


Longitudinal observations of typical English voicing acquisition in a 2-year-old child: Stability of the contrast and considerations for clinical assessment

Elaine R. Hitchcock & Laura L. Koenig

To cite this article: Elaine R. Hitchcock & Laura L. Koenig (2015): Longitudinal observations of typical English voicing acquisition in a 2-year-old child: Stability of the contrast and considerations for clinical assessment, Clinical Linguistics & Phonetics

To link to this article: <http://dx.doi.org/10.3109/02699206.2015.1083617>

 View supplementary material 

 Published online: 29 Oct 2015.

 Submit your article to this journal 

 View related articles 

 View Crossmark data 

Longitudinal observations of typical English voicing acquisition in a 2-year-old child: Stability of the contrast and considerations for clinical assessment

1896

ELAINE R. HITCHCOCK¹ & LAURA L. KOENIG^{2,3}

¹*Department of Communication Sciences and Disorders, Montclair State University, Bloomfield, NJ, USA,* ²*Haskins Laboratories, New Haven, CT, USA,* and ³*Long Island University, Brooklyn, NY, USA*

(Received 16 April 2015; revised 5 August 2015; accepted 11 August 2015)

Abstract

Early assessment of phonetic and phonological development requires knowledge of typical versus atypical speech patterns, as well as the range of individual developmental trajectories. The nature of data reporting in previous literature on typical voicing acquisition left aspects of the developmental process unclear and limited clinical applicability. This work extends a previous four-month group study to present data for one child over 12 months. Words containing initial /b p d t/ were elicited from a monolingual English-speaking 2-year-old child biweekly for 25 sessions. Voice onset time (VOT) was measured for each stop. For each consonant and recording session, we measured range as well as accuracy, overshoot and discreteness calculated for means and individual tokens. The results underscore the value of token-by-token analyses. They further reveal that typical development may involve an extended period of fluctuating voicing patterns, suggesting that the voiced/voiceless contrast may take months or years to stabilise.

Keywords: Assessment, English, speech development, voice onset time, voicing

Introduction

Valid assessment of speech sound disorders in children requires a clear understanding of the time course of typical phonetic and phonological development, and the variety of paths to acquisition that individual children may take. With increased recognition that early intervention may prevent or mitigate later speech and language difficulties (e.g. ASHA, 2008), clinicians must determine when existing data on early speech development are sufficient to determine that a young child's speech production differs from that of typically-developing peers.

This study focuses on the English stop voicing contrast. As reviewed in the next section, many authors have carried out instrumental evaluations of stop voicing acquisition in English and other languages, most typically using voice onset time (VOT; Lisker & Abramson, 1964). Nevertheless, a review of that literature led us to the conclusion that our understanding of consonant voicing accuracy in children between 2 and 3 years of age was insufficient to allow for confident clinical

assessment of voicing in this important age range, when extensive phonological development occurs (Stoel-Gammon & Dunn, 1985). Specifically, much past work has presented VOT data in highly reduced form (e.g. means and standard deviations). Results based entirely on averaged data are difficult to use in clinical application, since speech–language pathologists typically use a percent of correct productions to support the identification of a speech sound disorder and determine a treatment plan (e.g. the percentage of consonants correct [PCC] measure; Shriberg, Austin, Lewis, McSweeney, & Wilson, 1997). To see the potential issues here, consider this hypothetical example. A clinician elicits a number of monosyllabic words from a child, attempting to sample across all English syllable-initial consonants. Suppose that this sample contains four words with initial /p/, and their VOT values are as follows: 14, 12, 10 and 130 ms.¹ The first three values correspond to typical VOTs for English /b/; only the last corresponds to an aspirated stop as expected for syllable-initial English /p/. An analysis of individual productions (token-by-token) would find three of these four productions to be inaccurate productions of /p/, yet the mean of the four productions, 41.5 ms, is within the range of English aspirated /p/.

These realisations about the limitations of reduced data led to a longitudinal study of 10 typical children (Hitchcock & Koenig, 2013), and the single-speaker, year-long study reported here. The current paper presents longitudinal data on the voicing behaviour of one typically-developing child from 2;4 to 3;4, giving particular attention to how stable her voicing categories were over time. As before, one component of the analysis assessed production accuracy, defined with reference to adult labelling behaviour, i.e. perceptual judgements. This measure is of interest because it yields insight into how a child’s speech would be classified in a typical clinical evaluation. Moreover, the classic staging of VOT development (see next section) was based, in part, on how an adult is likely to perceive a child’s productions.

Studies of typical voicing using VOT and stages of voicing acquisition

Previous work established the following developmental sequence for English (Macken & Barton, 1980; cf. also Kewley-Port & Preston, 1974; Scobbie, Gibbon, Hardcastle, & Fletcher, 2000; Zlatin & Koenigsnecht, 1976):

- Stage I. VOT values are unimodal, in the short-lag range (appropriate for /b d g/) for all stop consonants.
- Stage II. A statistically significant VOT contrast develops. It may be subperceptual (or “covert”), within the short-lag range.
- Stage IIIa. Some long-lag or /p t/ VOT values appear; they show exaggerated VOT lag (“overshoot”).
- Stage IIIb. Long-lag VOTs shorten to more adult-like values over time. Clearly separated short- and long-lag VOT distributions are the last stage of development (“discreteness”).

Despite general consensus on the staging outlined above, the timescale of the English voicing acquisition process and the variety of developmental schedules across children have remained somewhat unclear (cf. Hitchcock & Koenig, 2013, for fuller discussion). Part of the difficulty is that past studies have varied widely in sampling method (longitudinal versus cross-sectional), sample size, age range, elicitation method, number and frequency of data collection sessions and statistical measures. For example, one of the most widely-cited longitudinal studies, Macken and Barton (1980), collapsed data across multiple recording sessions for reporting purposes,

¹Although hypothetical, this example is not outside the realm of possibility; two of the current participant’s sessions (Session 1 for alveolars and 3 for bilabials; see Supplemental Material) showed data clustered in the short-lag range with one very long-lag outlier; such cases can also be observed in Macken and Barton’s (1980) data (pp. 56 and 64).

potentially obscuring short-term changes. Also, comparing across studies, it was difficult to determine whether an overshoot phase (Stage IIIa) was universal to all children learning a language that uses voiceless aspirated stops. Thus, to gain a clearer and more clinically-applicable picture of how typical 2-year-olds learn to produce VOT contrasts, Hitchcock and Koenig (2013) collected VOT data on 10 children recorded biweekly for four months. Two goals of that work were to explore how young children's productions continue to differ from adults' after a statistically significant VOT contrast is observed, and to evaluate how the evident VOT staging varied depending on analysis method. VOT was analysed using token-by-token measures and summary statistics (means and ANOVA results), using four criterion measures, described schematically here and in more quantitative detail in the methods section below:

(a) Range of VOT values per target consonant. This measure was by nature restricted to the token-by-token analysis; it is simply the difference, in ms, between the highest and lowest VOT value for each consonant in a given recording session. Range was included, following Zlatin and Koenigsnecht (1976), as an index of the token-to-token variability that characterises child speech (e.g. Eguchi & Hirsh, 1969 and many subsequent studies).

(b) Accuracy, defined as VOT values appropriate to the target based on adult production and perception data, for means and individual productions (token-by-token). This measure was implemented following Kewley-Port and Preston (1974) and Macken and Barton (1980). Those authors proposed that adults would categorise (i.e. label) children's productions based on adult values of production and perception.² Note that Macken and Barton's Stage II was designated a "subperceptual"³ contrast based on the logic that values in the short-lag range would be perceived as phonemically identical. Moreover, as pointed out above, clinical assessment of children's speech primarily relies on perceptual evaluation by adult listeners. Macken and Barton (1980, p. 48) defined Stage II in terms of VOT means; we also carried out token-by-token measures of accuracy to assess how this would affect VOT staging.

(c) Discreteness, or degree of overlap between the voiced/voiceless categories, for means and individual productions (token-by-token). Discreteness measures for child speech were reported following Zlatin and Koenigsnecht (1976), who evaluated all tokens in measuring discreteness, and Bailey and Haggard (1973, 1980), who reported an analogous "distinctiveness" measure for means. Bharadwaj and Graves (2008) have suggested that more discrete categories (i.e. those with little or no overlap) may be associated with listener judgements of higher contrast typicality in the speech of children with cochlear implants. Further, Romeo, Hazan, and Pettinato (2013) recently reported that acoustic measures of consonants (VOTs for /b p/ and centroids for /s ʃ/) showed overlapping values in children, especially boys, more frequently than adults, supporting the general conclusion that measures of consonant distinctiveness provide a window into contrast development.⁴

²One may ask whether adult listeners have comparable VOT perception boundaries for child speech and adult speech. We know of no published studies addressing this question, but we have conducted one pilot study of two 2-year-olds (one being the participant of this study) and a subsequent, more extensive study with a larger number of children. Results of the pilot study, using four listeners, yielded 50% cross-overs (i.e. phoneme boundaries) of 25 ms for bilabials and 35 ms for alveolars, similar to what has been observed for synthetic stimuli based on adult models (e.g. Kuhl & Miller, 1978). The follow-up study (Hitchcock & Koenig, 2015) yields comparable results. Thus, the traditional practice of using adult VOT values to infer adult perception of a child's voicing categories appears to be valid.

³Macken and Barton specifically state that, in this stage, "the contrast that the child is making is presumably not perceptible to adults" (1980, p. 48).

⁴Romeo et al. (2013) did observe small amounts of overlap, i.e. lack of discreteness, in some adults, and also reported that category separation was greater in women than men. Thus, factors like speaking style can also come into play here and yield variation among adult speakers. We do not claim that category overlap never occurs in typical adult speech; our point is simply that extensive and persistent overlap is not a feature we would associate with mature speech. Along similar lines, adult productions, particularly in rapid or connected speech, could occasionally be produced with inaccurate VOTs, i.e. values appropriate to the opposite category (cf. Lisker & Abramson, 1967), but we expect such tokens to be rare on the whole.

(d) Overshoot, or exaggerated aspiration (very long positive VOTs), for means and individual productions (token-by-token) of /p t/. This measure was intended to capture the occurrence of Macken and Barton's "overshoot" phase (Stage IIIa).

Findings of Hitchcock and Koenig (2013) included the following: (a) Children's stop voicing characteristics could show extreme variations between adjacent recording sessions, i.e. over two-week intervals. (b) The timescale of development varied across children; some showed rapid, abrupt gains whereas others made more gradual progress over time. (c) Contrasts evidently acquired at one recording session were not always maintained at the next. (d) Measures of individual productions (i.e. the token-by-token analysis) presented a more nuanced picture of development than the summary statistics (mean values and ANOVA results). Hitchcock and Koenig observed that the token-by-token analyses revealed patterns (e.g. an overshoot phase for all children) not evident in the means. These results suggest that a battery of measures carried out on a token-by-token basis, is more sensitive than analyses based on session means in showing changes in a child's VOT contrast over time. Further, the token-by-token analysis provides a firmer foundation from which one can assess how children's productions continue to differ from adults' after a statistically significant VOT contrast is achieved. Thus, the staging of Macken and Barton (1980), outlined above, was revised to include the four measured parameters (Table 1).

This paper extends the work of Hitchcock and Koenig (2013) by following one child for a full year of biweekly recording sessions, providing a unique window into the stability of contrasts in early phonological acquisition. Past longitudinal studies either did not record the children for this long or this frequently, or if they did they grouped the data for analysis purposes, thus potentially masking short-term variation in the voicing contrast. Our elicitation methods were designed to ensure that each consonant was represented by a sufficient number of tokens (15+) to establish valid distributional characteristics for all sessions. As before, we present data on VOT range, as well as accuracy, overshoot and discreteness, on a token-by-token basis and for session means. We give particular attention to VOT accuracy and discreteness, and their stability over time. Once a child is observed to demonstrate a VOT contrast, do the categories remain accurate and discrete, or is regression observed (e.g. in the form of greater overlap or reduced accuracy)? To our knowledge, no longitudinal studies have investigated this question.

Case study

Methods

The participant inclusion/exclusion criteria, recording protocol and analyses were identical to those used in Hitchcock and Koenig (2013). The subject of this study, KC, was first recorded biweekly for four months (as were nine other children). Her biweekly recording schedule was then extended for eight more months following the same protocol. At the beginning of the original four-month study, KC demonstrated no VOT contrast at either place of articulation (according to the accuracy criteria defined in the next section). She showed discrete and largely accurate categories at the end of that study, suggesting that she had rapidly achieved the voicing contrast. The current work investigates how that apparently established contrast was maintained over time. In order to show development over the full 12-month duration, the results section here includes all data points from the previous study along with the newly-collected data. The previously-reported (Hitchcock & Koenig, 2013) data are indicated in the plots by showing the first eight sessions with a grey background.

Table 1. Revised stages of English voicing acquisition, following Macken and Barton (1980), adding the four additional measures suggested in the text (Hirchcock & Koenig, 2013).

Stage	Approximate age range (years; months)	VOT values	Range size	Accuracy	Discretteness	Overshoot
I	Typically <2;0	Short-lag	Small (≤ 100 ms)	/b d/ good; /p t/ bad	Low: Extensive overlap between /b p/ and /d t/ pairs Low: Extensive overlap	Absent/minimal
II	1;6–2;6	Primarily short lag; expanding into long-lag range	Increased from Stage I	/b d/ good; /p t/ fair	Low: Extensive overlap	Absent/minimal
IIIa	2;0–3;0	VOT values in the short-lag and long-lag range	Increased from Stage II	/b d/ either stable or moving slightly in the long lag range; /p t/ good	Moderate–high: overlap still present but decreased from Stages I to II High: No overlap present between the categories; separation between voicing categories is ≥ 30 ms as seen in adults	Present for /p/ and/or /t/
IIIb	2;6–3+	VOT values in the short-lag and long-lag range	Decreased from Stage IIIa but longer than adults	/b d/ good; /p t/ good		Absent

Participant

KC was a monolingual, American-English-speaking child who was 28 months old at the time of her initial enrollment, and 32 months at the beginning of the study extension. At the conclusion of the first study, all parents were asked if they wanted to continue their child's participation; only KC's mother responded to our request. KC's mother provided informed consent for her child to enroll in the study. Following the selection criteria of the original study, KC's general health and birth history were unremarkable per parent questionnaire. She passed a free-field pure-tone hearing screening conducted in a soundproof audiometric suite (20 dB thresholds or less at 250, 500, 1000, 2000 and 4000 Hz), demonstrated no gross structural or functional abnormality in the oral mechanism and scored within one standard deviation of the age level mean on standardised tests of speech and language skills. Speech production was measured by the *Goldman–Fristoe Test of Articulation-2* (Goldman & Fristoe, 2000). KC received a standard score of 89 on the GFTA-2, which placed her in the 37th percentile. Language functioning was measured by the *Preschool Language Scale-3* (Zimmerman, Steiner, & Pond, 1992).⁵ KC received a standard score of 111, which placed her in the 77th percentile. All assessments were performed by the first author, a certified speech–language pathologist. Finally, the presence/absence of a syllable-initial stop voicing contrast was assessed prior to the study by recording a minimum of 10 single-word utterances per target consonant (/b d p t/) over two sessions. KC demonstrated less than 25% acoustic accuracy (defined in the Measurements and Analysis section) for voiceless target consonants, indicating that she did not have a perceptual or subperceptual VOT contrast at either place of articulation and was therefore in Macken and Barton's (1980) Stage I.

Recording and stimulus materials

Recordings were made in a quiet room in the child's home using a Sony (MZ-N707, Tokyo, Japan) portable minidisk recorder and a sampling rate of 44.1 kHz. A Kay Pentax (Montvale, NJ) lapel microphone (Model #3502) was attached to the child's clothing approximately six inches from the mouth for a favourable signal to noise ratio and minimal feedback distortion.

A minimum of 15 tokens of each bilabial and alveolar target in initial/prevocalic position was attempted in each recording session to establish VOT distributional characteristics for all stops. Single words were targeted as being age-appropriate, and to facilitate accurate production (Hitchcock & Koenig, 2013; cf. also Morrison & Shriberg, 1992). Stimuli were black and white drawings of two minimal pairs of CV words, *boo-poo* and *doe-toe*. The velar place of articulation was excluded for two reasons. First, we wanted to constrain the stimulus set because recording sessions for children of this age need to be kept short, and multiple repetitions of each consonant were needed to establish distributional characteristics. Further, the velars tend to be acquired later than alveolars and bilabials (e.g. Sander, 1972), which can lead to markedly smaller sample sizes for this place compared to the others. In fact, KC showed some velar → alveolar substitutions in the initial standardised and unstandardised assessments; tokens without the correct place of articulation were excluded in our analysis.

Caregivers were given the stimulus pictures a week before the initial recording session in order to familiarise KC with the pictures and the word targets. (Once recording sessions began, the stimulus pictures were removed from the house and parents were asked not to elicit the study words from their child between sessions.) The picture prompts were presented in random order using verbal cues to elicit the target response (e.g. "Winnie the ____" or "The ghost says ____"). A model was given if KC did not respond to the verbal/picture stimuli. Responses

⁵At the time of data collection, the *Preschool Language Scale-3* was the most current version of this standardised assessment.

provided in direct imitation of the model were discarded given that imitation has been shown to affect production accuracy (Hodson & Paden, 1991). KC typically produced the target words in isolation. Repeated productions on a single breath (e.g. *boo, boo, boo, boo*) were excluded from analysis.

Measurements and analysis

Data were transferred to the Kay Pentax Computerized Speech Lab (Model 4500, Montvale, NJ) for analysis, with a sampling rate of 11 025 Hz. Inclusion criteria for productions were as follows: CV word shape; the initial C was a stop with the appropriate place of articulation, produced in response to a known picture target; and the token was not acoustically distorted by ambient noise, yelling or singing. VOT, defined as the interval between the release of vocal tract occlusion and the onset of glottal vibration, was measured using an acoustic waveform and wideband spectrogram. The final data set for KC, including the first four months of data collection, consisted of 2246 tokens (600 from the previous study). This corresponds to an average of 22.46 productions of each of the four consonants over 25 recording sessions, although later sessions averaged slightly higher due to additional spontaneous repetitions during data collection. The percentage of tokens discarded due to noise/distortion for this participant was 7.7%.

The four measures of assessment were calculated as described below. For reference, VOT histograms are provided for both places of articulation and all sessions in the Supplemental Material, along with the means and standard deviations for all sessions. Readers are also referred to Hitchcock & Koenig (2013) for more extensive justification and examples.

(a) Range. The lowest VOT value subtracted from the highest VOT value for each consonant target. For example, in session 21 (cf. Supplemental Materials), KC's VOTs for /b/ ranged from -47 to 26 ms, yielding 73 ms as her range value for that sound and session.

(b) Accuracy. This was based on thresholds of 20 ms for bilabials and 30 ms for alveolars, following the adult data of Lisker and Abramson (1964) and Zlatin and Koenigsknecht (1976). The different boundaries for bilabials and alveolars reflect the fact that VOT values tend to increase with more posterior places of articulation (e.g. Lisker & Abramson, 1964), and the perceptual boundaries vary accordingly (e.g. Kuhl & Miller, 1978). These thresholds were applied to mean VOT values for each session, and also to each production in a session for the token-by-token analysis. For example, in session 2 (see Supplemental Materials), KC had mean VOTs of 16 ms for /b/ and 20 ms for /p/; following the 20 ms bilabial threshold /b/ was considered accurate (<20 ms), whereas /p/ was inaccurate (not above the threshold value, i.e. too short). The token-by-token accuracy measure designated each production as accurate or inaccurate based on the same thresholds; we then calculated the percentages of each consonant in that session that were accurate. The accuracy criterion did not exclude exaggerated but "correct" productions; that is, long prevoicing for /b/ or long aspiration for /p/ did not render a production inaccurate.

(c) Discreteness. The difference, in ms, between voiced and voiceless targets at each place of articulation. For the means analysis, this was simply the difference between the session means, following Bailey and Haggard (1973). For the token-by-token analysis, this was measured as the difference between the highest voiced target VOT and the lowest voiceless target VOT, following Zlatin and Koenigsknecht (1976). For example, in session 13, KC's lowest VOT value for /p/ was 16 ms, whereas her highest VOT value for /b/ was 19 ms – an overlap of 3 ms in the token-by-token analysis. On the other hand, in session 12, her lowest /p/ VOT was 54 ms, and her highest /b/ VOT was 29 ms, yielding no overlap. Note that the discreteness measure was calculated based on the data from individual sessions, unlike the calculation of accuracy, which was based on fixed values for all sessions. Negative discreteness values indicate overlap between voicing categories, whereas positive values indicate separation between categories. A separation of circa 30 ms between

contrasting VOT categories has been reported for American-English speaking adults (Zlatin, 1974, but cf. footnote 4). A lack of discreteness should correspond to perceived ambiguity between the voicing categories in a given session.

(d) *Overshoot*. This quantified the occurrence of very long-lag values (>100 ms) for /p/ and /t/, and was included to permit evaluation of whether the child went through Macken and Barton's (1980) Stage IIIa. Such long VOT values are not typical in most adults, even in citation speech (Lisker & Abramson, 1964; Zlatin, 1972). Overshoot was obtained for session mean values and as a percentage of all productions of /p t/ in a recording session in a manner analogous to that used for accuracy. Specifically, means of /p t/ greater than 100 ms were designated as showing overshoot. For the token-by-token analysis, each production was labeled as showing overshoot or not, and overshoot was then expressed as a percentage of /p t/ productions exceeding 100 ms in each session. When 50% of the distribution was >100 ms, the participant was said to be in an "overshoot phase". Notice that overshoot is not redundant with range, because range values were based on the lowest as well as the highest value for a particular phoneme in a given session. Overshoot, in contrast, quantified target voiceless sounds with very long positive VOTs only.

Measurement reliability

All tokens were measured either by the first author or a trained graduate student. A second trained researcher measured randomly selected tokens across all recording sessions. Measurements were considered in agreement if they were within 5 ms of each other (exceeding the 10 ms value recently suggested by Fabiano-Smith & Bunta, 2012). Inter-rater reliability was based on a comparison of 18% of the tokens (409/2246). The measures showed 92% agreement with a mean difference of 2.2 ms and a standard deviation of 5 ms; further, VOT values for the original and re-measured tokens were highly correlated ($r = 0.987$, $p < 0.0001$).

Statistical analysis

Many researchers have used the presence of a statistically significant VOT difference between the voicing categories as an indication that a child has acquired a voicing contrast (Bailey & Haggard, 1980; Lowenstein & Nittrouer, 2008; Macken & Barton, 1980; Snow, 1997; Tyler & Saxman, 1991; Zlatin & Koenigskecht, 1976). Thus, within-subject repeated measures analyses of variance (ANOVAs) were applied to the /b p/ and /d t/ data for each session to establish the presence or absence of a statistically significant difference between the voicing categories and allow for a comparison of statistically significant contrasts with the four measures of assessment listed above. Given the large number of ANOVAs (two places of articulation \times 25 sessions), significance was assessed not only with a standard α -level of 0.05 but also with a conservative α of 0.001. Cohen's d was also computed for each /b p/ and /t d/ pair to provide an estimate of effect size (two places of articulation \times 25 sessions = 50 values). Following Cohen (1988), effect sizes of $d = 0.2$ were considered small, $d = 0.5$ were considered medium and $d = 0.8$ were considered large.

Results

Statistical significance

Repeated measures ANOVAs revealed a statistically significant difference in the expected direction for 84% of both /b p/ and /d t/ pairs across the 25 recording sessions (21 of 25 sessions

Table 2. Results (*p* values) of repeated measures ANOVAs for KC per session.

Session #	/b p/	/d t/
1	0.1321	*
2	0.2514	0.1226
3	0.9506	0.7168
4	0.3417	0.7163
5	*	0.5453
6	**	*
7	**	**
8	**	**
9	**	**
10	**	**
11	**	**
12	**	**
13	**	**
14	**	**
15	**	**
16	**	**
17	**	**
18	**	**
19	**	**
20	**	**
21	**	**
22	**	**
23	**	**
24	**	**
25	**	**

For each place of articulation, VOT values for all voiced and voiceless tokens were entered into the ANOVA to evaluate whether KC had a significant VOT contrast in that session for /b p/ and /d t/. Asterisks indicate significance using a standard α -level of 0.05 (*) and a conservative α of 0.001 (**). The conservative α represents a correction for the large number of ANOVAs (1 child \times 25 sessions \times 2 places of articulation).

for each pair, i.e. a total of 42 of 50 comparisons) using a criterion of $p < 0.05$. Within this group, 93% (39 of the 42) met the more stringent significance criterion of $p < 0.001$ (Table 2). These results indicate that VOT values were generally lower for /b/ versus /p/ and /d/ versus /t/. Session effect sizes for /b p/ and /d t/ were usually large (41 of 50 sessions); one was medium and eight were small. All effect sizes were large in sessions 5–25 for /b p/ and in sessions 6–25 for /d t/.

These statistical results could be taken to suggest that KC had acquired the voicing contrast and that a perceptual evaluation would show high accuracy for this phonological feature. However, the analyses reported below demonstrate KC's continuing process of refining the /b p/ and /d t/ contrasts, and the ongoing presence of ambiguous/inaccurate productions and category overlap. As in our earlier study, the token-by-token analyses ultimately presented a considerably richer picture of voicing development than the analysis of means; they are also more straightforwardly applicable for purposes of clinical assessment. As such, the former will be given primary attention in the rest of the results presentation.

Range

Range sizes for each phoneme are shown in Figure 1. Adult range sizes for English voiced and voiceless stops have been reported to be up to 180 ms (including prevoiced tokens) and 120 ms,

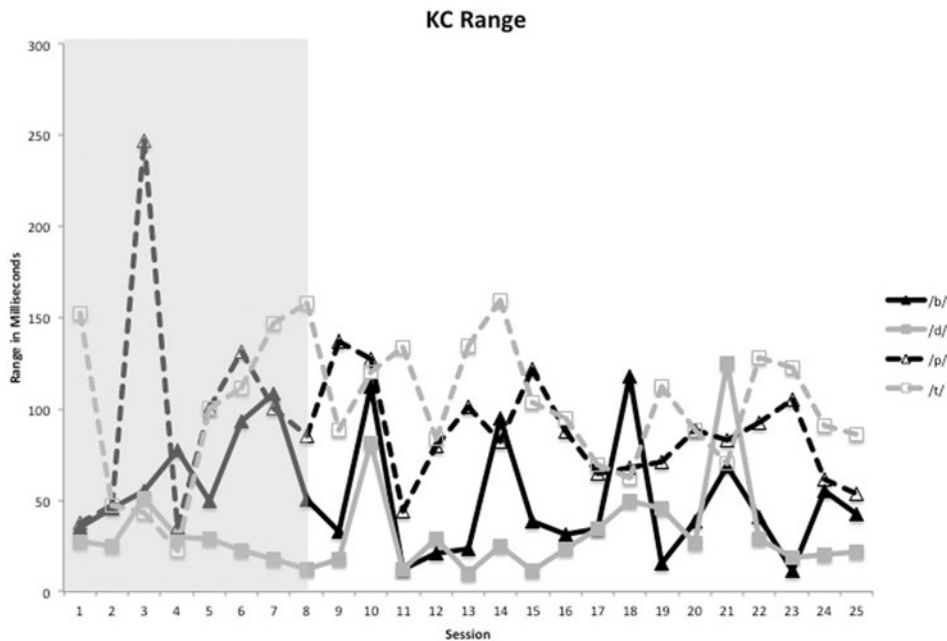


Figure 1. Range in ms of VOT values for /b/, /d/, /p/, /t/ for KC over 25 sessions. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

respectively (Lisker & Abramson, 1964; Zlatin & Koenigsnecht, 1976). Although ranges for all four phonemes varied greatly over time, KC's range sizes were larger overall for /p t/ than for /b d/, suggesting greater stability for the voiced category. In total, just 8% of the /b/ and /d/ sessions showed a range of >100 ms, whereas 42% of the /p/ and /t/ sessions showed a range size of >100 ms. The range data indicate that KC showed considerable token-to-token variability in her production, as typically reported for young children. She also varied extensively in this measure from session to session.

Accuracy

Mean VOT values for /b/ were accurate for all sessions. Those for /d/ were accurate in 23/25 sessions (all but sessions 14 and 18); upon closer inspection, those two session means barely crossed the accuracy threshold of 30 ms (30.54 and 30.51 ms). All /p/ and /t/ means were accurate in sessions 5–25 and 6–25, respectively. Such high and apparently stable accuracy did not always hold for individual productions, however, as shown by the token-by-token analyses.

Token-by-token accuracy measures are graphed in Figure 2. All phonemes had some productions that were inaccurate, but the numbers varied across consonants. All /b/ distributions were >80% accurate; all /d/ distributions with the exception of sessions 18 and 19 were >80% accurate. Early in the original study, KC's VOT values showed marked, rapid improvement in accuracy for /p/ and /t/. This progress occurred in tandem for the two phonemes (/p/ slightly ahead of /t/), with both rising above 80% accuracy after the sixth session. Subsequently, accuracy for /p/ and /t/ generally remained above >80%, albeit with a noticeable dip in accuracy (63%) for /t/ at session 23.

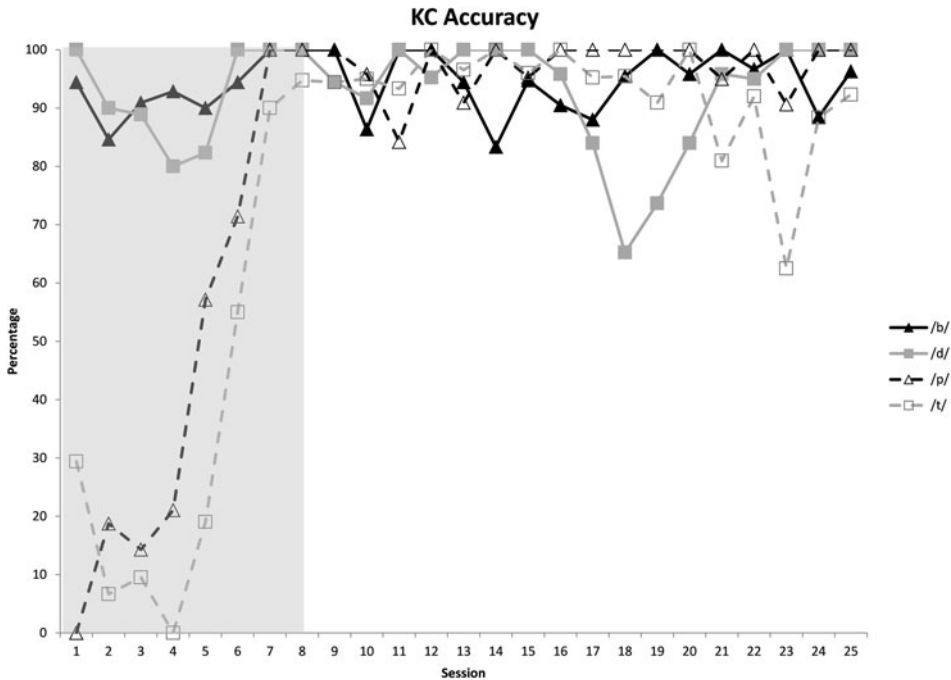


Figure 2. Accuracy of stop consonants for KC over 25 sessions, measured according to whether VOT values fell within adult boundaries. The solid lines show token-by-token accuracy percentages for voiced targets; dashed lines reflect accuracy percentages for voiceless targets. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

Accuracy was compared across consonants by taking the average accuracy per phoneme across all sessions. Greatest accuracy was observed for /b/ which measured an average of 94%, followed by /d/ at 93%, /p/ at 81% and /t/ at 75%. These data are consistent with expectations that the short-lag/voiced category is established earlier than the long-lag/voiceless category (Kewley-Port & Preston, 1974; Macken & Barton, 1980).

Comparing the token-by-token measures of accuracy (Figure 2) with the results from the means and the ANOVA results (Table 2) reveals some of the limitations of relying solely on reduced data and tests of statistical significance to determine contrast acquisition. In the current data, 42/50 sessions showed a statistically significant voiced/voiceless difference, seemingly indicating good production accuracy. However, just 12% (5/42) of these /b p/ and /d t/ comparisons were associated with 100% accuracy of both phonemes in the pair. Sixty-nine percent (29/42) of the remaining /b p/ or /d t/ comparisons were associated with accuracy between 75% and 100%, and 19% (8/42) of the /b p/ or /d t/ comparisons that reached statistical significance were associated with <75% accuracy (i.e. inaccuracy rates of 25% or more). As stated above, the mean VOT data provided no indication of inaccuracy in KC’s productions. Thus, mean values and statistically significant differences between voicing categories do not correspond well to high production accuracy as assessed for individual productions, as is commonly done in clinical practice (e.g. the PCC measure).

Discreteness

Figure 3 shows discreteness data calculated both for the session means and for individual tokens. As was observed for accuracy, discreteness based on the means was markedly higher than the

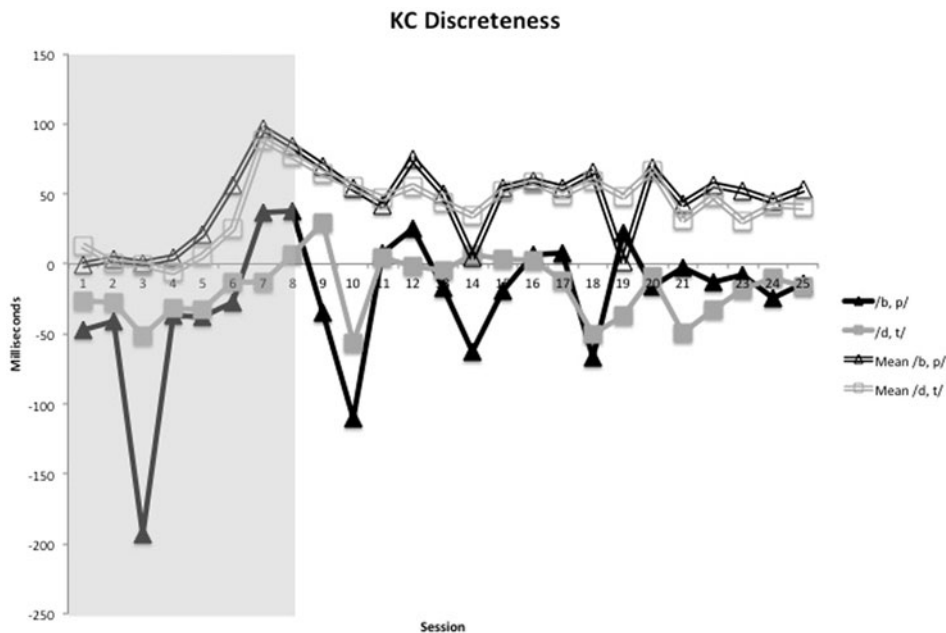


Figure 3. Discreteness values for /b p/ and /d t/ for KC over 25 sessions indicating overlap between the phoneme pairs. Plot lines indicate overlap as a negative value and separation between the pairs as a positive value (single lines for the token-by-token analysis; double lines for means analysis). The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

token-by-token results: For the means, discreteness was found in 23/25 (92%) and 22/25 (88%) sessions for /b p/ and /t d/. Of these discrete mean VOT comparisons, 83% (19/23) for /b p/ and 86% (19/22) for /d t/ were separated by >30 ms, the separation observed for adults (Zlatin, 1972). In short, mean discreteness suggested well-separated distributions.

In contrast, the token-by-token analysis showed that the VOT distributions did not remain discrete over time. Although discreteness was observed by session 7 for /b p/ distributions and session 8 for /d t/, overlap, or lack of discreteness, was observed for /b p/ in 12 of the 18 remaining sessions and for /d t/ in 12 of the 17 remaining sessions. This included the last six sessions for /b p/ and the last nine for /t d/. A review of the findings over 12 months showed that /b p/ were discrete in only 28% (7/25) of sessions; /t d/ were discrete in 24% (6/25). Only two of these sessions for /b p/ and one for /d t/ showed discreteness of ~30 ms.

Previous studies do not clearly show persistent overlap between voicing categories once a bimodal distribution is established. Thus, further analysis was completed to explore the percentage of overlapping tokens for /b p/ and /t d/ distributions. All tokens that fell within the highest /b/ and lowest /p/ target productions in a given session were considered to be within the overlap span. The number of productions within this group was summed and then divided by the total number of productions in the combined /b p/ set. For example, if the lowest VOT value for /p/ in a given session was 16 ms, a VOT value of 19 ms for /b/ would be within the overlap region (Supplemental Histograms: session 13). On the other hand, if the lowest /p/ VOT was 54 ms, a VOT of 29 ms for /b/ would not be in the overlap region (Supplemental Histograms: session 12).

The results of this analysis showed large proportions of overlap (>60%) for /b p/ and /d t/ contrasts from sessions 1–5 (Figure 4). Over time, smaller percentages of overlap were generally

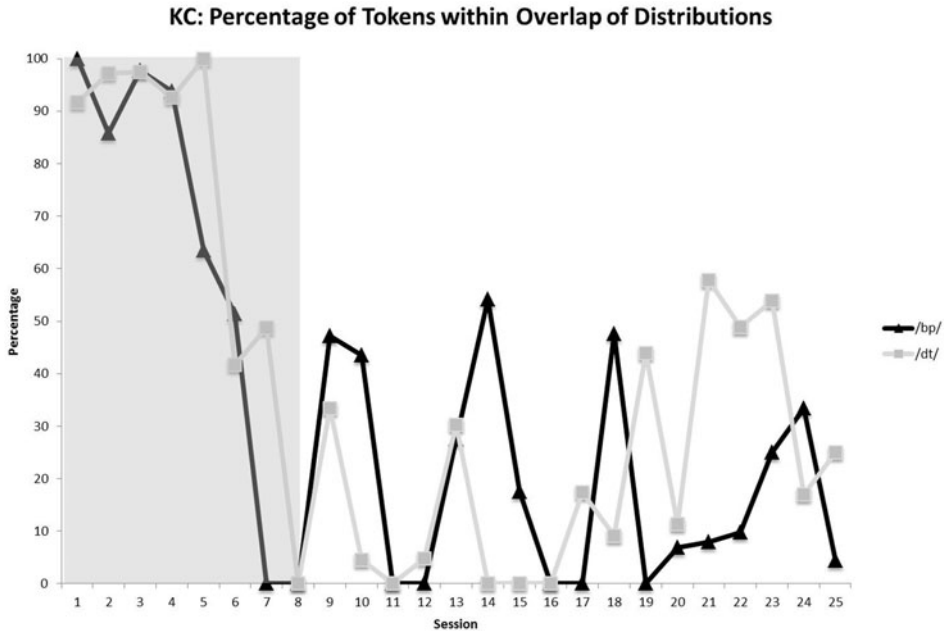


Figure 4. The percentage of overlapping tokens for /b p/ and /t d/ distributions for KC over 25 sessions. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

observed between the highest voiced token and the lowest voiceless token, but there was considerable variation across sessions. Overlap percentages ranged from 0% to 54% for /b p/ and 0% to 58% for /t d/ in sessions 6–25.

Overshoot

Based on mean VOT values, KC did not appear to demonstrate /p/ or /t/ overshoot for any of the 25 sessions; all session means were <100 ms. In contrast, token-by-token overshoot, calculated as a percentage of the VOT values ≥ 100 ms per distribution (Figure 5), shows that KC experienced a brief overshoot phase over a two-session, i.e. one-month period. Both /p/ and /t/ follow a similar trajectory over the 12 months with /t/ trailing approximately one session behind /p/ during the observed overshoot period. The token-by-token analysis also indicates that exaggerated VOT values (≥ 100 ms) consistently occurred in later sessions for a modest percentage of productions. That is, although the percentage of exaggerated values per VOT distribution peaked between sessions 7 and 8, small percentages of exaggerated aspiration persisted throughout the 12-month period.

Discussion

Similar to Hitchcock and Koenig (2013), the current VOT means analysis showed no overshoot, and consistently high accuracy and discreteness over time. The token-by-token analyses, on the other hand showed persistent VOT variability for labial and alveolar productions after KC had apparently established the voicing contrast. These differences across analysis methods provide further support for the premise of Hitchcock and Koenig (2013) that descriptions of development

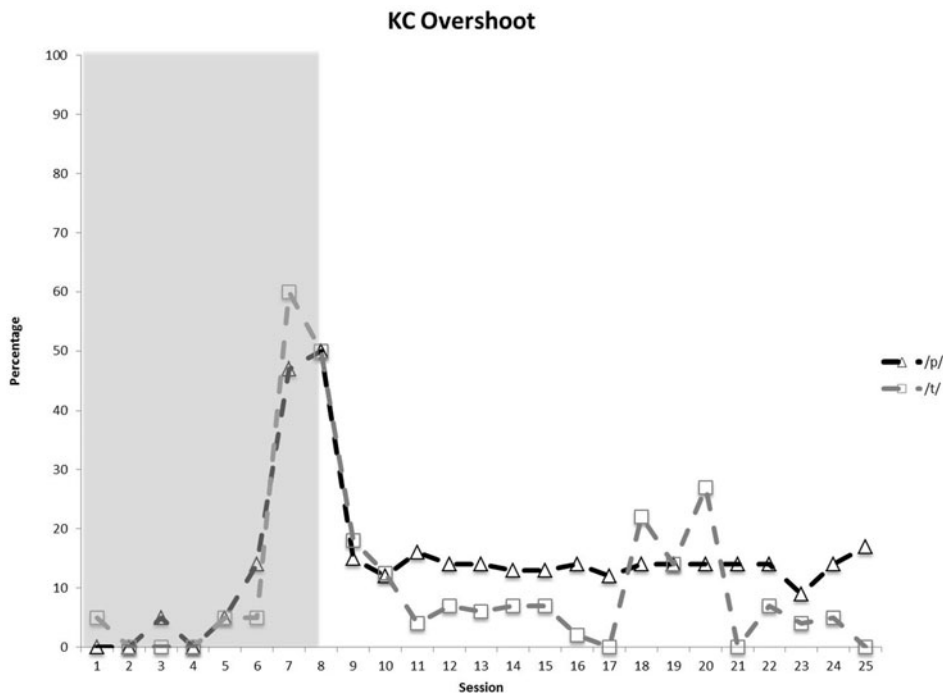


Figure 5. Overshoot of voiceless aspirated stop consonants /p/ and /t/ for KC over 25 sessions. The dashed lines show token-by-token overshoot percentages for /p/ and /t/. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

based on highly reduced data yield not only a limited, but in some cases inaccurate representation of a child's speech production development.

Subsequent to the rapid improvement in /p t/ accuracy, token-by-token accuracy for all four phonemes generally remained $\geq 80\%$, although occasional dips were observed. In other words, *inaccuracy* rates exceeded 20% at times. For discreteness, we observed large proportions of overlap ($>60\%$) in /b p/ and /t d/ VOTs from sessions 1 to 5. Subsequently, smaller percentages of overlap between voicing categories were observed, but 50% or more overlap was still occasionally observed in later sessions. Finally, the token-by-token measures (but not the means) showed a brief overshoot phase followed by the consistent presence of exaggerated VOT values (≥ 100 ms) for a modest percentage of productions. For this child, increased accuracy, emergence of discreteness and the overshoot phase occurred close in time across both places of articulation. Hitchcock and Koenig (2013) observed several such cases of parallel development across bilabial and alveolar places, as well as some cases of divergent development, e.g. discreteness improving for alveolars in a session where it worsened for bilabials. Thus, the acquisition of obstruent voicing spreads across places of articulation in a way that is child-specific.

Interactions across measures

It is instructive to consider potential relationships among measures. Interactions are obscured by the coarse nature of the means data, but such patterns are evident in the token-by-token analyses. For example, when /p/ and /t/ are initially established for KC with long-lag VOTs (between sessions 4 and 7; see Figure 2), discreteness and overshoot measures developed in

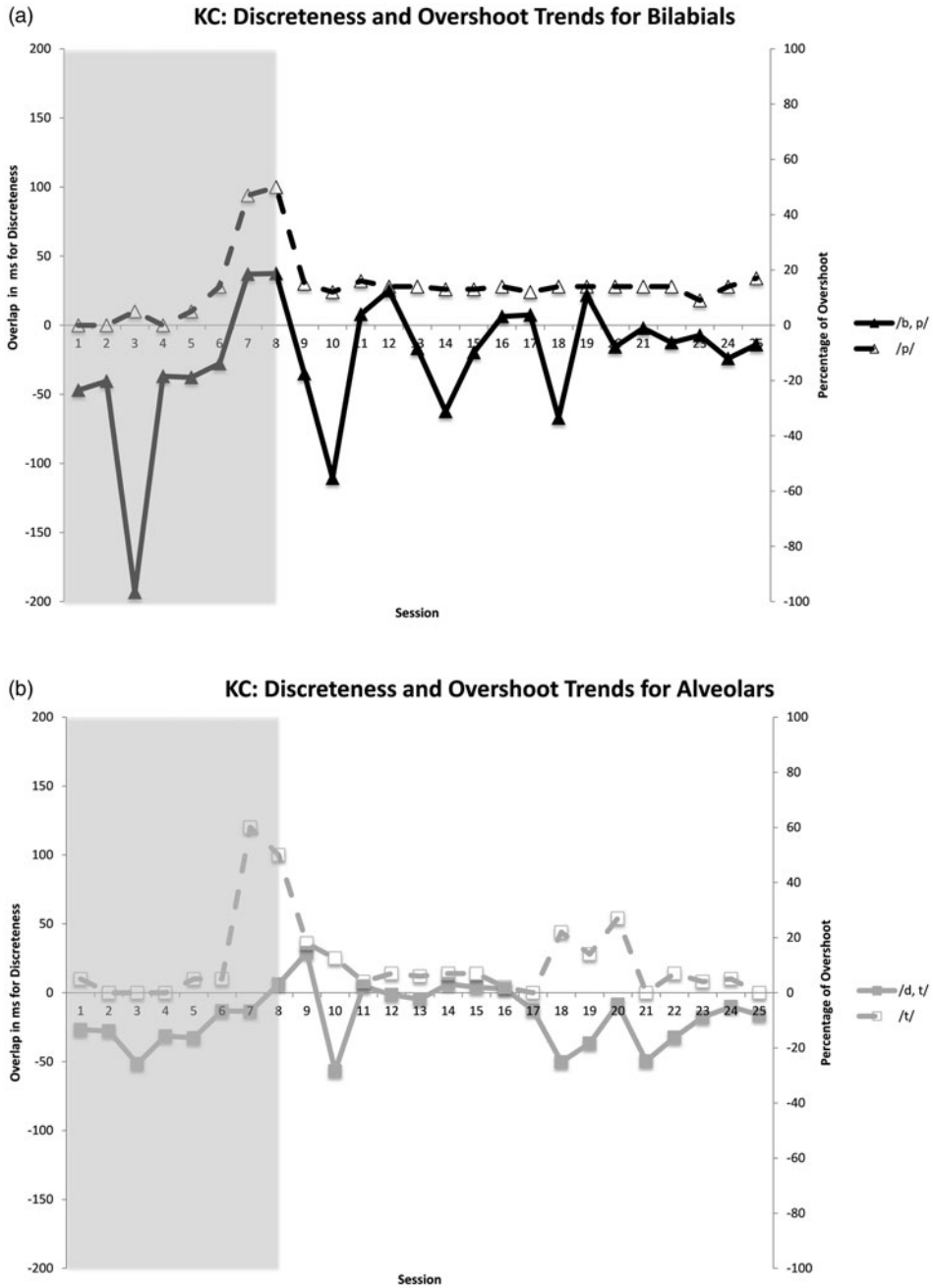


Figure 6. (a) Interactions across token-by-token measures for discreteness and overshoot: bilabial productions. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013). (b) Interactions across token-by-token measures for discreteness and overshoot: alveolar productions. The data in the grey-shaded sessions (1–8) were presented as part of Hitchcock and Koenig (2013).

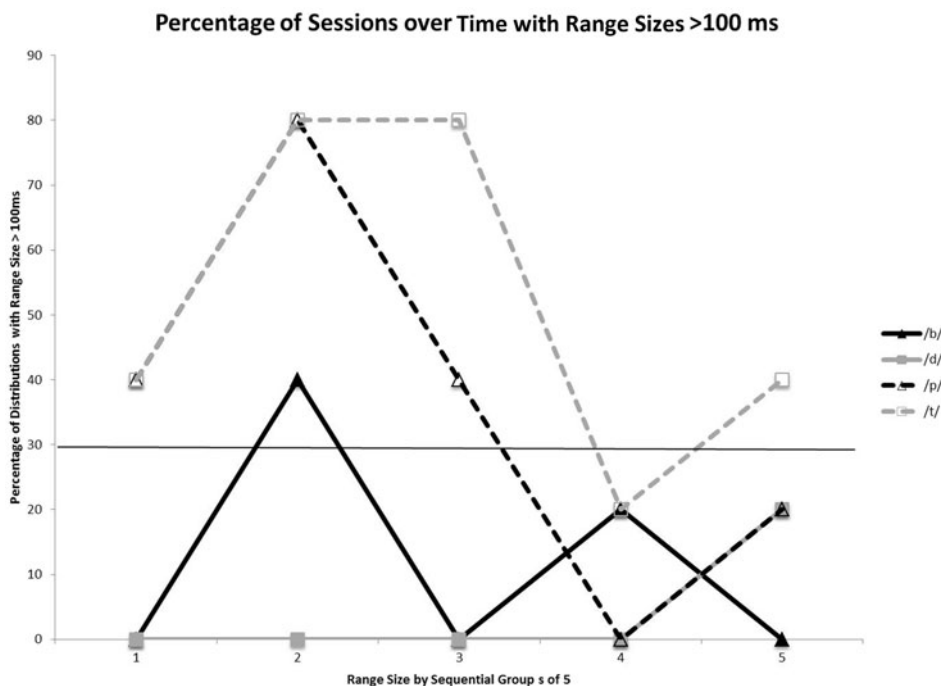


Figure 7. Grouped range size per phoneme across 25 sessions. The solid lines show the token-by-token percentage of sessions with range size ≥ 100 ms for voiced targets; dashed lines reflect percentage of sessions with range size ≥ 100 ms for voiceless targets.

parallel (Figure 6a and b); discrete voiced/voiceless categories were observed at both places of articulation at sessions 7–8, coincident with the sessions where /p/ (session 7) and /t/ (sessions 7–8) showed overshoot ($\geq 50\%$ of the VOT values over 100 ms). Thus, overshoot contributed to discreteness in these sessions. However, over time, low levels of overshoot persisted and discreteness was extremely variable.

Interactions can also be seen between range and token-by-token overshoot and accuracy measures. As stated above, overshoot was present in session 8 for /p/ and sessions 7–8 for /t/. An increase in range size can be seen leading up to, during, and slightly after the presence of overshoot (Figure 7). This is consistent with previous studies suggesting that production variability may increase during periods of learning (e.g. Goffman, Ertmer, & Erdle, 2002). When viewed together, these comparisons suggest that subtle increases in range size, together with the presence of overshoot and improved accuracy, represent evidence of evolving voicing acquisition *via* refinement of articulatory movements. Once /p/ and /t/ became largely accurate (after session 8), smaller range sizes were generally observed although intermittent sessions of exaggerated ranges persist and occasional marked dips in accuracy occurred for individual phonemes. These examples show the value of assessing child speech using multiple measures in combination (cf. Hitchcock & Koenig, 2013).

Stages of voicing acquisition

According to the means analysis, KC's voicing acquisition appeared to be abrupt: Both /b p/ and /d t/ categories moved from unimodal (Stage I) to bimodal, adult-like distributions (Stage IIIb),

Table 3. Stages of acquisition for KC based on mean VOT measures and token-by-token measures.

	Stage I	Stage II	Stage IIIa	Stage IIIb
Analysis of means	✓			✓
Token-by-token analysis	✓	✓	✓	

with little evidence of overlap and no overshoot (i.e. no Stage II or IIIa). Furthermore, the means analysis suggested nearly uniform adult-like discreteness (>30 ms separation between the voicing categories) from sessions 7 to 25, with only two cases of /b p/ separation <30 ms. Table 3 (top row) shows the apparent progression according to the means.

The token-by-token analyses imply a very different pattern of voicing development (Table 3, bottom row). Here, KC's VOT productions proceed regularly through the acquisitional schedule outlined by Macken and Barton (1980) and refined by Hitchcock and Koenig (2013), including a brief overshoot phase. At the end of the study, KC appears to have moved through Stages I, II and IIIa over the course of the 12-month period. In the token-by-token analysis, high accuracy might suggest movement into Stage IIIb, but accuracy values did not remain stable, falling below 80% for /d/ and /t/ in sessions 18 and 23, respectively. Moreover, the persistence of overlap throughout the duration of the study indicates that KC had not yet moved to Stage IIIb. These findings, viewed together, demonstrate that a child may show progress in one area (e.g. accuracy) but not in others (e.g. discreteness), and movement between stages may not occur simultaneously for all four assessment measures. Gains may also be variable, with movement towards adult-like productions followed by periods of regression (e.g. adult-like discreteness for a short period of time followed by a return to overlapping distributions in subsequent sessions). In sum, following Hitchcock and Koenig (2013), a child must demonstrate the range, accuracy, overshoot and discreteness criteria defined for each stage in order to be said to have achieved that level of voicing development. More generally, the results contribute to a view of acquisition more as an ongoing process than a discrete event. Indeed, they support a suggestion made by Macken and Barton (1980, p. 73): "Although considerable progress is usually made by 2;0, it may be many more months (or even years) before children acquire sufficient articulatory skill to consistently produce adult-like voicing".

Although not always framed in terms of phonological contrasts *per se*, decades of work demonstrate that speech motor control develops for a number of years, with even school-age children showing, for example, longer durations and higher token-to-token variability than adults (see, e.g. Koenig, Lucero, & Perlman, 2008 and references therein). The literature on motor learning characterises acquisition as arising from practice or experience that facilitates development of the target motor skill (e.g. Goffman et al., 2002; Wu, Miyamoto, Castro, Ölveczky, & Smith, 2014). The premise that children engage in practice, or demonstrate trial and error productions, during acquisition suggests that development is a continuously evolving exercise toward a goal (e.g. a voicing target). Given the precise laryngeal and supralaryngeal control required to produce accurate voiced and voiceless plosives it is not surprising that KC demonstrated considerable movement within and between measures as her voicing output became more adult-like. It is likely that the protracted course of speech perception development (e.g. Flege & Eefting, 1986; Hazan & Barrett, 2000; Nittrouer & Studdert-Kennedy, 1987; Zlatin & Koenigsknecht, 1975) also contributes to ongoing modification of production. What is striking about the current results is the magnitude of evidence, obtained at short time intervals, that KC engaged in extended practice and refinement of the voicing contrast over the course of many months. To date, the literature has not provided longitudinal data of this extent charting not only the process of voicing acquisition, but also the degree to which accuracy and discreteness

remain stable. Our results suggest that the English voicing contrast may not stabilise for some time. The persisting overlap between categories (lack of discreteness) observed here is consistent with a recent study of stop voicing and fricative contrasts in children ages 9–14 and adults (Romeo et al., 2013). Those authors found that overlap was more prevalent in children (particularly boys) than adults. Along similar lines, a 10-month follow-up study by the first author and colleagues showed extensive production variability in stop consonant voicing in 3–8-year-old English-speaking children across all three places of articulation (Hitchcock, Ochs & Gencarelli, 2013).

The occasional reductions in accuracy and return to overlapping distributions in KC's voicing contrast recall findings of developmental phonologists, who observed that acquisition is not simply a linear course of improvement towards adult-like forms, but may involve apparent regression, a phenomenon that some have attributed to phonological reorganisation (e.g. Ferguson & Farwell, 1975; Macken, 1979; see also theoretical review in Vihman, 1996). Evidence of regression can also be seen in studies of developing motor control for both speech (e.g. Goffman et al., 2002) and non-speech actions (e.g. Thelen, Corbetta, & Spencer, 1996). The current study provides further evidence for the long-term nature of phonological contrast acquisition (which we consider to be both a linguistic and motoric process). Valid interpretation of the results of any single-session assessment of a child's speech production must recognise the dynamic and sometimes nonlinear nature of the process.

Clinical relevance of contrastive voicing acquisition

When conducting a comprehensive phonological evaluation, many clinicians and researchers include an inventory analysis to assess a child's speech sound productions (e.g. Bauman-Waengler, 2004; Bleile, 2004; Craghead, Newman, & Secord, 1989; Pena-Brooks & Hedge, 2007; Smit, 2004; Vellemen, 2003; Williams, 2003). For example, a phonetic inventory may be used to assess the age appropriateness of a child's speech (Bleile, 2004) or determine treatment targets (Gierut, 2005). In order to do so, an accuracy criterion may be added for attributing sounds to a child's inventory (Ingram & Ingram, 2001; Stokes, Klee, Carson, & Carson, 2005). Perceived errors in voicing may contribute to misdiagnosing or overestimating the severity of a child's speech sound disorder if the nature of typical voicing development is not well-understood.

For example, childhood apraxia of speech (CAS), characterised by difficulty in sequencing speech movements in the absence of muscle weakness, has long been recognised as a speech sound disorder for which differential diagnosis is challenging (ASHA, 2007). Many researchers have attempted to determine a standard set of diagnostic criteria for CAS. Forrest (2003) asked 75 speech–language pathologists to report three diagnostic markers they use in order to diagnose a child with CAS. In total, 50 different characteristics were given by the participants, supporting previous literature that emphasises the ambiguity of a CAS diagnosis. Among this set of 50, voicing errors emerged as a potential, albeit less common, marker for CAS. This finding is supported by Shriberg, Potter, and Strand (2011), who reported voicing errors as one of 10 possible markers of CAS based on Strand's 10 point checklist (Shriberg, Potter, & Strand, 2009). According to this protocol, if a child presents with a minimum of four out of the 10 possible markers assessed across three tasks, the child is given a diagnosis of CAS. It is important to point out that Shriberg et al. (2011) and Murray, McCabe, Heard, and Ballard (2015), who also used Strand's 10 point checklist as one of three possible assessments for differential diagnosis of CAS, did so with children ages 3 and older. Given the information presented in the current study, clinicians should be aware that using these protocols with children under three may be inappropriate. We do not necessarily challenge their use as diagnostic markers for older children, but do note that robust, longitudinal research regarding the stability of voicing in 3-year-olds is sparse. More generally, as indicated above, we suggest that results of one-time assessments of

voicing contrasts in young children should be viewed with some caution. Other phonetic/phonological features may well show similar patterns of variability in early development. It is also clear that extended instability of an emerging contrast can be seen in typically-developing children, and can therefore likely be expected in at least some children undergoing intervention as well.

Limitations and future directions

Case-study methodology always raises questions of how well the population is represented by the participant(s). Since KC was in our earlier group study (she was C8 in Hitchcock & Koenig, 2013), we are able to verify that she was not an obvious outlier among that group of 10, granting that (as stated earlier) there was considerable variation across children. For example, KC tended to show larger ranges for /p t/ than /b d/, as observed for most other children. Four of the 10 children in that study, KC being one, showed rapid gains in accuracy (over one or two sessions, i.e. 2–4 weeks) for at least one phoneme; for KC and one other child this was observed for the voiceless sounds. Like most children, KC did not show discrete VOT distributions for most of her early recording sessions. Finally, KC entered her overshoot phase rapidly (again, over the course of 2–4 weeks), as was true for eight of the 10. Based on those initial four months at least, KC does not appear to be an atypical child.

It is also the case that the current study collected data for a very small set of words that were elicited in focused recording sessions. The total duration of recording (about 12.5 h, i.e. 25 sessions of about 30 min) represents a very small portion of KC's second year of life and development; indeed, parental checklists showed that during the first four months of the group study KC's lexicon grew from 155 to 257 words, and her syntax expanded to three-word utterances. Nevertheless, one could argue that these four stimulus words were rather well-practiced by the end of the study. This would suggest that our data may represent a "best-case" scenario for development and stabilisation of a voicing contrast; under typical acquisitional conditions one might observe even more variability and instability than what was found here.

Finally, for practical reasons, our stimulus set excluded the velar place of articulation. As noted above, the bilabial and alveolar places of articulation showed cases of parallel development for the current participant, as well as for several participants in our group study. The literature provides examples of development in tandem or close in time across places of articulation including the velars (Macken & Barton, 1980); nevertheless, since velars often emerge later than other places of articulation, the velar voicing contrast may, on average, lag behind more anterior places of articulation. Future longitudinal and cross-language work is needed to determine the variety of patterns that children may show in establishing voicing characteristics across places of articulation.

Granting these necessary limitations of our study design, however, these methods allowed us to present data obtained at short, systematic intervals for a duration that would have been prohibitively long for a large group study. In this respect, the current study complements our previous group study of 10 children, and presents a more complete picture of short-term changes in voicing development than is available in any previous study. The recent popularity of single-subject designs in clinical research (e.g. Gierut, Morrisette, & Dickinson, in press) has made frequently-obtained longitudinal data available on numerous children undergoing a course of therapy. Interpreting such data requires comparable data from typically-developing children. The results presented here may be useful in this regard. Additional studies collecting many productions of target consonants per session scheduled at systematic, short time intervals, over a long time period, will contribute to understanding how and when children's voicing contrast stabilises. Finally, applying the proposed token-by-token analyses to more children and expanding the age

range would provide much needed data to further refine our understanding of patterns of voicing development, and contribute to more sensitive differentiation between typical and atypical development.

Conclusions

The current findings provide additional support for the schedule of voicing acquisition proposed by Hitchcock and Koenig (2013), and for analysing child productions on a token-by-token basis. A finding of interest is that modest percentages of exaggerated VOT values for /p t/, i.e. overshoot, persisted long after KC's brief "overshoot phase". Discreteness measures also showed an extended period of non-adult-like behaviour. These results clearly indicate that /b p/ and /d t/ distributions did not develop in a linear progression from unimodal to bimodal distributions but instead moved from short lag to extremely long VOT values before moving to less exaggerated long-lag VOT values for voiceless targets. However, in this child overlap returned after the overshoot phase, suggesting that she had not yet acquired the fine motor control required to produce discrete, bimodal distributions without using exaggerated voiceless productions. It is noteworthy that in the absence of token-by-token analysis and short intervals between recordings, KC would not have appeared to go through Stage II or IIIa, nor would she have revealed persistent overlap or inaccurate productions. This research provides clear evidence of an extended duration of fluctuating voicing patterns for KC, suggesting that the voiced/voiceless contrast, once achieved, may take months or possibly even years to stabilise. Given these findings, clinical use of voicing errors as an indication of a phonological disorder or CAS would not be justified until after the age of at least 3. Future research with a larger sample size should be carried out to allow for greater generalisation of the results.

Acknowledgements

The authors gratefully acknowledge the contributions of the following individuals: Patricia Gencarelli, Laura Ochs, Kurt Keena, Julie Irwin and Christy Van Jackael, student research assistants. We would also like to thank KC and her family for their commitment and cooperation throughout the extended duration of the study. Aspects of this research were presented at the annual convention of the American Speech–Language Hearing Association in Atlanta, GA (2012).

Declaration of interest

The authors have no financial or nonfinancial disclosures to report.

References

- American Speech–Language–Hearing Association. (2007). *Childhood apraxia of speech* [Position statement]. Retrieved from <http://www.asha.org/policy>.
- American Speech–Language–Hearing Association. (2008). Roles and responsibilities of speech–language pathologists in early intervention: Guidelines. Retrieved from <http://www.asha.org/policy>.
- Bailey, P. J., & Haggard, M. P. (1973). Perception and production: Some correlations on voicing of an initial stop. *Language and Speech*, 16, 189–195.
- Bailey, P. J., & Haggard, M. P. (1980). Perception–production relations in the voicing contrast for initial stops in 3-year-olds. *Phonetica*, 37, 377–396.

- Bauman-Waengler, J. (2004). *Articulatory and phonological impairments: A clinical focus* (2nd ed.). Boston, MA: Pearson.
- Bharadwaj, S. V., & Graves, A. G. (2008). Efficacy of the discreteness of voicing category (DOVC) measure for characterizing voicing errors in children with cochlear implants: A report. *Journal of Speech, Language, and Hearing Research, 51*, 629–635.
- Bleile, K. (2004). *Manual of articulation and phonological disorders: Infancy through adulthood* (2nd ed.). Clifton Park, NY: Delmar.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Creaghead, N. A., Newman, P. W., & Secord, W. A. (1989). *Assessment and remediation of articulatory and phonological disorders* (2nd ed.). Columbus, OH: Merrill.
- Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngologica, 257*, 1–51.
- Fabiano-Smith, L., & Bunta, F. (2012). Voice onset time of voiceless bilabial and velar stops in 3-year-old bilingual children and their age-matched monolingual peers. *Clinical Linguistics & Phonetics, 26*, 148–163.
- Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51*(2), 419–439.
- Flege, J. E., & Eefting, W. (1986). Linguistic and developmental effects on the production and perception of stop consonants. *Phonetica, 43*, 155–171.
- Forrest, K. (2003). Diagnostic criteria of developmental apraxia of speech used by clinical speech–language pathologists. *American Journal of Speech–Language Pathology, 12*(3), 376–380.
- Gierut, J. A. (2005). Phonological intervention: The how or the what? In A. G. Kamhi & K. E. Pollock (Eds.), *Phonological disorders in children: Clinical decision making in assessment and intervention* (pp. 201–210). Baltimore, MD: Brookes.
- Gierut, J. A., Morrisette, M. L., & Dickinson, S. L. (in press). Effect size for single-subject design in phonological treatment. *Journal of Speech, Language, and Hearing Research*, (forthcoming).
- Goffman, L., Ertmer, D. J., & Erdle, C. (2002). Changes in speech production in a child with a cochlear implant: Acoustic and kinematic evidence. *Journal of Speech, Language, and Hearing Research, 45*(5), 891–901.
- Goldman, R., & Fristoe, M. (2000). *Goldman–Fristoe test of articulation* (2nd ed.). Circle Pines, MN: American Guidance Service.
- Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28*, 377–396.
- Hitchcock, E. R., & Koenig, L. L. (2013). The effects of data reduction in determining the schedule of voicing acquisition in young children. *Journal of Speech, Language, and Hearing Research, 56*(2), 441–457.
- Hitchcock, E.R., & Koenig, L.L. (2015). Adults' perceptual voicing boundaries of 2-year-olds' citation form speech. Poster to be presented at the 170th meeting of the Acoustical Society of America, Jacksonville, FL.
- Hitchcock, E. R., Ochs, L., & Gencarelli, P. (2013). Patterns of contrastive voicing in typically developing preschool and school age children. *Poster session presented at the American Speech–Language–Hearing Association National Convention*. Chicago, IL.
- Hodson, B., & Paden, E. (1991). *Targeting intelligible speech* (2nd ed.). Austin, TX: Pro-Ed.
- Ingram, D., & Ingram, K. D. (2001). A whole-word approach to phonological analysis and intervention. *Language, Speech and Hearing Services in Schools, 32*, 271–283.
- Kewley-Port, D., & Preston, M. S. (1974). Early apical stop production: A voice onset time analysis. *Journal of Phonetics, 2*, 195–210.
- Koenig, L. L., Lucero, J. C., & Perlman, E. (2008). Speech production variability in fricatives of children and adults: Results of functional data analysis. *Journal of the Acoustical Society of America, 124*, 3158–3170.
- Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63*, 905–917.
- Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–442.
- Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10*, 1–28.
- Lowenstein, J. H., & Nittrouer, S. (2008). Patterns of acquisition of native voice onset time in English-learning children. *Journal of the Acoustical Society of America, 124*, 1180–1191.
- Macken, M. A. (1979). Developmental reorganization of phonology: A hierarchy of basic units of acquisition. *Lingua, 49*, 11–49.
- Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice onset time in word initial stop consonants. *Journal of Child Language, 7*, 41–74.
- Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35*, 259–273.

- Murray, E., McCabe, P., Heard, R., & Ballard, K. J. (2015). Differential diagnosis of children with suspected Childhood Apraxia of Speech. *Journal of Speech, Language, and Hearing Research*, 58, 43–60.
- Nittrouer, S., & Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research*, 30, 319–329.
- Pena-Brooks, A., & Hedge, M. N. (2007). *Assessment and treatment of articulation and phonological disorders in children* (2nd ed.). Austin, TX: Pro-Ed.
- Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *Journal of the Acoustical Society of America*, 134, 3781–3792.
- Sander, E. K. (1972). When are speech sounds learned? *Journal of Speech and Hearing Disorders*, 37, 55–63.
- Scobbie, J. M., Gibbon, F., & Hardcastle, W. J., Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. *Papers in Laboratory Phonology V: Language Acquisition and the Lexicon*, 5, 194–203.
- Shriberg, L., Austin, D., Lewis, B. A., McSweeney, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, 40, 708–722.
- Shriberg, L. D., Potter, N. L., & Strand, E. A. (2009). *Childhood apraxia of speech in children and adolescents with galactosemia. Paper presented at the American Speech-Language-Hearing Association National Convention*, New Orleans, LA.
- Shriberg, L. D., Potter, N. L., & Strand, E. A. (2011). Prevalence and phenotype of childhood apraxia of speech in youth with galactosemia. *Journal of Speech, Language, and Hearing Research*, 54, 487–519.
- Smit, A. B. (2004). *Articulation and phonology resource guide for school age children and adults*. Clifton Park, NY: Delmar Learning.
- Snow, D. (1997). Children's acquisition of speech timing in English: A comparative study of voice onset time and final syllable vowel lengthening. *Journal of Child Language*, 24, 35–56.
- Stoel-Gammon, C., & Dunn, C. (1985). *Normal and disordered phonology in children*. Austin, TX: PRO-ED.
- Stokes, S. F., Klee, T., Carson, C. P., & Carson, D. (2005). A phonemic implicational feature hierarchy of phonological contrasts for English speaking children. *Journal of Speech, Language, and Hearing Research*, 48, 817–833.
- Thelen, E., Corbetta, D., & Spencer, J. P. (1996). Development of reaching during the first year: Role of movement speed. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1059–1076.
- Tyler, A. A., & Saxman, J. H. (1991). Initial voicing contrast acquisition in normal and phonologically disordered children. *Applied Psycholinguistics*, 12, 453–479.
- Velleman, S. L. (2003). *Childhood apraxia of speech resource guide*. Clifton Park, NY: Delmar Learning.
- Vihman, M. M. (1996). *Phonological development: The origins of language in the child*. Cambridge, MA: Blackwell Publishers Inc.
- Williams, A. L. (2003). *Speech disorders resource guide for preschool children*. Boston, MA: Allyn & Bacon.
- Wu, H. G., Miyamoto, Y. R., Castro, L. N. G., Ölveczky, B. P., & Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nature Neuroscience*, 17, 312–321.
- Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Zlatin, M. A. (1972). Development of the voicing contrast: A psychoacoustic study of voice onset time (Unpublished doctoral dissertation). *Northwestern University*, Evanston, IL.
- Zlatin, M. A. (1974). Voicing contrast: Perceptual and productive voice onset time characteristics of adults. *Journal of the Acoustical Society of America*, 56, 981–995.
- Zlatin, M. A., & Koenigsnecht, R. A. (1975). Development of the voicing contrast: Perception of stop consonants. *Journal of Speech and Hearing Research*, 18, 541–553.
- Zlatin, M. A., & Koenigsnecht, R. A. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research*, 19, 93–111.

Supplementary material available online.