribers could hear the toddler's voice only. The three female adults were each asked to read randomly presented written versions of the word lists. The Korean and Ukrainian speakers were given the list of English words prior to the recording and asked to translate the words into their respective native languages. Thereafter, each speaker recorded the words in her native language. All of these samples were digitally recorded via TF32 (Time-Frequency Analysis, 32 Bits for Windows (http://userpages.chorus.net/cspeech, by Paul Milenkovic), and separate files were clipped out to represent just the words for the transcription study. The wave files were imported into LIPP where the items could be played during transcription.

For this study, the samples of utterances were small because of the labor-intensive nature of the effort, which included individual segment-by-segment alignment and checking for consistency with the transcription protocol. In the future, further automation of these approaches will make much larger samples possible.

Coders and Coding Protocol

Eight coders participated in the present study, all from the School of Audiology and Speech-Language Pathology at The University of Memphis. The six graduate student coders, including the second author, had previously been trained intensively in the first author's phonetic transcription class, while the other two transcribers were doctoral faculty members, including the first author, with many years of experience as phoneticians. All coders transcribed the infant sample. Six of the eight transcribed the additional samples from the 2-year-old and the adults. All the transcribers worked at coding stations in the infant vocalizations laboratory, and had the option of adjusting gain levels for each utterance while they coded. The coders were encouraged to transcribe in circumstances free of noise and were encouraged to use high-fidelity headphones if any one else was in the laboratory at the time.

Prior to initiation of the study, a meeting was held for introduction to the coding protocol. Practice utterances from infants were used in group coding during the initial meeting, to ensure that everyone understood the procedure. Because this effort was part of a broader study to evaluate predictors of transcription reliability, the coders were asked to make a variety of judgments for each utterance regarding syllable structure, canonicity, and their confidence in the judgments. For the present article, only the transcriptions themselves were considered. As per instruction, the transcriptions included IPA elements (see review in Ladefoged, 2001) from a variety of languages, and included a level of detail that the coders felt portrayed the utterances as accurately as the IPA would permit. They were instructed to avoid as much as possible being biased by their familiarity with the English phonemic system, a point that had been previously emphasized in classroom training on phonetic transcription. Consequently, although all of the transcribers were native English speakers, the data reported next often included symbols not present in the English phonemic alphabet or commonly heard in English speech (e.g., velar voiced fricatives and unrounded labial glides).

After the initial group training session, all coders were given practice on the protocol through transcription of a different 10-utterance pilot sample drawn from recordings of the same infant who provided the 30-utterance sample. Transcriptions were based on auditory perception only. Transcriptions of the 10 utterances were reviewed to ensure that the procedure had been clearly understood before the formal transcription study began.

For the formal investigation, all coders were asked to transcribe the 30-utterance experimental sample of infant vocalizations, according to the protocol that had been taught, again based on auditory-only judgments. Each transcriber worked completely independently and was allowed to listen to each utterance for a total of 12 times in order to make all of the structural, canonicity, confidence, and transcription judgments that were required for the infant sample within the broader study of which this one was a component. The coding protocol was abbreviated for all other transcription samples. For the toddler and adult speakers, coders were asked to listen to the words produced, no more than four times, and to transcribe them. No judgments of structure, canonicity, or confidence were required for these samples because they were not part of the broader effort, but were designed particularly for the present study. Again, the judgments were based on auditory information only.

The infant sample was transcribed first by all the coders, because that was the original focus of the work. Months later, the three adult samples were recorded and transcribed. Six transcribers coded all of the samples. Four of these six coded the toddler sample last, 9 to 21 days after the adult samples. The other two transcribers coded the toddler and adult samples on the same day. The data showed no notable differences in patterning of agreement across samples from the various transcribers.

Aligning Codes

Transcriptions differed along various dimensions: the number of segments transcribed, the choice of particular segments coded, and the level of detail portrayed. These differences were often so large that it was not obvious which symbols corresponded to each other between transcriptions. To analyze agreement among coders, decisions had to be made about alignment. This is a general problem of reliability assessment for temporally ordered data of any kind. In what follows, we specify the principles of alignment we utilize and recommend.

After pairs of coded transcriptions were merged in LIPP, the second author aligned the codes. For each utterance in each sample, a single standard transcription was adopted for the bulk of the analysis in order to simplify the comparisons. This work makes no assumption that the standard coder's transcription was the "correct" one. A single coder was used as a standard to simplify a very complex analysis, and by default the most experienced coder's transcriptions were chosen to serve that role. Later, the basic patterns of the data were verified to be essentially identical if a different randomly chosen coder was treated as the standard (see the Calculations section).

Our lack of commitment to a single correct transcription owes to the fact that infant utterances are often phonetically ambiguous, such that several differing IPA transcription options are compatible with the perceptions of listeners. Thus, more than one transcription can often be deemed correct. Even the individual listeners in this work noted (as is usual in transcription experience) that they could often hear an individual utterance in multiple ways on different playbacks. This study is not focused, then, on degree of correctness of transcriptions, but on degree of agreement among listeners, who can be thought of as auditory/perceptual equals.

The standard coder's (first author's) transcription was assigned to the target row (the upper row) in LIPP, and a comparator coder's transcription was assigned to the transcription row (the lower row) for each utterance. The 30-utterance transcriptions from each of seven comparators were thus aligned in seven LIPP files with the 30 transcriptions of the standard coder.

Alignment was conducted according to four principles. First, no segment reordering on either row was allowed. This is the *strict order principle*. Second, if the two transcriptions showed the same number of margin (consonant-like) and nucleus (vowel-like) segments in the same order, they were simply aligned according to order with no empty slots. This is the *matched segment principle*. The notions "margin" and "nucleus" were interpreted strictly in accord with IPA symbology—syllabic consonants (which are indicated in IPA by consonant symbols marked with the syllabic diacritic, as in the syllabic liquid [l̩]) were treated as nuclei, and nonsyllabic vowels (which are phonetically interpreted in the IPA as glides and are indicated by vowel symbols marked with the nonsyllabic diacritic, as in the schwa glide [ə̯]) were treated as consonants. Nuclei were, in accord with this approach, preferentially aligned with other nuclei, whether their base symbols were consonants (diacritically marked as syllabic) or vowels, and margins were preferentially aligned with other margins, taking fully into account both base symbol and diacritic indicators of margin status. Margins were aligned with nuclei if and only if there were no margin-to-margin or nucleus-to-nucleus options open for alignment that did not violate the strict order principle.

If, on the other hand, different numbers of margin or nucleus segments occurred across the transcriptions, or if they were not ordered in the same way, then without reordering any segments on either row, segments were aligned so as to minimize the phonetic discrepancy (in phonetic features, see the Featural Agreement section) between the two transcriptions. This is the *minimum discrepancy principle* recommended by Cucchiarini (1996). However, the minimal discrepancy principle was applied in accord with an additional principle here. Nuclei (vowels or syllabic margins) were aligned first with other nuclei, and then margins (consonant-like elements) were aligned with margins in such a way that nucleus alignments were never shifted as a result of any subsequent margin alignments. This is the *nucleus alignment first principle*, which was instituted on the basis of the recognition that a syllable in a natural language must have a nucleus but can exist without a margin and on the corresponding theoretical assumption that syllable nuclei constitute the perceptual centers of phonetic processing (see, e.g., Clements & Keyser, 1983). In cases of transcription disagreement in which the number of segments attributed to an utterance was identical, but the number of nuclei differed, margin-to-nucleus alignment could occur—if and only if none of the previously mentioned principles was violated in the process.

Once aligned, an utterance consisted of a series of "segment columns" (or slots) spanning the two rows. In many slots, there were segments transcribed in both rows, but in many others, there was an "orphan" segment in one of the two rows (when the two transcribers did not agree on the occurrence of the segment). The number of segment columns (or slots) for each utterance–transcription pairing was determined by the number of cases in which one transcriber or the other included a segment in the transcription. In Example 1, there are five segment columns for both transcribers, with the same number of nuclei and margins in the same order, thus the alignment is straightforward. In Example 2, on the other hand, there are six segment columns even though each transcriber indicated just five segments, consequently there are two orphan segments (the glides [w] and [ʋ]).

Example 1:	Standard transcription	[ə	ʋ	a	z̻	a]	
	Comparator transcription	[ə	ʋ	aː	β	ə]	
Example 2:	Standard transcription	[b		a	ʋ	p	ə]
	Comparator transcription	[m	w	æ		b	ʌ]

Calculations

Once aligned, the weighted transcription agreement was calculated by the LIPP program, which automatically compared each coder's transcription with that of the standard coder. The weighted values of agreement can be viewed as a general measure of proportion of transcription similarity. Utterance by utterance and segment by segment, the program assesses the degree of similarity between the standard coder's transcriptions and those of each comparator coder. Calculations are provided by the program at both the line and session levels.[2] Because transcribers entered one utterance per line in LIPP, the calculations that were automatically provided corresponded to the utterance and the session levels for each transcriber pairing. Segment-level calculations were not directly utilized in this study. In what follows we provide an outline of the principles by which the agreement analysis works. A full portrayal of the program would take many additional pages, but the code of the program can be obtained in full by request from the authors.

The program produces two component measures for each utterance (or session) along with an overall measure. The first component measure, termed *global structural agreement*, represents the proportion of segment slots in the aligned utterances in which both transcriptions include a segment. The second component measure, termed *featural agreement*, represents the proportion of phonetic information shared in segments that are present in the same slot in the two transcriptions. The third measure is the *overall transcription agreement*, which is obtained by multiplying the global structural agreement value by the featural agreement value.

Global structural agreement. To determine the global structural agreement, each coded segment slot is assigned an agreement value of one if the slot includes both a referent and a comparator segment. Otherwise (i.e., for slots with orphan segments), the global structural agreement value is set to zero. The global structural agreement value for the utterance is the mean of the structural agreement values for all of the segment slots. Consider Example 3:

Example 3:	Standard transcription	[pʰ	ĩ	n	]
	Comparator transcription	[b	i	d	i]

There are four slots with three shared (the comparator includes an orphan [i] in the fourth slot), so the global structural agreement is .75.

Any slot in a transcription comparison where either transcription includes an off glide or a glottal consonant that ends a syllable is treated in the current version of the LIPP program as a half-weighted slot. Thus, if one coder transcribes an utterance as a CV, and another transcribes the utterance as a CV plus a nasal consonant, the global structural agreement is .667 (2 agreed slots divided by 3 total slots). However, if one coder transcribes an utterance as a CV, and another transcribes the utterance as a CV plus an off glide, the global structural agreement would be .8 (2 agreed slots divided by 2.5 total slots), because the orphan off glide is treated as pertaining to a half-weighted slot. This provision takes account of the idea that a syllable skeleton can have more central and more peripheral elements, with off glides and glottal offsets playing less central roles than other consonants in syllable structure. Other provisions of this sort could be added to the program at a later date to take fuller account of emerging theory of internal syllabic organization.

Featural agreement. To determine featural agreement, only slots that are filled in both rows of the comparison are considered. For these paired (or nonorphan) slots, the featural agreement value is determined by the proportion of shared phonetic features. If the segments are identically transcribed, then all features are shared and the value assigned to the slot is one. If no phonetic features are shared, the value assigned is zero. If some, but not all features are shared, featural disagreements are weighted such that each disagreement produces a reduction in agreement score where a fixed proportion is subtracted from one.

The weightings of reliability score reductions for featural disagreements abide by the following general principles. First, for both nuclei (vowels) and margins (consonants), the value for each slot is distributed across all of the features that can differentiate segments of that type (nucleus or margin). This is the *totally distributed weights principle*, by which each feature that contributes to a segment's character must have a weight greater than zero, and by which the sum of all weights possible for features contributing to the segment's full value must equal one. In accord with the second, or *equal steps principle*, score reductions for each featural disagreement type are distributed equally, unless there is a theoretical reason (e.g., based on feature geometry embedding or markedness)

[2] Throughout this article, reliability is assessed point to point, in accord with the alignment procedures. Session-level reliability thus refers to point-to-point agreement averaged across utterances within samples (either 30 infant utterances or 20 utterances from any of the other speakers). In other publications (e.g., Oller & Eilers, 1988), the notion of session-level reliability is differently defined: In such cases, each coder's results are analyzed separately across utterances in a sample, and then the summary statistics from the two coders are compared.

to assign a higher score reduction to one type of featural disagreement than another. Thus, if half the features required to characterize a segment differ in the standard and comparator transcriptions, the featural agreement score is set to .5, if one fourth of the features differ, .75, and so on. If there are only three features deemed relevant to specifying the phonetic nature of a particular segment type (and there is no theoretical reason to assign them different weights), each is assigned (in accord with the equal steps principle) a weight of .33, and so on.

Where appropriate, the most global features that characterize a segment encompass contrasts among subfeatures. Place, for example, is one of three global consonant features recognized by the program, and the distinction between labial and labiodental represents a contrast within the place feature. "Big" disagreements within place correspond to maximal contrasts (e.g., labial vs. uvular or labial vs. glottal) within the global feature and receive the total score reduction applicable to the global feature (.33), whereas smaller disagreements (e.g., labial vs. labiodental) are assigned a half score reduction (.167), in accord with the equal steps principle. Similarly, score reductions based on differences in manner (the second global consonant feature) are scaled to reflect sonority. A difference between a stop and a glide is treated as a big difference (an obstruent alternating with a sonorant), corresponding to a maximal score reduction within manner (.33), whereas a difference between a stop and an affricate (an obstruent alternating with another obstruent) is treated as a small difference within manner (score reduction = .167). The program in its current form treats all differences between transcriptions on the global feature voicing (encompassing both voicing lag and voicing lead) as big differences (score reduction = .33).

Consider again Example 3, where the voicing disagreement between the standard transcription [p^{h}] and the comparator yields a .33 score reduction (featural agreement for the segment = .67). An additional .33 score reduction would occur if, in addition to the voicing disagreement, there were a big disagreement in place of articulation. Thus, if [p^{h}] → [g] (defined in the program as a big place disagreement), then featural agreement would be .33. If, in addition to the voicing disagreement, there were a big disagreement in manner of articulation, there would also be an additional .33 score reduction. Thus, if [p^{h}] → [m] (an obstruent/sonorant alternation), then featural agreement would also be .33. If all three big disagreement types occurred in the slot (voicing, manner, and place, e.g., [p^{h}] → [ŋ]), then featural agreement would be zero for the segment.

Returning to Example 3, the big manner disagreement on [n] in the example (which is transcribed by the comparator as [d]) produces a total score reduction of .33 (featural agreement = .67). Smaller disagreements of place or manner are treated as embedded within the place or manner features according to a simple phonetic feature geometry (see, e.g., Goldsmith, 1995). These smaller disagreements are weighted as subcomponents of the higher order features by the program in equal steps within the .33 value that is assigned to both manner and place. Thus, a small place disagreement (corresponding, e.g., to [k^{h}] → [c^{h}], a velar stop alternating with a palatal) would result in a score reduction of .0833, because this disagreement can be viewed as representing a step embedded within the medium place disagreement, [k^{h}] → [t^{h}] (which would result in a score reduction of .167), and even further embedded within the big place disagreement, [k^{h}] → [p^{h}] (which would result in a score reduction of .33). The logic of the approach is based on the simple articulatory facts of distance between places of articulation, with maximal distances yielding maximal score reductions, and with equal steps of reduction at each level of the feature geometry in accord with the equal steps principle.

The equal steps principle yields weightings based on the number of features that characterize a segment. In some cases, these weightings are additionally adjusted in accord with the *markedness principle*, to account for such factors as frequency of occurrence of segments and of contrasts between segments in natural languages. For example, some vowels are common and tend to be universal across languages. Empirical evidence suggests that there is a physical basis for this pattern of commonness (viz., discriminability). The most commonly occurring vowels, the most unmarked ones ([a, e, i, o, u]), are the most discriminable from each other (Lindblom, 1992; Lindblom & Maddieson, 1988). Nasalized vowels, on the other hand, are less common (more marked) in languages than oral vowels (Greenberg, 1966), and, predictably, contrasts between nasalized and oral vowels at the same height and frontness (e.g., [o] vs. [õ]) are less discriminable than between unmarked oral vowels (e.g., [o] vs. [u]). Thus, in our weightings, a nasalization disagreement in vowels (e.g., [o] vs. [õ]) corresponds to a marked contrast and produces a lower score reduction (.1) than a "small height" (e.g., [o] vs. [u]) disagreement (.2), which corresponds to an unmarked contrast. The differences in weighting of disagreements correspond to theoretically different degrees of importance of nasalization disagreements as opposed to height or frontness disagreements, a pattern of importance corresponding to degree of markedness of the contrasts. In Example 3, the second standard transcriber segment [ĩ] is transcribed without nasalization by the comparator coder, resulting in a score reduction of .1 (featural agreement for the segment = .9). The equal steps principle and weighted adjustments based on the markedness principle are never allowed to violate the totally distributed weights principle for each segment. Thus, the sum of possible score reductions for each segment type is always one.

Featural agreement scores are also adjusted for alternations that might be thought of as crossing the consonant–vowel boundary. For example, a syllabic consonant, such as the *n* in the word *button* pronounced in the standard American style, is transcribed in accord with IPA using a diacritic that indicates the "consonant" [n̩] functions as a syllabic nucleus (or vowel-like element) in this case. In our alignment procedure, a syllabic consonant from one transcription corresponds to a vowel in the other transcription if the two elements can be interpreted as pertaining to the same syllabic nucleus (in accord with the strict order and matched segment principles). For example, one transcriber might indicate a nasalized vowel in a location where another transcriber might indicate a syllabic nasal consonant. Here, because both transcribers indicate a syllabic element, score reductions are only half as great (.5) as they would be if one transcriber indicated a nasal consonant (nonsyllabic) and the other a (mid or low) nasalized vowel (yielding the maximum agreement score reduction of 1.0). In the latter circumstance, the discrepancy is treated as more severe, according to the phonological principles adopted here, because the transcribers differ not only in the types of segments invoked in the transcriptions, but also in the number of syllables attributed to the utterance. A difference among transcribers in number of syllables, resulting in an orphan slot after alignment, always results in the maximum score reduction (1.0). Disagreement in type of nucleus segment indicated at a slot can produce much smaller score reductions.

The values of score reductions for margin-nucleus disagreements are always high (ranging from .5 to 1.0), but the precise values are determined in accord with a sonority principle; thus, obstruent-vowel alternations (which show very large sonority differences) produce maximal score reductions (1.0) whereas alternations of high vowels with high glides (which show much smaller sonority differences) result in the minimal score reduction for margin-nucleus alternations (.5).

The mean featural agreement for all of the paired slots of each aligned utterance represents the featural agreement for an utterance. In the case of Example 3, the mean value is .74; that is, the mean of the featural values for the three nonorphan slots [(.667 + .9 +.667) / 3].

Overall agreement. The overall agreement for each utterance is computed by multiplying the global structural agreement value by the featural agreement value. For Example 3, .75 (global structural) times .74 (featural) equals .56, the overall agreement value for the utterance.

The product of the two values represents a proportion of a proportion. The first proportion is based on the number of segments that are shared in any way among the transcribers, and the second is based on the proportion of information that is shared within the segments that are shared in any way. To gain an intuitive understanding of why multiplication is appropriate to combine the scores, consider the following example. If half of the segment slots are filled on both rows of a transcription comparison, the global structural agreement is .5, indicating that the two transcribers share only half the segments. Half of the phonetic information is literally missing from one transcription to the other. If all the shared (nonorphan) segments differ additionally by a mean of .2 of the features included in them, the total overall agreement for the utterance would be half of .8 (i.e., .5 × .8), because the overall agreement should be thought of as the proportion of features shared on paired segments as a proportion of slots shared. Thus, the overall agreement after multiplication would be .4. The temptation to compute the overall reliability by taking the mean of the global and featural values must be resisted, because that would result in an overall value greater than .5, clearly a misleading value, since only half of the segments are shared at all by the two transcribers in the example, and the ones that are shared show additional disagreements.[3]

The overall transcription reliability measure provided here represents an attempt to establish a stable method for comparing transcriptions that is adapted to a wide variety of degrees of discrepancy, even to circumstances in which very large differences may exist between transcriptions (e.g., in transcription of infant vocalizations or very disordered speech). While the procedure includes score reductions for the great bulk of unmarked feature disagreements, some featural disagreements that would represent especially marked details do not produce reliability score reductions in the present form of the program. Many highly marked details proved unnecessary to account for because the transcribers in the study did not utilize all of the diacritic options that would have invoked such details (which may or may not indicate the infants did not produce sounds corresponding to those diacritics). Score reductions for such disagreements could of course be easily added to the program at a future point. In addition, the program does not attempt to take into account all characteristics of phonological theory, partly because there is substantial disagreement among phonologists about details of theory and partly because we assume there may be a point of diminishing returns in tweaking the program with additional details of phonological theory. This article presents a first

[3] By simple algebra, it is clear that the multiplication of global structural agreement by featural agreement is equivalent to taking the mean of the values across slots after assigning a zero to all orphan slots and assigning an appropriate featural value to all other slots. If N = the number of slots, X = the number of nonorphan slots, and ΣF = the sum of the featural values computed at each nonorphan slot, then

$$(X/N)(\Sigma F/X) = \Sigma F/N.$$

Because we can cancel the two Xs on the left side of the equation, the equivalency is obvious.

approximation to a weighted measure that relies on tenets of theory that are tried and true, and it can be adjusted at a later point to incorporate additional theoretical features as they become more stably accepted among phoneticians and phonologists.

Clearly, the decisions about how to weight score reductions represent a combination of judgments based on the markedness principle, plus the numerical scaling of the weights based on the equal steps principle and the totally distributed weights principle. In essence, the LIPP program translates an ordinal scale of theoretical featural and markedness relations for individual segments to an interval scale using the weighting principles.

The new measure provides this transformed scale at every level of analysis, from the segment, through the utterance, through the recording session—and all of these levels are computed for the data to be analyzed below in the LIPP program. For the present purposes, segment-level computations remain internal to the LIPP computations, but both utterance-level agreement scores and session-level scores (computed as the mean across utterances within each sample for each transcriber pairing) are reported.

The traditional measure of percentage agreement is also typically treated as an interval scale for agreement on utterance transcription or recording session transcription. However, traditional percentage transcription agreement is based on binary judgments at the segment level. We reason that the failure of the traditional measure to richly represent degrees of agreement at the segment level may hamper its applicability in measuring transcription agreement at every level of analysis in infant vocalizations and in other particularly difficult cases of transcription. The approach suggested here (as well as the similar approach discussed in Cucchiarini, 1996) refines the scale of proportion of information shared across transcriptions in a way that is consistent with phonological theory regarding featural relations and markedness, and thus is expected to provide a firmer foundation for transcription agreement analysis, especially in cases such as those presented by infant vocalizations.

Results

Distribution of Agreement Values for Weighted and Unweighted Measures on the Infant Sample

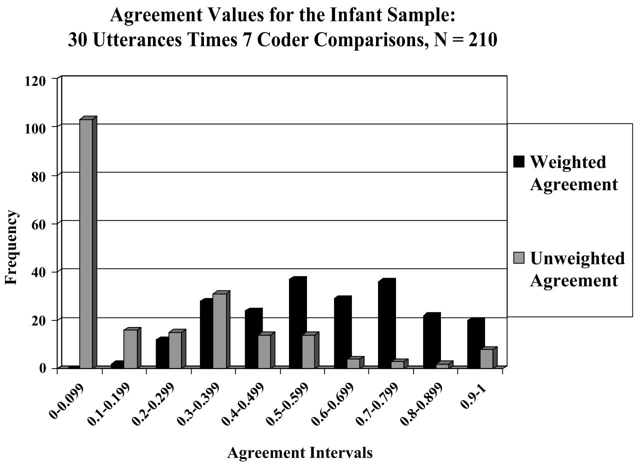
The agreement data for the eight transcribers were analyzed in seven pairings of the standard transcriptions with the comparator transcriptions. All transcriptions were also compared with a second standard referent, chosen randomly from the other seven coders, and extremely similar results were obtained on overall weighted transcription agreement. In fact, the values for the two standard transcribers were identical to two significant digits for the overall agreement measure averaged over the 30 utterances and seven transcriber pairings. Therefore, the data presented here will be referenced to a single standard coder.

Each of the 30 infant utterances was assigned an overall weighted transcription agreement value for the seven pairings with the standard coder's transcriptions through the LIPP analysis program described previously, yielding 210 agreement values at the utterance level. In addition, another LIPP analysis program was implemented to produce overall unweighted transcription agreement values based on the absolute match criterion, yielding an additional 210 agreement values.

Frequency distributions of the 210 values obtained illustrate scaling properties of the two measures. Figure 1 shows that the weighted measure produced agreement values spanning the range from under .2 to 1, with many values at every interval above .2. Figure 1 also shows that the unweighted measure is severely skewed. Nearly half of the values were zeroes, reflecting the fact that the absolute match criterion does not permit discrimination of degrees of disagreement between transcriptions of individual segments and skews the distribution of utterance-level agreement.

For the unweighted measure, the low number of values at any interval above the first shows that much of the potential of the scale is wasted in characterizing degree of transcription agreement for individual utterances. A sample of vocalizations such as this one, with many examples of utterances that are difficult to

Figure 1. Frequency distribution of utterance scores for 210 comparisons (7 transcriber comparisons with the standard transcriber times 30 utterances) under the weighted and unweighted agreement measures for the infant sample.



transcribe, and thus many cases of transcriber disagreement, produced many zeroes under the absolute match criterion. The weighted measure, on the other hand, scaled differences more evenly across the range, with the exception that no values occurred in the first interval and only a few in the second.

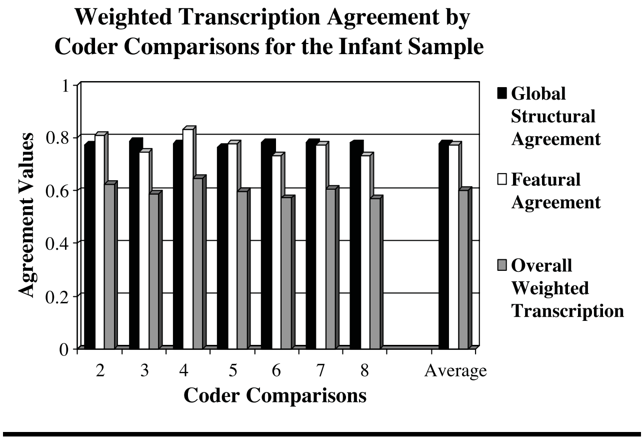
Another way to consider the scaling for the two measures is to consider the frequency distributions for the 30 infant utterances computed as a mean across the seven transcriber pairings. The lowest overall weighted transcription agreement, computed this way, for any of the 30 utterances was .332, and the highest average overall weighted transcription agreement was .850. For the unweighted measure, the lowest and highest utterance mean agreement values computed across the seven transcriber comparisons were 0 and .59, respectively. If we consider the mean weighted agreement score for the 30 infant utterances, the value (.60) again falls in the middle range of the scale, whereas the value for the unweighted measure is very low (.21), and not far from the floor of the scale. The mean agreement values for the seven transcriber comparisons computed across the 30 utterances were all consistent with this observation; the values ranged from .56 to .64 for the seven transcriber comparisons on the weighted measure and from .15 to.28 on the unweighted measure. Furthermore, all seven transcriber comparisons showed many zeroes on the unweighted measure, ranging from 12 zeroes out of 30 utterances to 17 out of 30. There were six utterances for which all seven transcriber comparisons showed zeroes on the unweighted measure. Thus, when we consider the mean data for each utterance across the group of transcriber comparisons, the weighted measure utilized the middle of the scale for the sample much more effectively than the unweighted measure for characterization of individual utterance agreement. Again, we see the skewing impact of the floor effect for the unweighted measure.

The 103 zeroes out of 210 comparisons in the unweighted data in Figure 1 illustrate a floor effect. None of these utterances actually showed a total lack of agreement across the transcriptions—there were typically agreements on the number of consonant and vowel segments in each utterance and other agreements on featural information related to segments. The floor effect illustrates the fundamental disadvantage of the unweighted measure in assessing agreement among transcribers for samples such as this one, in which utterances with a wide range of transcribability are included.

As a concrete illustration of the disadvantage of the unweighted approach, consider again Example 2. The two transcriptions have much in common. Both indicate that the utterance consists of two syllables, both beginning with a labial stop. The nuclei transcribed are similar, both low and central to front in the first syllable, and both central in the second, one mid and one mid to low; thus, the vowels differ by only a small featural step in each case. The major difference between the two transcriptions is that an unrounded labial glide [ʋ] was heard after the first nucleus in the standard transcription, whereas a rounded labial glide [w] was heard before the first nucleus in the second case. There are six slots in the comparison, and because no slot shows absolute match, the unweighted measure yields an agreement value of zero. The weighted measure on the other hand gives credit for agreement to many features ignored by the unweighted measure, including all those indicated above, yielding a value of .567.

The stability of the weighted transcription agreement measure is illustrated in another way in Figure 2. The three calculations of main interest (global structural agreement, featural agreement, and overall weighted transcription agreement) are displayed for each of the seven transcription pairings with the standard transcription. Each individual coder, as compared with the standard coder (Coder Number 1), is displayed along the *x*-axis. Therefore, the comparisons begin with Coder 2 compared with Coder 1, indicated by the number two in the figure and so on. The mean overall weighted transcription agreement for the infant vocalization sample was .60, with similar results across all coder pairings; the average standard deviation of the mean for the seven pairings was only .028. The pairing that showed the score that differed most from the mean differed by only .047, suggesting relative concordance of agreement values regardless of transcriber pairing. The average global structural agreement was .778, and the average featural agreement was .772, again with relatively little variation across the seven transcriber comparisons (mean standard deviation for the coder pairings = .007 and .039, respectively). These data thus illustrate substantial stability of the weighted measure across coder comparisons.

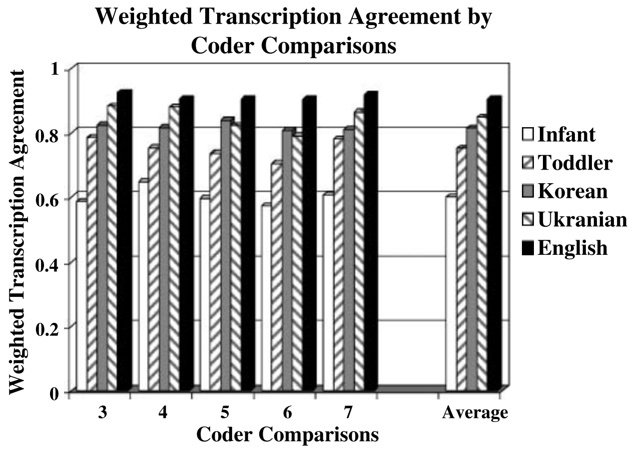
Figure 2. Agreement values for the weighted measure across the pairings of the standard transcriber with the seven comparator transcribers for the infant sample.



Orderliness of Agreement Values for the Weighted and Unweighted Measures Across Different Samples of Vocalization

Figure 3 provides a comparison of overall transcription agreement across the various samples of vocalization transcribed in the study for both the weighted and the unweighted measures. Both measures show an orderly relation among the values of agreement obtained across the samples, with the infant sample showing the lowest and the adult English sample showing the highest values. Because all of the transcribers were native English speakers, it is to be expected that the agreement values for the Ukrainian and Korean samples would be somewhat lower than those for the English sample.

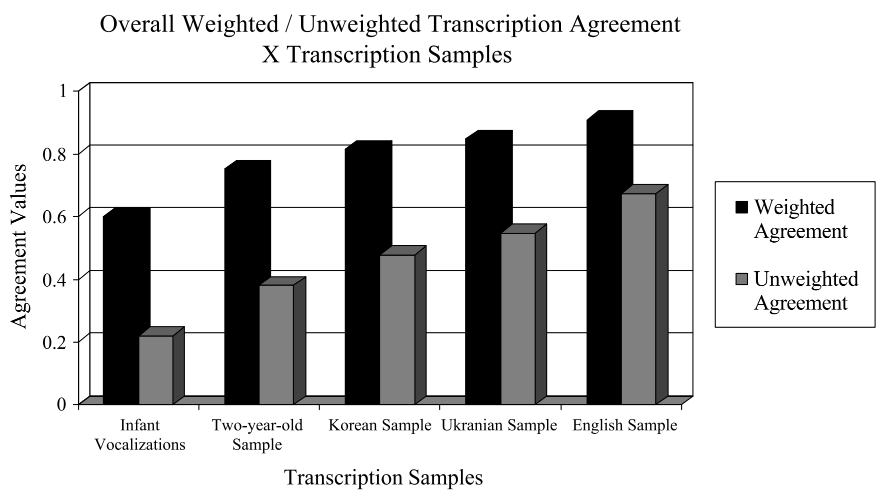
Figure 4. Agreement values for the weighted measure across the five transcriber comparisons with the standard transcriber that were available for all five samples.



The agreement calculations for the weighted measure averaged across all five coder comparisons available for all samples are displayed in Figure 4. The figure again demonstrates that there are logical, predictable relations between agreement values for the samples, and that the orderliness applies to all of the coder comparisons. Each individual coder, as compared with the standard coder, is displayed along the x-axis, which begins with Coder 3 compared with Coder 1 and ends with Coder 7 compared with 1 (because only Coders 3–7 and the standard coder transcribed all five vocalization samples). The infant vocalization sample yielded the lowest overall weighted transcription agreement across all of the coder comparisons, and the adult English sample yielded the highest, again in all cases. Furthermore, the toddler sample always showed the second lowest value. For these five samples of vocalizations, the variation of agreement values across the transcriber pairings was always small, with a mean SD = .026, indicating again that the weighted measure maintains substantial stability across various comparisons of transcribers. In contrast, the unweighted measure showed higher variability (mean SD = .046) across the transcriber pairings, even though for the infant sample the mean standard deviation was quite low (mean SD = .024) presumably as a result of the floor effects noted earlier.

Another way to illustrate the scaling stability of the weighted measure across transcriber pairings is to note that for every pairing, the lowest agreement value pertained to the infant sample, the second lowest to the toddler sample, and the highest to the English adult

Figure 3. Average agreement values between comparator transcribers and the standard transcriber for the weighted and unweighted measures across the five sample types.



sample. The unweighted measure showed the same predictable relations. For both weighted and unweighted measures, the Korean and Ukrainian samples were intermediate between the English and toddler samples for every pairing. Even the absolute values for all of the available pairings of transcribers across all the samples (seven samples for the infants, five for the toddlers, six for Korean, Ukrainian, and English) showed orderliness and stability across all transcriber pairings. Under the weighted measure, all showed an agreement value lower than .65 for the infant sample, whereas every pairing for the toddler sample yielded a value higher than .65. All showed agreement values for the Korean and Ukrainian samples that were higher than the highest toddler agreement value, and the English agreement values were higher than the highest Korean and Ukrainian values in every comparison but one. Thus, 29 of 30 comparisons followed a pattern indicating that the weighted measure successfully described differences in transcribability that were both predictable and intuitively satisfying.

Discussion

Conceptual and Empirical Outcomes

The present work was inspired by a long-term and continuing effort in the study of the foundations of phonetic capabilities as manifest in the vocalizations of infants and young children, both monolingual and multilingual, as well as in speakers with various sorts of disorders of communication (Eilers, Oller, & Benito Garcia, 1984; Oller, 1995; Oller & Eilers, 1982, 1988; Oller, Wieman, Doyle, & Ross, 1975). Such work has led to novel approaches to description of sounds that prove particularly difficult to transcribe (in particular, in protophone and infraphonological coding, see Nathani & Oller, 2001), but neither we nor other investigators in the field have abandoned the selective utilization of phonetic transcription as a systematic indicator of auditory perception of speech and speechlike sounds. Our intention all along has been to understand better the limitations on such auditory perception and thus to better grasp the utility of phonetic transcription in difficult cases (Oller & Eilers, 1975). In the recent pursuit of research on predictors of transcription reliability (in particular, effects of degree of canonicity of vocalizations and transcriber confidence judgments), currently under study in our laboratories (Ramsdell & Oller, 2005), we were led to develop improvements in tools for assessment of reliability. Without an assessment measure that could characterize proportion of transcription agreement consistent with principles of phonetic and phonological theory, and without a measure with stable and well-distributed scaling properties, we reasoned we might fail to discern predictors of reliability simply because the measure of reliability itself could be too weak to support stable correlations.

In this context, we set about to specify criteria for analysis of transcription agreement according to principles founded in widely accepted tenets of phonological theory. Both alignment of transcriptions to be compared and weighting of discrepancies between aligned segments were specified in basic accord with theoretical phonological principles. The principles were laid out earlier as explicitly as possible in the space allotted.

The alignment principles are of sufficient simplicity that it will be possible for any phonologically knowledgeable programmer to implement them with little or no ambiguity in transcription and phonetic analysis software, and we expect our collaborators to provide such programming in the not too distant future in LIPP. The principles of weighting for assignment of score reductions to different types of transcription discrepancy are not so unambiguous, however. The description of the weighting principles provided here portrays the flavor of the method, without specifying precisely how all possible weightings are implemented. There is insufficient space to do so, and consequently we have offered to supply the program on request. Furthermore, there are many different options for weighting that are consistent with the principles, depending in part on the particular set of phonetic features one chooses to employ. For example, in implementing the totally distributed weights principle, decisions must be made about the level of detail to be used in featural definition of segments. Distribution of weights cannot be fully implemented until the total characterization of each segment is in place, and any change of decision about the number of features to employ for a segment may require readjustments to ensure comparability of weighting across all segments in the inventory that might share the feature or features being changed and to ensure consistency with the totally distributed weights principle and the equal steps principle.

Because there exists no final fixed theory of phonetics, it seemed wise for the present effort to concentrate on creating a method consistent with only the most widely shared aspects of phonetic and phonological theory while making the tool flexible enough to be easily adapted to potential changes in theory or empirical evidence about relations and perceptibility of speech sounds. The programming in LIPP that underlies the present work has just this sort of flexibility. Not only can weights be easily reassigned, but also featural definitions can be revised by reformulating or even adding or subtracting features. Even new segments can be easily added to the transcription and transcription analysis approach within the program environment.

Particularly important potential improvements in the weighting procedures can be envisioned in order to incorporate more information related to a hypothesized skeletal tier of syllable organization (McCarthy, 1985). Within this model, consonants have different roles in syllables depending on whether they are onsets, beginning a syllable, or codas, ending a syllable, and depending on whether they form part of clusters, which can similarly be broken down based on sonority principles. Vowel elements also have different roles depending on whether they are pure nuclei or diphthongs. Such principles of syllable organization are becoming increasingly standard in phonological theory, but the weighting principles implemented in the current version of the LIPP program developed here represent these skeletal syllable features only partially. For example, consider the provision that an orphan off glide or a glottal consonant ending a syllable is treated as pertaining to a half-weighted slot in the current version of the LIPP program. This provision takes account of theoretical differences in the weighting of different syllable skeleton elements, but it is not the only such adjustment that might be considered. Other provisions of this sort could be added to the program to take fuller account of emerging theory of internal syllabic organization.

Furthermore, in part because nonlinear syllable-level and higher order rhythmic-level information was not directly assessed in the transcriptions provided here, the program does not currently offer weightings for transcription disagreements on stress, tone, or duration of syllables. In general, this version of the weighted transcription agreement procedure can be thought of as almost entirely "linear" (for an introduction to the distinction between linear and nonlinear phonologies, see Goldsmith, 1976, 1995). However, the approach does not need to stay that way in the future, because the programming tools of LIPP can support easy incorporation of both more metrically sophisticated transcription and more metrically sophisticated analysis.

In addition, the alignment of segments in the transcription agreement approach utilized in this study is relatively linear and nonmetrical. The nucleus alignment first principle constitutes a nonlinear enhancement that was introduced in the present version of the alignment principles, taking the centrality of nuclei (as opposed to margins) into account. A more complete nonlinear, metrical approach to alignment, consistent with a wide variety of theoretical approaches to phonology, can easily be envisioned. In particular, it seems clear that alignment should be based (assuming that stress is transcribed) on stressed nucleus alignment first, followed by alignment of unstressed nuclei, and last by margin alignment (with the constraint that no-nucleus alignment can be overridden in subsequent alignment operations). Such an approach would take account of the perceptual saliency of stressed nuclei and of the fact that stressed nuclei regularly attract intonational tone and length in phonological operations (see Hayes, 1995; Liberman & Prince, 1977). Again, it will be easy in the future to adapt the approach utilized here to include more metrically friendly principles of alignment.

On Optimal Scaling Characteristics

The data demonstrate that the scaling properties of the weighted agreement measure are robust, even with this small sample. Still, it is clear that in typical vocal samples there are important variations session to session within subject, across infants at the same age, and across different circumstances of recording (see, e.g., Delack, 1976; Legerstee, 1991; Lewedag, 1995; Oller & Bull, 1984). Generalizability of the outcomes reported here will clearly be enhanced with further investigation of more diverse samples.

Even in its relatively simple current form in terms of application of principles of phonological theory, the weighted measure indeed seems to offer considerable advantages in scaling over traditional unweighted percentage agreement. For the infant sample investigated here, the weighted method produced values across individual utterances that were distributed fairly evenly from .3 through 1. At the same time, the values for the same utterances were drastically skewed when analyzed with the traditional unweighted measure, with approximately half of the utterances for every transcriber pairing producing zero values. This floor effect would appear to severely limit the light that the traditional measure could shed on transcription agreement in cases as difficult to transcribe as the infant case studied here. Experience suggests that severely disordered speech can be equally difficult to transcribe.

The scaling problems of the unweighted measure do not pertain merely to especially difficult cases of transcription, however. In general, an ideal measure produces values across the entire scale for a wide range of samples within the populations in which the measure is expected to be used. If the scale produces less than the full range, its potential is partially wasted, and it should be redesigned. The unweighted measure appears to be skewed rightward in general, as indicated by many zeroes for the infant sample, and by the fact that even in an optimal case of intelligibility and well-formedness (citation form adult English, with English-speaking transcribers), transcriber agreement scores at the utterance level showed an apparent ceiling at less then .7 (Figure 3), a clear indication that the top end of the scale is squandered in the unweighted approach because of the disproportionate occurrence of very low scores.

In the present study, even the weighted measure failed to produce an outcome spanning the entire scale at the utterance level of analysis. No values occurred in the first interval for the infant data, as indicated in Figure 1, and only a few in the second interval. However, this limitation in range of reliability scores was not caused by an inherent limitation of the weighted measure; the measure is capable in appropriate circumstances of yielding scores across the entire range. Consider first that agreement values presented at the utterance level are based on the averaging of segment-by-segment data. With utterance-level averaging, scores tend to inch up above zero every time there is any degree of agreement between segments within the utterance. At the segment level, however, the entire range of possible agreement values from zero to one did occur for the weighted measure; zeroes occurred, for example, in every case of orphan segments, in which one transcription included a segment that was absent in the other. Furthermore, even featural comparisons yield zero values with the weighted measure (when all of the features characterizing a segment are portrayed maximally differently in the two transcriptions). With the unweighted measure, the problem of floor effects and consequent skewing at the segment level is even more severe than at the utterance level, and the proportion of zeroes is even higher.

Second, we reiterate that the infant sample studied here was more than 40% canonical. Transcription of samples with lower canonical babbling ratios would produce even lower agreement values than those obtained here. The weighted measure leaves room for assessment even in cases of yet poorer agreement than those seen in this study. The unweighted measure, on the other hand, leaves little room for any assessment of transcription agreement in samples that are less canonical than this one, because even with this sample, the values for the unweighted measure were very close to the scale's floor.

Another point favoring the present form of the weighted measure, even though it produced few values under .2 at the utterance level in this study, is related to the context of transcription in general. In this study, transcribers were presented with preselected utterances. They did not have to find utterances to transcribe naturalistically from amid the flow of infant vocalization and surrounding noise. In maximally challenging comparisons of transcriber agreement, coders do face the naturalistic flow of vocalization. They hear recordings from which they must not only transcribe identified utterances, but from which they must also decide which elements of the recording are utterances to transcribe. They must exclude extraneous noises or utterances from speakers other than the target of the study, and they must decide whether to include utterances that are excluded by definition in most studies (e.g., crying utterances, vegetative utterances, or utterances produced at extremely low amplitude). In fact, locating utterances to transcribe in such cases is subject to notable disagreements, just as transcription itself is subject to disagreements. As an indication of the difficulty found in utterance identification, note that prior research on volubility in infant vocalizations has included extensive training of coders to prevent utterance disagreements from exceeding 15% of total utterances counted by different coders (Oller, Eilers, Steffens, Lynch, & Urbano, 1994). Of course, in an overall scheme of assessment of transcription agreement, failures of agreement on occurrence or nonoccurrence of utterances should be taken into account, and the weighted transcription agreement measure proposed here would assign zeroes to whole-utterance disagreements. Consequently, if the weighted measure is to be applied completely generally, it needs to have room for the very low (even zero) utterance-agreement values that would be expected to result from samples coded naturalistically from recordings and in which disagreements on occurrence of utterances are consequently inevitable. The weighted measure offers that sort of room.

At the top end of the scale, the weighted measure for the English data yields an agreement value exceeding .9 for citation-form utterances produced by a mature native English speaker. Clearly, the weighted measure uses the top end of the scale more effectively than the unweighted, yet why was the value not even closer to 1.0 for the weighted measure, if this represents an optimal case of transcription reliability? A firm answer would require additional experimentation and analysis, but it seems clear that in some cases in the transcriptions from the present study there were simple mistakes of application of IPA (e.g., use of a relatively unfamiliar symbol inappropriately) and others in which the transcriber may have simply failed to enter all of the detail that should have been included (e.g., failure to include the aspiration or nasalization diacritic). Other cases of disagreement appeared to be associated with the fact that syllabification of the English utterances was not always easy. The speaker for the English sample was a southerner whose vowels could sometimes be heard as one or two syllables depending on the listener's perceptual set. The listeners were from the South, the North, and the West. Thus, the value of .9 obtained with the weighted measure may be fairly reasonable for this sample, given who the transcribers were and who the speaker was.

Still, both the weighted and unweighted measures showed orderly relations among agreement scores obtained across the five samples of vocalizations studied here. This fact should temper somewhat any expectation that a weighted agreement measure is always advantageous. The more important point, as indicated here, is that the weighted measure showed orderly relations in a more sensible way when we consider the absolute values obtained between the two measure types. The weighted measure leaves room at the bottom end for lower agreement values than those obtained here (and it is clear such lower values will occur in naturalistic transcription of vocal samples that are not preselected) while providing high reliability scores (around .9) for optimally intelligible samples. The unweighted measure produces values that are impractically low at both ends of the scale. For the infant sample, as noted previously, the unweighted average of .21 is so low as to leave little room for characterizations of further reductions in reliability that would be expected with samples of utterances from even younger (and consequently precanonical) infants or very disordered speakers, or for reductions in reliability that occur in naturalistic coding that includes utterance identification. On the other end of the scale for the unweighted measure, the apparent ceiling value of (which the data suggest to be around .7) for transcription of mature English speech is also impractically low, and much of the scale's potential at the top is wasted. The pattern of outcomes suggests that differentiations of degrees of agreement that are possible within the weighted approach are compressed and less differentiable within the unweighted. Additional research comparing the two measure types in detail with samples from normal speakers would be useful to illustrate this point.

With improved scaling properties of the weighted measure in hand, we are hopeful that transcription reliability research will be able to reveal important predictors of transcription agreement through correlational research. Many individual examples of syllables with noncanonical transitions have been examined over the years in our laboratories, and in general we have seen that these present vexing problems for auditory phonetic interpretation. With improved scaling of measurement, we have found that stable correlations of transcription agreement with factors such as degree of canonicity of syllables can be appropriately assessed and reliably discerned (Ramsdell & Oller, 2005), and we are continuing with the correlational research, utilizing the weighted measure.

On Optimal Reliability Levels and Optimal Reliability Assessment Procedures

It has been pointed out to us that the primary concern of many who utilize phonetic transcription is not the nature of the appropriate measure to use, weighted or unweighted, but rather the level of reliability that should be considered adequate regardless of the measure. In fact, we know of no conceptual rationale that has ever been published for recommended acceptable reliability levels in phonetics. Determination of a single such level is not possible because of factors that have not been discussed, to our knowledge, in prior phonetics literature about acceptable level of reliability. It is important to consider the fact that reliability of measurement interacts systematically with other factors that influence significance of outcomes in both research and clinical evaluation. Reliability (or more precisely lack of reliability) contributes to error variance in quantitative evaluation, and consequently plays into the power to detect real differences between samples. Consequently, an acceptable level of reliability has to be determined on a study-specific basis.

Consider an example clinical evaluation: In attempting to detect a statistically significant change across time in a client's speech articulation, the reliability of the phonetic transcription measure of articulation is a source of error variance in the assessment, in a similar way to other sources of error variance based on sampling from any population of events or individuals. Thus, if transcription reliability is poor for the samples taken from the client, the ability of the assessment to detect significant change in articulation across time is curtailed, in a similar way to the curtailment of power that would occur if the real change (the effect size) is small or the number of utterances transcribed is small (the sample size). The power or sensitivity of the assessment procedure is limited by the phonetic transcription reliability, just as it is limited by the sample size or the effect size, or any other source of error variance.

The same sort of limitation in power of assessment is imposed by measurement reliability in group comparisons. So, for example, if speech articulation measured through phonetic transcription in clients with Down's syndrome is compared with clients who have apraxia of speech, lack of transcription reliability will again limit the power of the evaluation to detect differences between the two groups. This is true whether we consider single transcribers or multiple transcribers, because lack of reliability in both cases (intratranscriber and intertranscriber) contributes to error variance. If we fail to consider the role of reliability in statistical assessment (and in many areas of research and clinical practice, phonetics included, the role of reliability is usually ignored), we run the risk of Type II error. That is, we run the risk of failing to detect a real effect, and not realizing that the source of the failure is the unreliability of our measure.

The relation between reliability and power is well known in psychometrics (Cleary, Linn, & Walster, 1970; Nicewander & Price, 1983; Sutcliffe, 1958; Zimmerman, Williams, & Zumbow, 1993), but has never previously been mentioned to our knowledge in publications related to phonetic transcription reliability. Although high reliability values are always desirable (because they always correspond to higher power), low reliability values are not routinely fatal in research, because loss of power associated with low reliability can be made up for by favorable values on other factors that influence power (sample size, effect size, and other sources of error variance).

Consequently, the often-mentioned goal to determine a generally acceptable reliability level for phonetic transcription is unrealistic. The appropriate level of reliability for any study or any clinical evaluation depends on the relation of reliability with all of the factors that influence power in the evaluation. Reliability needs can therefore vary wildly. Modeling proves that even reliability levels of .9 and higher can easily be inadequate to prevent Type II error if other power factors are unfavorable, and even a reliability level of .5 can be adequate to detect differences across samples if other power factors are favorable (Bacon, 2004).

The psychometric research cited here is devoted in large measure to providing the tools needed to calculate measurement reliability levels necessary to achieve given power levels within varying circumstances of sample size and effect size. While the relevant psychometric studies have been conducted primarily with research in mind, they are equally relevant to interpretation of clinical assessment. The appropriate approach, in our opinion, to determining an acceptable level of reliability is study specific, and should utilize tables that indicate power outcomes based on the relation between effect size, sample size, and reliability values. Such tables are available in the psychometrics literature cited here.

As a research approach, the use of such tables to calculate acceptable reliability levels is plausible (and hopefully will begin to be used in our field in the near future), but in clinical practice it is difficult enough to generate transcriptions, let alone to evaluate the relation of reliability to other factors related to power. So it seems important for someone with appropriate training in psychometrics to develop simplified clinical recommendations about how to choose appropriate reliability levels on a practical basis. Even in phonetics research, simplification and clarification of the needs of reliability would seem to be a high priority. Phonetics, both from the standpoint of research and clinical work, is in fundamental need of input from the field of psychometrics to help improve measurement procedures. Our effort in this article does not directly contribute to that need, but only provides a precursor to it: The weighted measurement tool we have described should offer a more stable and well-scaled approach for evaluating reliability in anticipation of development of guidelines for appropriate levels of reliability in research and in the clinic.

With increased awareness of the relation between reliability and test sensitivity, it should become more evident that procedures (whether in research or clinical practice) that acquire data without an appropriate measure of reliability are very risky indeed. Interpretation of data acquired without a reliability measure (regardless of the number of transcribers) would seem to be based on the tacit assumption that reliability is 100%, an entirely implausible state of affairs. In cases in which for practical reasons (especially in the clinic) only one transcriber is possible, it may be important to acquire reliability information based, at least, on a subsample of the kind of data evaluated in the clinic. Such reliability checking can reasonably be done quickly at broad intervals as long as transcription procedures do not change from sample to sample. Some sort of estimate of reliability is fundamentally important in order to enhance interpretation of assessment measures.

Applicability of the Present Evaluation Tool Beyond Reliability Assessment

The discussion here has addressed reliability only and has not considered the related topic of validity of transcription. Intertranscriber agreement measures will be affected inevitably by any phonetic bias of listeners, and any bias can be thought of as a challenge to the validity of data acquired in phonetic transcription. Related to bias is the matter of generalizability of results obtained in transcription. Even when transcription reliability is high between two coders, it does not necessarily mean that the reliability would generalize to other coders. It could be instead an indication that the two coders did not make perceptual judgments independently or that the two simply shared (because they were accidentally similar listeners) phonetic perceptions a significant portion of the time. The coders could have been biased in similar ways, and if so, their results would not generalize to coders with different biases. The uncertainty in generalization of transcription reliability assessment values across listeners is suggested by research indicating that dialect differences among transcribers affect their perceptions and consequent reliability (Coussé, Gillis, Kloots, & Swerts, 2004). In an ideal world, transcriptions would represent the varying perceptions of the gamut of potential listeners, in order to accurately reflect the total error variance of the transcription measure. Again, practicality enters the picture. No one can afford regularly to sample at random from the population of potential listeners in transcription-based studies of vocalization. However, it is sensible to consider transcriber selection from the standpoint of generalizability and to interpret results in a way that takes transcriber selection into account.

The measure developed here does not directly contribute to validity of transcription. However, it may contribute to more appropriate and more convenient reliability assessment. With improved tools for reliability assessment, we should at least be in a better position to research the role of generalizability of agreement results (and the coder biases that are implied by lack of generalizability) in transcription validity.

Beyond the goals of agreement research, it should be noted that the computer-based analysis developed here is applicable to additional circumstances. In fact, it constitutes a general weighted scheme of comparison between any two transcripts, and it yields a phonetic or phonological similarity score expressed as a proportion of matching elements. Because one transcript in the analysis can represent a well-formed target, and the other a pronunciation by a learner or a disordered speaker, one of the applications of this approach is to supply a general weighted measure of the degree to which learners or disordered speakers successfully pronounce words or sentences. It has thus already been treated as a weighted *proportion of phonetic elements correct* (PPEC) score (Ramsdell & Oller, 2005). The PPEC score should offer a more subtle and revealing measure than traditional *percentage of consonants correct* or other measures that are generally implemented with an absolute match criterion (Shriberg & Kwiatkowski, 1980; Smith, 1975). The PPEC score might offer a phonologically general supplement to clinical assessment tools that seek overall measures of pronunciation adequacy (see, e.g., Hodson, 1980). Furthermore, if the improvements in scaling properties that are found by substituting a weighted measure for an unweighted one meet their promise, they should provide a much more stable basis for correlational research on the relation between degree of correctness in pronunciation as assessed by transcription and auditory intelligibility, a topic that has been the subject of research for several generations (Hudgins & Numbers, 1942; Steffens, 1994). Additional applications are easy to imagine in dialectology, second language acquisition, and accent assessment.

Furthermore, a weighted assessment approach for reliability of phonetic transcription has yet another potential utility that has been considered informally for years within a number of laboratories. The lower the transcription reliability for a sample, the less well formed the sample is presumed to be. Consequently, reliability measures themselves have been treated as indicators of well-formedness. If this interpretation is valid, it offers an answer to the question: What is the use of data produced by measures with very low reliability? In fact, the whole range of reliability across the weighted scale might be thought to correspond to a range of speechlike quality and potential intelligibility. The interpretation clearly requires empirical support that has not yet been provided. As far as we know, there is no published report to date providing systematic empirical support for the widespread belief that well-formedness (or canonicity) is directly related to transcription agreement. With the weighted agreement measure in place, it would appear that the time is ripe to test that belief.

Acknowledgment

This work was supported by the National Institute on Deafness and Other Communication Disorders (R01DC006099; D. Kimbrough Oller, principal investigator, and Eugene Buder, co-principal investigator) and by the Plough Foundation.

References

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research and Evaluation, 9*. Retrieved January 31, 2006, from http://PAREonline.net/getvn.asp?v=9&n=3.

Bakkum, M. J., & Plomp, R. (1995). Objective analysis versus subjective assessment of vowels pronounced by deaf and normal-hearing children. *Journal of the Acoustical Society of America, 98,* 745–761.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.

Cleary, T. A., Linn, R. L., & Walster, G. W. (1970). Effect of reliability and validity on power of statistical tests. *Sociological Methodology, 2,* 130–138.

Clements, G. N., & Keyser, S. J. (1983). *CV phonology*. Cambridge, MA: MIT Press.

Coussé, E., Gillis, S., Kloots, H., & Swerts, M. (2004). The influence of the labeller's regional background on phonetic transcriptions: Implications for the evaluation of spoken language resources. In M. Lino, M. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (Vol. IV, pp. 1447–1450). Paris: ELRA.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics, 10,* 131–155.

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

Davis, B. L., Morrison, H. M., von Hapsburg, D., & Warner Czyz, A. D. (2005). Early vocal patterns in infants with varied hearing levels. *Volta Review, 105,* 7–27.

Delack, J. (1976). Aspects of infant speech development in the first year of life. *Canadian Journal of Linguistics, 21,* 17–37.

Eilers, R. E., Oller, D. K., & Benito Garcia, C. R. (1984). The acquisition of voicing contrasts in Spanish and English learning infants and children: A longitudinal study. *Journal of Child Language, 11,* 313–336.

Firth, J. R. (1957). *Papers in linguistics: 1934–1951*. Toronto, Ontario, Canada: Oxford University Press.

Goldsmith, J. (1976). An overview of autosegmental phonology. *Linguistic Analysis, 2,* 23–68.

Goldsmith, J. (1995). *The handbook of phonological theory*. Cambridge, MA: Blackwell.

Greenberg, J. H. (1966). *Language universals*. The Hague, The Netherlands: Mouton.

Gussenhoven, C., & Jacobs, H. (1998). *Understanding phonology*. London: Arnold.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Hodson, B. W. (1980). *Assessment of phonological processes*. Danville, IL: Interstate Press.

Hudgins, C. V., & Numbers, F. C. (1942). An investigation of intelligibility of speech of the deaf. *Genetic Psychology Monograph, 25,* 289–392.

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

Irwin, O. C., & Chen, H. P. (1941). A reliability study of speech sounds observed in the crying of newborn infants. *Child Development, 12,* 351–368.

Irwin, O. C., & Curry, T. (1941). Vowel elements in the crying vocalization of infants under ten days of age. *Child Development, 12,* 99–109.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Kent, R. D., Weismer, G., Kent, J., & Rosenbek, J. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54,* 482–499.

Koopmans-van Beinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom & R. Zetterstrom (Eds.), *Precursors of early speech* (pp. 37–50). New York: Stockton Press.

Ladefoged, P. (2001). *A course in phonetics*. Boston: Heinle and Heinle.

Legerstee, M. (1991). Changes in the quality of infant sounds as a function of social and nonsocial stimulation. *First Language, 11,* 327–343.

Lewedag, V. L. (1995). *Patterns of onset of canonical babbling among typically developing infants*. Unpublished doctoral dissertation, University of Miami.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8,* 249–336.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development* (pp. 131–164). Timonium, MD: York Press.

Lindblom, B. E. F., & Maddieson, I. (1988). Phonetic universals in consonant systems. In L. M. Hyman & C. N. Li (Eds.), *Language, speech, and mind* (pp. 62–78). New York: Routledge.

Louko, L. J., & Edwards, M. L. (2001). Issues in collecting and transcribing speech samples. *Topics in Language Disorders, 21,* 1–11.

McCarthy, J. J. (1985). *Formal problems in Semitic phonology and morphology*. New York: Garland.

Miller, G. A., & Nicely, P. E. (1955). The analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27,* 338–352.

Nathani, S., & Oller, D. K. (2001). Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, & Computers, 33,* 321–330.

Nicewander, W. A., & Price, J. M. (1983). Reliability of measurement and the power of statistical tests: Some new results. *Psychological Bulletin, 94,* 524–533.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Vol. 1. Production* (pp. 93–112). New York: Academic Press.

Oller, D. K. (1992). Description of infant vocalizations and young child speech: Theoretical and practical tools. *Seminars in Speech and Language, 13,* 178–193.

Oller, D. K. (1995). Development of vocalizations in infancy. In H. Winitz (Ed.), *Human communication and its disorders: A review* (Vol. IV, pp. 1–30). Timonium, MD: York Press.

Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Erlbaum.

Oller, D. K., & Bull, D. (1984, April). *Vocalizations of deaf infants*. Poster presented at the International Conference on Infant Studies, New York.

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

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

Oller, D. K., & Eilers, R. (1981). A pragmatic approach to phonological systems of deaf speakers. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 6, pp. 103–141). New York: Academic Press.

Oller, D. K., & Eilers, R. E. (1982). Similarity of babbling in Spanish- and English-learning babies. *Journal of Child Language, 9,* 565–578.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59,* 441–449.

Oller, D. K., Eilers, R. E., Steffens, M. L., Lynch, M. P., & Urbano, R. (1994). Speech-like vocalizations in infancy: An evaluation of potential risk factors. *Journal of Child Language, 21,* 33–58.

Oller, D. K., & Steffens, M. L. (1994). Syllables and segments in infant vocalizations and young child speech. In M. Yavas (Ed.), *First and second language phonology* (pp. 45–61). San Diego, CA: Singular.

Oller, D. K., Wieman, L., Doyle, W., & Ross, C. (1975). Infant babbling and speech. *Journal of Child Language, 3,* 1–11.

Pollock, K. E., & Hall, P. K. (1991). An analysis of the vowel misarticulations of five children with developmental apraxia of speech. *Clinical Linguistics and Phonetics, 5,* 207–224.

Prince, A., & Smolensky, P. (1993). *Optimality theory: Constraint interaction in generative grammar*. New Brunswick, NJ: Rutgers University Press.

Ramsdell, H., & Oller, D. K. (2005, June). *Reliability in transcription and coding of infant vocalizations*. Presentation to the International Child Phonology Conference, Fort Worth, TX.

Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997). Developmental apraxia of speech: I. Descriptive and theoretical perspectives. *Journal of Speech, Language, and Hearing Research, 40,* 273–285.

Shriberg, L. D., & Kwiatkowski, J. (1980). *Natural process analysis: A procedure for phonological analysis of continuous speech samples*. New York: MacMillan.

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

Smith, C. (1975). Residual hearing and speech production of deaf children. *Journal of Speech and Hearing Research, 18,* 795–811.

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. Y. Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology* (Vol. 1, pp. 73–90). New York: Academic Press.

Steffens, M. L. (1994). *Assessment of phonological development utilizing multidimensional measures of phonological* *adequacy*. Unpublished doctoral dissertation, University of Miami.

Stockman, I. J., Woods, D. R., & Tishman, A. (1981). Listener agreement on phonetic segments in early infant vocalizations. *Journal of Psycholinguistic Research, 10,* 593–617.

Stoel-Gammon, C. (1992). Prelinguistic vocal development: Measurement and prediction. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 439–456). Parkton, MD: York Press.

Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in Language Disorders, 21,* 12–21.

Stoel-Gammon, C., & Cooper, J. (1984). Patterns of early lexical and phonological development. *Journal of Child language, 11,* 247–271.

Stoel-Gammon, C., & Herrington, P. B. (1990). Vowel systems of normally developing and phonologically disordered children. *Clinical Linguistics and Phonetics, 4,* 145–160.

Stoel-Gammon, C., & Otomo, K. (1986). Babbling development of hearing impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders, 51,* 33–41.

Sutcliffe, J. P. (1958). Error of measurement and the sensitivity of a test of significance. *Psychometrika, 13,* 9–17.

Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie*. Prague, Czech Republic: Cercle linguistique de Prague.

van der Stelt, J. M. (1993). *Finally a word: A sensori-motor approach of the mother–infant system in its development towards speech*. Amsterdam, The Netherlands: Uitgave IFOTT.

Vihman, M. M. (1986). Individual differences in babbling and early speech. In B. Lindblom & R. Zetterstrom (Eds.), *Precursors of early speech* (pp. 21–35). New York: Stockton Press.

Vihman, M. M., Macken, M., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A reassessment of the continuity issue. *Language, 61,* 397–445.

Warren, R. M., & Warren, R. P. (1966). A comparison of speech perception in childhood, maturity and old age by means of the verbal transformation effect. *Journal of Verbal Learning and Verbal Behavior, 5,* 142–146.

Zimmerman, D. W., Williams, R. H., & Zumbow, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17,* 1–17.

Received September 28, 2005

Accepted April 28, 2006

DOI: 10.1044/1092-4388(2006/100)

Contact author: D. Kimbrough Oller, School of Audiology and Speech-Language Pathology, The University of Memphis, 807 Jefferson Avenue, Memphis, TN 38105.
E-mail: koller@memphis.edu.

Appendix. Toddler and adult word list (English).

pencil
tea
neighborhood
dog
number
ant
television
cow
paper
loud
computer
owl
rain
bed
history
doctor
soda
science
mailman
water