JSLHR

1716

Article

The Effects of Data Reduction in Determining the Schedule of Voicing Acquisition in Young Children

Elaine R. Hitchcock[a] and Laura L. Koenig[b,c]

Purpose: In this study, multiple measures of voicing acquisition were used to evaluate the extent to which developmental patterns based on voice onset time (VOT) mean data differed from those based on token-by-token analyses in typically developing 2-year-olds.
Method: Multiple repetitions of words containing initial /b p d t/ were elicited from 10 English-speaking children biweekly for 4 months. VOT was measured for each stop. For each child, consonant, and recording session, means and ranges were obtained, as were measures of accuracy, discreteness, and overshoot calculated for session means and for individual tokens.
Results: The token-by-token analyses suggested lower accuracy and more category overlap than the session means and revealed an overshoot phase for all children. They also showed examples of both abrupt and gradual changes that were not always evident in the means. Measures of range, accuracy, discreteness, and overshoot all continued to change after statistically significant VOT differences were observed.
Conclusions: The findings suggest that some aspects of voicing development may not be evident in analyses that rely on VOT mean data and patterns of statistical significance. Token-by-token measures provide a more complete picture of stages of voicing development than those based solely on mean VOT values.

Key Words: voice onset time, contrastive voicing acquisition, discreteness

Extensive past research indicates that stop-consonant voicing contrasts are usually acquired in the first 2 to 3 years of life. However, as reviewed below, some details of this progression remain unclear. In this study, following much previous work, we investigated voicing development in syllable-initial stop consonants using *voice onset time* (VOT; Lisker & Abramson, 1964). Defined as the interval between the release of vocal tract occlusion and the onset of glottal vibration, VOT can be measured with high reliability from the acoustic signal (Francis, Ciocca, & Yu, 2003). Researchers in numerous studies have used VOT to assess voicing acquisition in English-learning children (Barton & Macken, 1980; Enstrom, 1982; Gilbert, 1977; Kewley-Port & Preston, 1974; Lowenstein & Nittrouer, 2008; Macken & Barton, 1980; Menyuk & Klatt, 1975; Preston & Yeni-Komshian, 1967; Preston, Yeni-Komshian, Stark, & Port, 1968; Snow, 1997; Tyler & Edwards, 1993; Tyler & Saxman, 1991; Zlatin, 1972). The goal of the present study was to gain greater insight into English VOT development in individual children by using multiple measures and frequent longitudinal sampling. We were particularly interested in how *data reduction*—that is, reliance on mean data as opposed to analyses based on token-by-token measures—might affect the identification of developmental patterns.

[a]Montclair State University, Montclair, NJ
[b]Haskins Laboratories, New Haven, CT
[c]Long Island University—Brooklyn, NY
Correspondence to Elaine R. Hitchcock: hitchcocke@mail.montclair.edu

Editor: Anne Smith
Associate Editor: Karen Forrest
Received July 7, 2011
Revision received January 20, 2012
Accepted July 15, 2012
DOI: 10.1044/1092-4388(2012/11-0175)

VOT in English Acquisition

In adult speakers of American English, the stressed, syllable-initial phonemes /b d g/ are usually produced without glottal vibration during the closure, although phonation (prevoicing) is optional. Consequently, these phonemes are most typically characterized by *short voicing lag,* where glottal pulsing begins at or shortly after the oral release (0–30 ms), or else by *voicing lead,* where the glottal pulsing begins prior to release of supraglottal occlusion. The phonemes /p t k/ in the same environment are produced without glottal vibration during closure and with aspiration (a delay in voicing onset after the release burst) and are characterized by *long voicing lag* (35–100 ms; Lisker & Abramson, 1964). The developmental data indicate that, cross-linguistically, the short-lag category is most prevalent in early child speech for all target categories (Eilers, Oller, & Benito-Garcia, 1984; Gandour, Petty, Dardarananda, Dechongkit, & Mukngoen, 1986; Kewley-Port & Preston,

1974; Macken & Barton, 1980; Preston et al., 1968; Whalen, Levitt, & Goldstein, 2007). Thus, a key component of achieving the phonemic contrast between the syllable-initial phonemes /b d g/ and /p t k/ in English is learning how to produce *aspiration,* or longer VOTs, for /p t k/.

Table 1 summarizes the literature on VOT in English-learning children. It is evident that past studies have varied widely in sampling method (longitudinal vs. cross-sectional), sample size, age range, method of eliciting target productions, time interval between data collections, number of data collection sessions, and statistical procedures. Such methodological differences may account for some of the questions that remain in this literature. One example is the nature of the progression from short-lag to adultlike values for voiceless aspirated stops. Zlatin and Koenigsknecht (1976), in a cross-sectional study of 10 children ages 30–36 months, reported that children showed a gradual progression from primarily short-lag stops for /p t/ to long-lag stops approximating the adult ranges. Macken and Barton (1980), based on longitudinal data from four children recorded over 8 months beginning at ages 1;4–1;7 (years;months), proposed a three-step developmental sequence for VOT. They observed, similar to Kewley-Port and Preston (1974), that English-speaking children in early phases of voicing acquisition produced unimodally distributed voiceless unaspirated stops, and designated this as Stage I. In the next stage (Stage II), children showed reliable differences between stops as a function of the phonemic voicing status of the target, but values remained in the short-lag range. The pattern was then marked by exaggerated long-lag values (overshoot) for /p t k/ (Stage IIIA) with regression to more adultlike VOT values over time (Stage IIIB). The overshoot phase of Macken and Barton (1980) contrasts with Zlatin and Koenigsknecht's (1976) conclusion that children achieved adultlike VOT values through a process of gradual approximation (cf. also Lowenstein & Nittrouer, 2008, on this observation). It may well be that overshoot is difficult to observe in a cross-sectional study if the children in a given age group differ in their current stages of voicing acquisition. On the other hand, Macken and Barton's sample was not very large, so the universality of an overshoot phase is not entirely clear. A complication of Macken and Barton's study is that they occasionally combined data from multiple sessions to arrive at four data sets spaced approximately 2 months apart. Thus, one question of the current study was whether a longitudinal analysis of a larger number of children—evaluated every 2 weeks, without combining data across sessions—would reveal a consistent overshoot phase.

A second question is the extent to which children's productions have VOT values appropriate to the target category on a token-by-token basis. It has been argued

Table 1. Methods used in voice onset time (VOT) studies of children 1–3 years of age.

Study	Year	*N* subjects	Age at study onset (years; months)	Study duration	Intervals between data collection	Total # of sessions	Data collection procedures	Data reporting
Bailey & Haggard	1980	41	3;3	Usually 1 session	1 month maximum	Up to 3x	Elicited from picture stimuli	Mean VOT
Bond & Wilson*	1980	10	1;10–3;4	1 session	N/A	N/A	Spontaneous and elicited speech	Grouped frequency distribution
Gilbert	1977	6	2;7–3;3	1 session	N/A	N/A	Spontaneous speech	Data from all children combined for mean VOT and frequency distribution
Kewley-Port & Preston	1974	3 (+ data from other children)	1;0	1 year	~3 mos. Data intervals grouped for presentation	7x – 2 subjects; 5x – 1 subject	Spontaneous speech	Mean VOT and frequency distribution
Lowenstein & Nittrouer	2008	7	1;0	2 years	2 mos.	~12x	Spontaneous speech	Mean VOT
Macken & Barton	1980	4	1;4–1;7	8 mos.	2 wks.; first few sessions combined for analysis	Approx. 16x	Spontaneous, elicited, imitated, repeated, modeled	Mean VOT
Preston & Yeni-Komishian	1967	14	1;0	1 year	2–4 wks.	Variable: min = 1x; max = 23x	Unknown	Unknown
Preston et al.	1968	6	1;0	1 year	2–4 wks.	Unknown	Unknown	Unknown
Snow	1997	10	1;0–1;8	9 mos.	3 mos.	4x	Semistructured play analysis	Mean VOT
Tyler & Saxman*	1991	3	1;9–1;11	3–8 mos.	1–2 wks.	6–18x	Spontaneous and elicited stimuli	Mean VOT
Zlatin & Koenigsknecht	1976	10	2;6–3;0	2 wks.	Variable	Variable	Elicited from picture stimuli	Mean VOT and frequency distribution

*Studies also included children with phonological disorders not recorded in this table.

that child production of VOT values within adult ranges is likely to correspond to adult perception of a phonemically accurate voicing contrast: Kewley-Port and Preston (1974) and Macken and Barton (1980) noted that children with a reliable mean VOT difference but with most values remaining in the short-lag range (i.e., children in Macken & Barton's Stage II) would have a contrast that adults might generally not be able to perceive (a subphonemic difference). Studies in which data reporting is limited to mean values and/or statements of statistically significant differences may mask the degree of *token-to-token accuracy,* defined as values appropriate to the target based on adult production and perception data. Thus, a second goal of this study was to track the accuracy of children's VOT values based not only on mean data but also on a token-by-token basis.

Related to the issue of accuracy is that of *discreteness,* or the extent to which contrastive categories show VOT overlap. Bailey and Haggard (1973) quantified the difference between mean /p b/ and /k g/ values in adult speech, and Bailey and Haggard (1980) referred to this as a measure of *distinctiveness.*[1] Zlatin and Koenigsknecht (1976) measured discreteness as the difference between the highest voiced target VOT value and the lowest target voiceless VOT value for individual speakers. Their results indicated that discreteness improved with age: Two-year-old speakers, as a group, showed overlap in their VOT distributions; overlap was minimal in 6-year-olds (a mean overlap of 1 ms or less); adults showed a wide VOT separation (30 ms or more) between voicing categories. To our knowledge, no past authors have evaluated patterns of discreteness longitudinally. However, a recent study on speech production in children with cochlear implants suggests that discreteness corresponds well with listeners' perceptions of whether such speakers had a typical voicing contrast (Bharadwaj & Graves, 2008). As such, discreteness, like accuracy, could provide additional insight into voicing contrast development.

Goals

The aim of this study was to explore voicing acquisition in 10 typically developing 2-year-old children, evaluating production variability (via measures of VOT range), accuracy, discreteness, and overshoot in biweekly sessions and giving particular attention to how the identification of developmental patterns may depend on whether assessments are based on VOT means versus token-by-token measures. The specific research questions were as follows:

1. What pattern or patterns of voicing acquisition may be observed in English-speaking children with typically developing phonology when followed longitudinally over short (biweekly) time intervals using an expanded set of measures?
2. How do patterns of development observed in VOT mean data compare with those observed in analyses of individual tokens, and do any such differences explain variations in the previously proposed stages of voicing acquisition?
3. Is the VOT acquisitional schedule a gradual movement toward adult values, or are there developmental nonlinearities such as the overshoot phase reported by Macken and Barton (1980)?

Method

Participants

Ten children (six girls, four boys) ages 27–29 months were selected for this study. At study onset, five were 27 months old (three girls, two boys); three were 28 months old (two girls, one boy); and two were 29 months old (one girl, one boy). Children are referred to by their participant number (C1–C10), where C1 was the youngest child at study onset, and C10 was the oldest. That is, lower participant numbers indicate younger ages. All participants were monolingual English speakers with typical speech and language development. Speech and language proficiency was evaluated by the first author, a certified speech-language pathologist, using the Preschool Language Scale—3 (Zimmerman, Steiner, & Pond, 1992) and the Goldman Fristoe Test of Articulation—2 (Goldman & Fristoe, 2000), as well as informal perceptual evaluation of conversational speech. A parent interview was conducted to verify that each child's medical and psychological history was unremarkable. A free-field pure-tone hearing screening was conducted in an audiometric soundproof suite. All children demonstrated normal hearing sensitivity over octave intervals from 250 Hz to 4000 Hz at 25 dB. 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. If a child demonstrated < 75% acoustic accuracy (defined in subsequent sections and in Table 2) for one or more of the target consonants, he or she was included in the study. This ensured that none of the children had fully accurate VOT characteristics at study onset.

Recording and Stimulus Materials

Recordings were made in a quiet room in the child's home using a Sony (MZ-N707) portable minidisk recorder (sampling frequency = 44.1 kHz). A Kay Pentax lapel microphone (Model 3502) was attached to the child's clothing approximately 6 in. from the mouth for a favorable signal-to-noise ratio and minimal feedback distortion.

Biweekly observations were carried out over 4 months for each child. A minimum of 15 productions each of /b/, /d/, /p/, and /t/ in initial/prevocalic position was attempted in each recording session to ensure that VOT distributional

[1]Bailey and Haggard (1973) was a study on production–perception correlations; they do not report discreteness data per se, but only how this and other production measures correlated with perceptual data. Similarly, although Bailey and Haggard (1980) mention the distinctiveness measure, they do not actually report distinctiveness data for their child speakers.

Table 2. Definitions of measures.

Measure	Definition
Range	The lowest VOT value subtracted from the highest VOT value for each target phoneme.
Accuracy	/b/: < 20 ms; /p/: ≥ 20 ms; /d/: < 30 ms; /t/: ≥ 30 ms.
Discreteness	The difference between the maximum voiced target VOT value and the minimum voiceless target value for the two places of articulation (Bailey & Haggard, 1973, 1980; Zlatin, 1972).
Overshoot	VOT values ≥ 100 ms for /p/ and /t/ were considered to represent overshoot (Gilbert, 1977; Macken & Barton, 1980).

characteristics could be accurately determined for all stops. Single-word utterances were targeted because of the young age of the participants and because emerging sounds are produced more accurately in single words than in continuous speech (Morrison & Shriberg, 1992). Furthermore, Lisker and Abramson (1967) have documented that VOT definitively separates the phonemic categories in adult citation form speech, whereas categories may show overlap in running speech. The stimuli consisted of black-and-white drawings of two minimal pairs of CV words: *boo–pooh* and *doe–toe.* Copies of the stimulus pictures were given to the primary caregivers 1 week prior to the initial testing session to familiarize the children with the picture labels. During recording, the researcher presented the picture stimuli to the children one at a time, in random order, using verbal prompts to elicit responses (e.g., "Who's this? It's Winnie the _____," or "What does the ghost say? The ghost says _____"). If the child did not label the stimulus following presentation, the experimenter provided an oral model as a reminder for the next elicitation turn. Productions in direct imitation of the experimenter's oral model were ignored given that imitation may affect production accuracy (Hodson & Paden, 1991).

Measurements and Analysis

Data were transferred to the Kay Pentax Computer Speech Laboratory (Model 4400) using a sampling rate of 11025 Hz. Inclusion criteria for measurement were that the production had a CV word shape, was produced in response to a known picture target, and contained a stop consonant produced at the expected place of articulation, and that the token was not acoustically distorted due to yelling or extraneous noise. In tokens that met these criteria, VOT was measured as the duration between the burst representing the release of the stop occlusion and the onset of glottal vibration (Francis et al., 2003; Lisker & Abramson, 1964), using a wide-band spectrogram and acoustic waveform. The final data set consisted of 6,339 tokens: an average of 19.81 productions of each of the four consonants, each child, and eight recording sessions (19.81 × 4 × 10 × 8 = 6,339). Fewer than 15 productions of a consonantal target were obtained in only eight cases, with the minimum of 11 obtained once. Thus, the goal of 15 productions of each stop per session was, on the whole, achieved.

All measures were originally made by the first author. Intrarater reliability was obtained many months after the original measures had been conducted by remeasuring a randomly selected 20 productions across all recording sessions for each child (a total of 200 productions). VOT values for the original and remeasured tokens were highly correlated (r = .998, p < .0001), with a mean difference of less than 1 ms (M = –0.54 ms, SD = 2.86 ms). Interrater reliability was obtained by having an acoustically trained graduate student remeasure a randomly selected 40 tokens across all recording sessions for each child (a total of 400 productions). VOT values for the original and remeasured tokens were again highly correlated (r = .996, p < .0001), with a mean difference of 2.06 ms (SD = 4.01 ms). These high levels of reliability and close agreement are comparable to what past authors have reported (Macken & Barton, 1980; Snow, 1997; Tyler & Saxman, 1991).

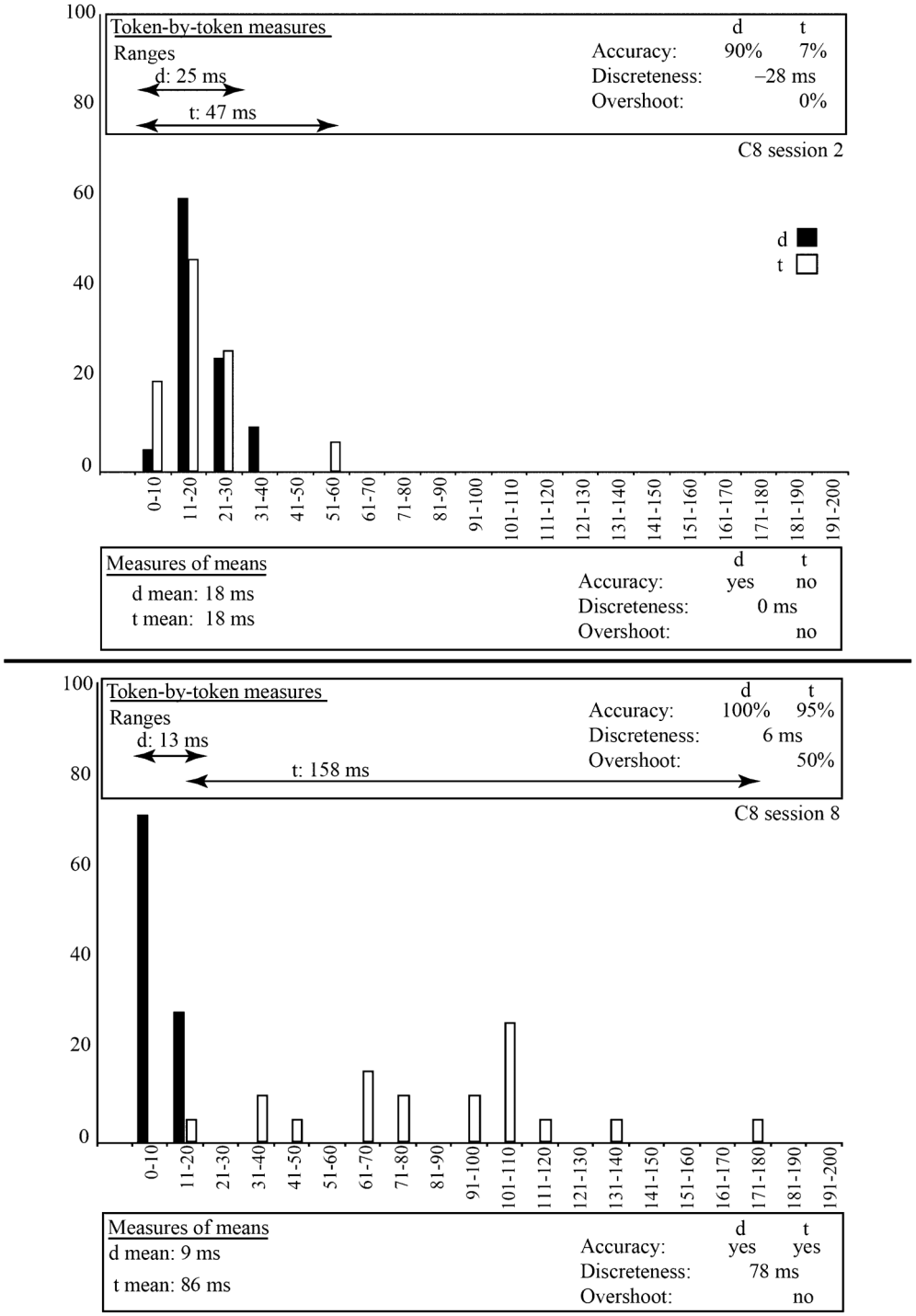
For each child, consonant, and session (10 children × 4 consonants × 8 sessions), four measures were obtained to chart VOT development: range, accuracy, overshoot, and discreteness. Accuracy, discreteness, and overshoot were computed on a token-by-token basis and also based on the session mean data for each consonant or consonant pair.[2] The definitions for the measures are provided in Table 2, and Figure 1 presents histograms from one early and one late session for Child C8 to show how the measures were obtained.

Range. Range, by definition, is a measure of the entire data distribution; it reveals the degree of token-to-token variability for each consonant. This measure was included because extensive past research indicates that high variability is a typical characteristic of child speech measures, including VOT (see, e.g., Eguchi & Hirsh, 1969; Kent, 1976). Zlatin and Koenigsknecht (1976) showed that VOT range sizes decreased as a function of age for voiceless aspirated stops; Eguchi and Hirsh (1969) found that VOT variability, measured by the standard deviation, decreased up to age 8 years for both English-voicing categories.

Accuracy. As indicated above, Macken and Barton (1980) and Kewley-Port and Preston (1974) considered adult VOT values and perceptual crossovers when describing what Macken and Barton termed "Stage II." For the current

[2]Koenig (2001) argued that child VOT values might be better characterized by median values than by means. Specifically, Koenig (2001) and Zlatin (1972) observed that children's VOT distributions could be skewed, in which case the median is closer to the bulk of the distribution. For this study, we did not evaluate medians because we were specifically interested in evaluating the staging provided by the classic VOT studies, and all of these studies considered means rather than medians.

Figure 1. Histograms from an early (top) and late (bottom) session for C8, showing the corresponding token-by-token and mean measures. The *y*-axes on the histogram show percentage occurrence; the *x*-axes show VOT in 10-ms bins.



work, voicing "accuracy" was formalized using a cutoff of 20 ms for the bilabials and a cutoff of 30 ms for the alveolars. According to this analysis, a target /b/ with a VOT of 40 ms would be classified as inaccurate because this value falls within the range of /p/ for adult productions. The specific values followed the adult production data presented by Lisker and Abramson (1964) and Zlatin and Koenigsknecht (1976) as well as the perceptual data of

Abramson and Lisker (1970).[3] Lisker and Abramson's stimuli included single words and sentence productions; Zlatin and Koenigsknecht's were single-word productions.

Figure 1 shows how accuracy measures can differ for token-by-token versus mean data. In the top histogram, token-by-token accuracy is high for /d/ and low for /t/: 90% of /d/ productions were in the short-lag range, and only 7% of /t/ productions were in the long-lag range. In this example, both /d/ and /t/ had mean values of 18 ms, so that, in aggregate, productions of /d/ would be judged as accurate (< 30 ms) and those of /t/ would be judged as inaccurate. Not surprisingly, the token-by-token analysis presents a somewhat more detailed picture of the child's productions. In the Results section, token-by-token accuracy is plotted as a percentage per phoneme, whereas mean data are characterized as "accurate" or "inaccurate."

Discreteness. Discreteness for the two places of articulation was obtained in ms for all productions of /b p/ and /d t/ following Zlatin and Koenigsknecht (1976). This token-by-token measure represents the overlap between voiced and voiceless distributions at each place of articulation. An analogous measure based on the closeness of the mean data was also obtained as the "mean discreteness" (equivalent to what Bailey & Haggard, 1980, referred to as "distinctiveness"; see also Bailey & Haggard, 1973). Note that discreteness, in contrast to all other measures, is based on voiced–voiceless pairs rather than on individual consonants. In the bottom panel of Figure 1, token-by-token discreteness of 6 ms reflects that the highest VOT value for a /d/ target was 14 ms, whereas the lowest VOT value for a /t/ target was 20 ms. In the histogram, this is evident as one token of /t/ falling into the 11–20 ms bin. For the mean data, discreteness of 78 ms reflects a wide separation between the voiced and voiceless means (9 ms and 86 ms, respectively; note that all values have been rounded to whole numbers).

Overshoot. Macken and Barton (1980) defined their Stage IIIA (the overshoot phase) as mean values being "considerably longer" than adult means (p. 48). Here, we formalized the overshoot boundary as being 100 ms. Adult speakers do not typically show mean values greater than 100 ms, nor a preponderance (> 50%) of such high values (Lisker & Abramson, 1964; Zlatin, 1974). For the token-by-token analysis, overshoot was determined as a percentage of productions of /t/ and /p/ exceeding 100 ms. Mean values greater than 100 ms were designated as showing overshoot as a binary decision, similar to Macken and Barton's (1980) designation of an overshoot phase. In the bottom panel of Figure 1, the token-by-token value of 50% overshoot for /t/ indicates that half of the child's productions exceeded 100 ms. The mean /t/ VOT of 86 ms was not high enough to show overshoot in the analysis of means.

Statistical analysis. In much past research, a statistically significant VOT difference between contrasting categories was used as the metric for determining that a child had acquired a voicing contrast (Bailey & Haggard, 1980; Lowenstein & Nittrouer, 2008; Macken & Barton, 1980; Snow, 1997; Tyler & Saxman, 1991; Zlatin & Koenigsknecht, 1976). As such, repeated-measures analyses of variance (ANOVAs)[4] were applied to each child's /b p/ and /d t/ token values for each session to (a) determine whether participants showed significant VOT differences at the two places of articulation and (b) permit an evaluation of how voicing development assessed using the four measures listed above compared with the appearance of significant VOT contrasts. Because this necessitated 160 ANOVAs (10 children × 8 sessions × 2 places of articulation), significance was assessed not only with a standard α-level of .05 but also with a conservative α of .0001. Cohen's *d* was also computed for each child's /b p/ and /d t/ pairs at the first and last sessions to provide an estimate of effect size (2 places of articulation × 2 sessions × 10 children = 40 values). Following Cohen (1988), effect sizes of $d = 0.2$ were considered small, effect sizes of $d = 0.5$ were considered medium, and effect sizes of $d = 0.8$ were considered large.

Results

Statistical Findings

Repeated-measures ANOVAs revealed that 91% (146 of 160) of the comparisons showed a significant difference between /b p/ and /d t/ distributions using a criterion of $p < .05$. Within this group, 93% (135 of 146) met the more stringent significance criterion of $p < .0001$ (see Table 3). These results, combined with the descriptive statistics, indicate that most mean VOTs were lower for /b/ than for /p/, and likewise for /d/ and /t/. In the first recording session, six of the 10 children exhibited significant /b p/ differences, and nine of the 10 showed significant /d t/ differences. By the study end, all 10 children showed significant differences between /b p/ and /d t/ pairs.

Cohen's *d* showed that the effect sizes for /b p/ and /d t/ VOT differences at the outset and the end of the study were usually large (35 of 40 cases); two were medium, and three were small, with one of the small effect sizes reflecting a 1-ms difference in the wrong direction (viz., VOT longer in /b/ than /p/). All effect sizes (20 of 20) were large in Session 8,

[3]Kewley-Port and Preston (1974) suggested 25 ms as the boundary between the "d-range" and the "t-range" (p. 209). Macken and Barton (1980) identified 0–20 ms as the short-lag range (p. 48) for bilabials and alveolars. Here we adopted different boundaries for the bilabials and alveolars following Lisker and Abramson's (1964) finding that VOT values for short-lag stops increase with more posterior places of articulation, as does the perceptual crossover (Abramson & Lisker, 1970).

[4]Some past studies (Koenig, 2001; Zlatin, 1972) have indicated that VOT data in children may not be normally distributed. To ensure that our results were robust in the face of possible threats to statistical validity, the full statistical treatment included several supplemental analyses. Shapiro-Wilks tests verified that the data were, on the whole, normally distributed, and the ANOVA results showed that within-child correlations over time were negligible. As an additional caution, two-sample *t* tests and nonparametric Wilcoxon Rank Sum tests were conducted. The results of these were almost entirely consistent with those of the repeated-measures ANOVAs.

Table 3. Results (*p* values) of repeated-measures ANOVAs for each child and session.

/b p/

Participant	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8
C1	**	**	**	**	**	**	**	**
C2	**	**	*	**	*	**	**	**
C3	.0747	*	**	*	**	**	**	**
C4	.6589	*	*	**	**	*	**	**
C5	**	**	**	**	**	**	**	**
C6	**	**	**	**	**	**	**	**
C7	**	**	**	**	**	**	**	**
C8	.1321	.2514	.9506	.3417	*	**	**	**
C9	.1844	**	**	**	**	**	**	**
C10	**	**	**	**	**	**	**	**

/d t/

Participant	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8
C1	**	**	**	**	**	**	**	**
C2	**	.0666	**	**	**	**	**	**
C3	**	**	**	.1081	**	**	*	**
C4	**	**	**	**	**	**	**	**
C5	**	**	**	**	**	**	**	**
C6	**	**	**	**	**	**	**	**
C7	**	**	**	**	**	**	**	**
C8	*	.1226	.7168	.7163	.5453	*	**	**
C9	.0643	**	**	**	**	**	**	**
C10	**	**	**	**	**	**	**	**

Note. For each place of articulation, VOT values for all voiced and voiceless tokens were entered into the ANOVA to evaluate whether the child had a significant VOT contrast in that session for /b p/ and /d t/. Asterisks indicate significant *p* values using a standard criterion of .05 (*) and a conservative criterion of .0001 (**). The conservative α represents a correction for the large number of ANOVAs (10 children × 8 sessions × 2 places of articulation).

and all but one were larger in Session 8 than in Session 1 (the exception being /d t/ for C7).

Range

Range sizes for each target phoneme and all children (C1–C10) are displayed in Figure 2. Developmental trends can be understood by comparing these data to adult ranges. Lisker and Abramson (1964) reported ranges of up to 180 ms for English /b/ and /d/ (including both prevoiced and voiceless unaspirated productions). For /p/ and /t/, adult speakers have ranges of up to 120 ms (Lisker & Abramson, 1964; Zlatin & Koenigsknecht, 1976). Figure 2 shows that the children often had larger ranges than adult norms, for both voiced and voiceless targets. In general, the children demonstrated larger ranges for /p t/ than for /b d/ (e.g., C1–C6 produced at least one /p/ or /t/ range between 300 ms and 400 ms). However, in some cases, ranges for /b d/ could also be as large as 300–400 ms (e.g., C1). Thus, one aspect of acquiring mature VOT patterns is reducing the large ranges found at young ages.

Accuracy

Token-by-token accuracy measures, based on adult VOT values for English (cf. Table 2), are graphed for all 10 children in the line plots of Figure 3 (the grids show results from the analysis of means, discussed below). Accuracy varied markedly across children, with some showing fairly high accuracy throughout (e.g., C1) and others producing rapid improvements from session to session (e.g., C9). Some children showed periods of inverse accuracy trends for the two voicing categories; as one pair (/b d/ or /p t/) improved in accuracy, the other became less accurate (e.g., C3 and C8).

To compare token-by-token accuracy across consonants, all 80 accuracy percentages for each phoneme were averaged across sessions and children. These results showed the least accuracy for /b/ (74%) followed by /t/ (83%), /d/ (91%), and /p/ (92%). The low accuracy for /b/ was unexpected, as previous studies have indicated that the short-lag category is well established at an early age (Enstrom, 1982; Kewley-Port & Preston, 1974; Lowenstein & Nittrouer, 2008; Preston & Yeni-Komshian, 1967; Preston et al., 1968; Whalen et al., 2007). Further, /b/ had substantially lower accuracy than did /d/ in some children (particularly large differences are evident for C4, C6, and C7). One possible explanation for this was that a lower VOT cutoff was used to determine accuracy for the bilabial place of articulation (20 ms vs. 30 ms for the alveolars; cf. Footnote 3). To assess whether a higher threshold would yield increased accuracy for /b/, the VOT values were reanalyzed for two randomly chosen children, C3 and C10, using 30 ms as the upper boundary (i.e., the same as for /d/). The results of this showed that accuracy was about 10% higher on average using the 30-ms cutoff; of the 16 comparisons (2 children × 8 sessions),

Figure 2. Range in milliseconds of VOT values for /b/, /d/, /p/, /t/ for each child (C1–C10, plotted individually) over eight sessions (/b d/ = solid lines; /p t/ = dashed lines).

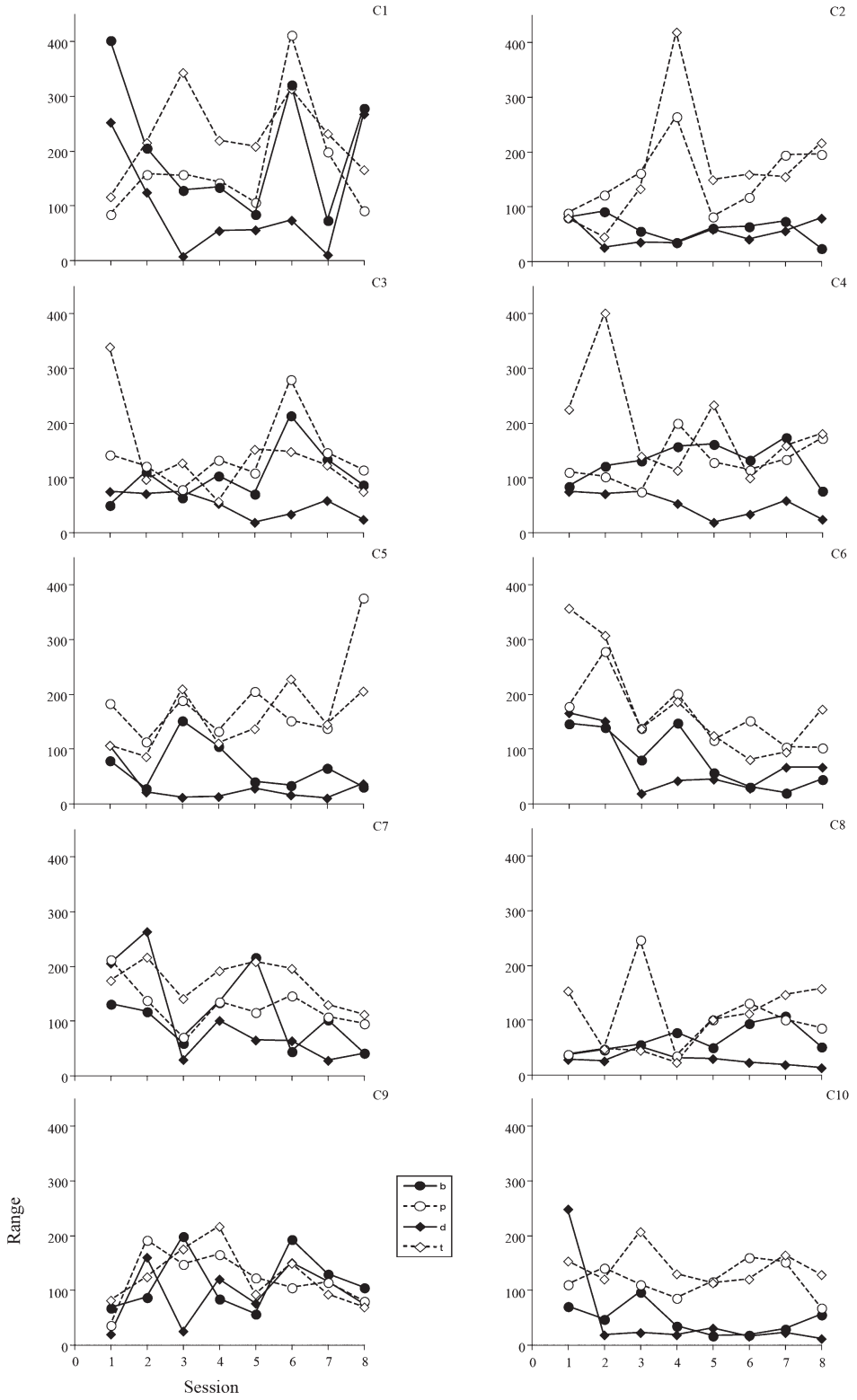
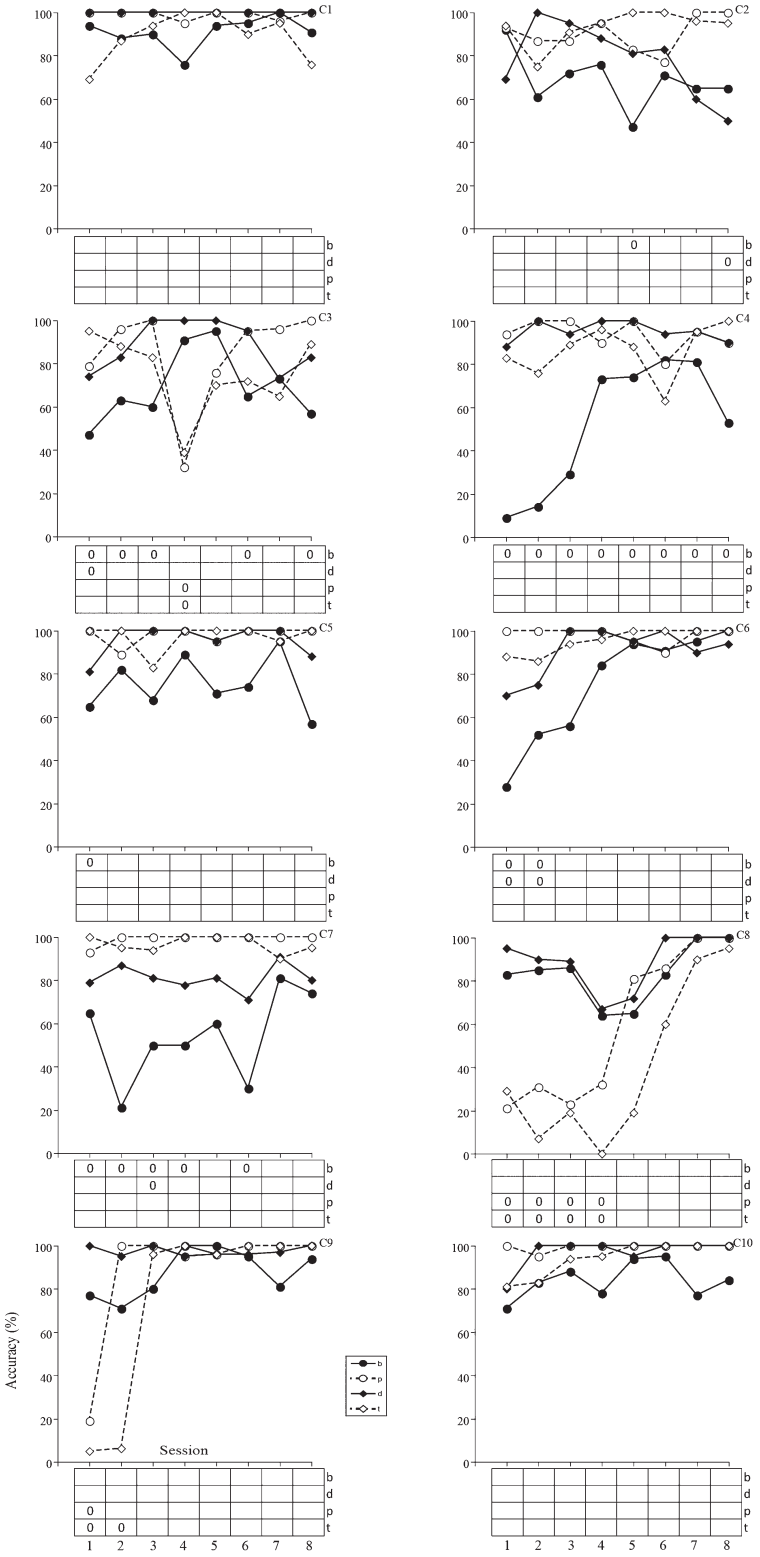


Figure 3. Accuracy of stop consonants for each child over eight sessions, measured according to whether VOT values fell within adult boundaries. The line plots in each panel show token-by-token accuracy percentages (/b, d/ = solid lines; /p, t/ = dashed lines). The grids below each panel indicate whether session mean values were accurate or inaccurate (with inaccurate means marked as ''0'').



three showed no difference, six differed by ≤10%, and only one showed a difference larger than 20%. Thus, the results using the higher threshold still yielded more voicing errors for /b/ than expected given previous literature.

It is instructive to compare the token-by-token measures of accuracy shown in Figure 3 with the patterns of statistical significance from the ANOVAs given in Table 3. Out of 146 /b p/ and /d t/ comparisons that reached statistical significance, 9% were associated with 100% token-by-token accuracy of both phonemes in the pair. In 84 of 160 significant /b p/ and /d t/ comparisons, accuracy was 75%–99%. Finally, 48 of 160 /b p/ or /d t/ comparisons reached statistical significance and were associated with < 75% accuracy. That is, about a third (48/160 = 30%) of the statistically significant differences co-occurred with *inaccuracy* rates of 25% or more when the data were evaluated on a token-by-token basis.

The grids in Figure 3 present mean accuracy results for each child, with 0 indicating inaccurate session means. Overall, accuracy based on mean data was higher than that obtained in the token-by-token analysis (compare, for example, the token-by-token and mean analyses for /b/ in speakers C2 and C5). Using the means, children showed accurate production of /b/ in 58/80 sessions (73%), followed by /t/ (91%), /p/ (93%), and /d/ (94%), respectively. Mean VOT values for C1 and C10 show perfect accuracy across all targets and sessions, suggesting that these children were fully adultlike from the study onset. Of the remaining children, mean values were accurate in 88% or more targets and sessions for C2, C5, C6, and C9; 84% for C7; and 75% for C3, C4, and C8. The phoneme most frequently in error differed across children (e.g., /b/ for C4 in all sessions vs. /p t/ in the first four sessions for C8).

The consonantal accuracy percentages based on individual tokens versus session means were similar (within 4%) with the exception of /t/, where the mean data showed 91% accuracy, compared to 83% in the token-by-token analysis. Some speaker patterns were also consistent across the two analyses; for example, C3 had reduced accuracy for /p t/ in Session 4, coinciding with increased/stable accuracy in /b d/. C8 had poor accuracy for /p t/ during the first half of the study in both analyses. However, her mean data suggest that /b d/ remained fully accurate throughout the study, whereas the token-by-token analysis shows a dip in accuracy in Sessions 4–5, as /p t/ were improving. The token-by-token measures also reveal that improved accuracy of /p/ preceded that of /t/ by 2 weeks. This detail is not evident in the VOT means, where /p/ and /t/ appear to improve simultaneously.

Discreteness of Voicing Categories

Discreteness is a measure of overlap between voiced–voiceless categories at a given place of articulation. Each child's token-by-token discreteness data are displayed in the line plots of Figure 4. None of the children demonstrated discrete voicing categories for either place of articulation at the onset of the study (i.e., all values are negative, indicating overlap). Discreteness improved over time, with five out of 10 children showing nonoverlapping /b p/ and /d t/ distributions by the end of the study. Discreteness also seems to develop with age: One of the five children age 31 months, two of the three children age 32 months, and both of the two children age 33 months produced nonoverlapping distributions for both contrasts at the end of the study. Nevertheless, after 16 weeks, overlap was still observed for /b p/ in five children (C1–C4 and C7), and /d t/ in four children (C1–C3 and C7). Moreover, at the last session, many children with discrete distributions did not demonstrate a VOT separation of > 30 ms between the highest valued voiced target and the lowest voiceless target, the degree of separation reported for adults (Zlatin & Koenigsknecht, 1976). Four children showed this adultlike level of discreteness for /b p/ (C5, C6, C8, C10). None of the children did for /d t/.

Whereas the token-by-token analysis frequently showed overlap (75% of /b p/ distributions; 70% of the /d t/ distributions), the mean VOT data rarely showed lack of discreteness or overlap (marked as O in the grids): It occurred in only 3% of the /b p/ pairs (Sessions 1 and 3 of C8) and 4% of the /d t/ pairs (Sessions 2, 3, and 4 of C8). Mean VOT values for all the other children are discrete, although there are some cases (marked with C, for "close") where values are separated by < 30 ms (nine sessions for /b p/ and eight for /d t/). Cases that are not marked with O or C have > 30 ms mean differences: 86% (69/80) of the /b p/ pairs and 86% (69/80) of the /d t/ pairs. Thus, discreteness calculated from the means is much greater than that calculated from individual tokens.

Overshoot

The percentages of VOT values ≥ 100 ms (i.e., the percentages of overshoot) calculated on a token-by-token basis are shown for each child in the line plots of Figure 5. These data show that all 10 children underwent a period of exaggerated VOTs (≥ 50% of values ≥ 100 ms) for /p/ and/or /t/ at some point during the study. Three children (C3, C6, and C7) were in an overshoot phase for one or both places of articulation at study onset. All of these children had lower overshoot by the last recording session. Two children (C8 and C10) did not show overshoot until the latter half of the study, suggesting that they began the study in an earlier stage of development for this measure. For several children, the VOT values for /p/ and /t/ consistently moved in the same direction (C5, C6, C8, C9, and C10). In two of these (C5 and C9), movement of VOT values for /p/ preceded that for /t/. In other cases, the VOT values for /p/ and /t/ moved in opposition to one another; see, for example, Sessions 4 and 5 for C1 and Sessions 3 and 4 for C7.

Cases of VOT overshoot in session means are indicated in the grids of Figure 5 with an "x". These calculations show an overshoot period for eight of the 10 children during the study. Three children (C5, C6, and C7) began the study in the overshoot phase; others (C1, C2, C4, C9, and C10) did not show overshoot until later sessions. Similar to the token-by-token analysis, the mean overshoot data reveal some cases of /p/ overshoot preceding /t/ (viz., for C9) and different patterns over time for /p/ and /t/ (C1). Finally, the total

Figure 4. Discreteness values for /b p/ and /d t/ for each child over eight sessions indicating overlap between the phoneme pairs. The line plots in each panel indicate overlap as a negative value and separation between the pairs as a positive value. The grids below indicate overlap between session mean values (O = overlap; C = overlap < 30 ms).

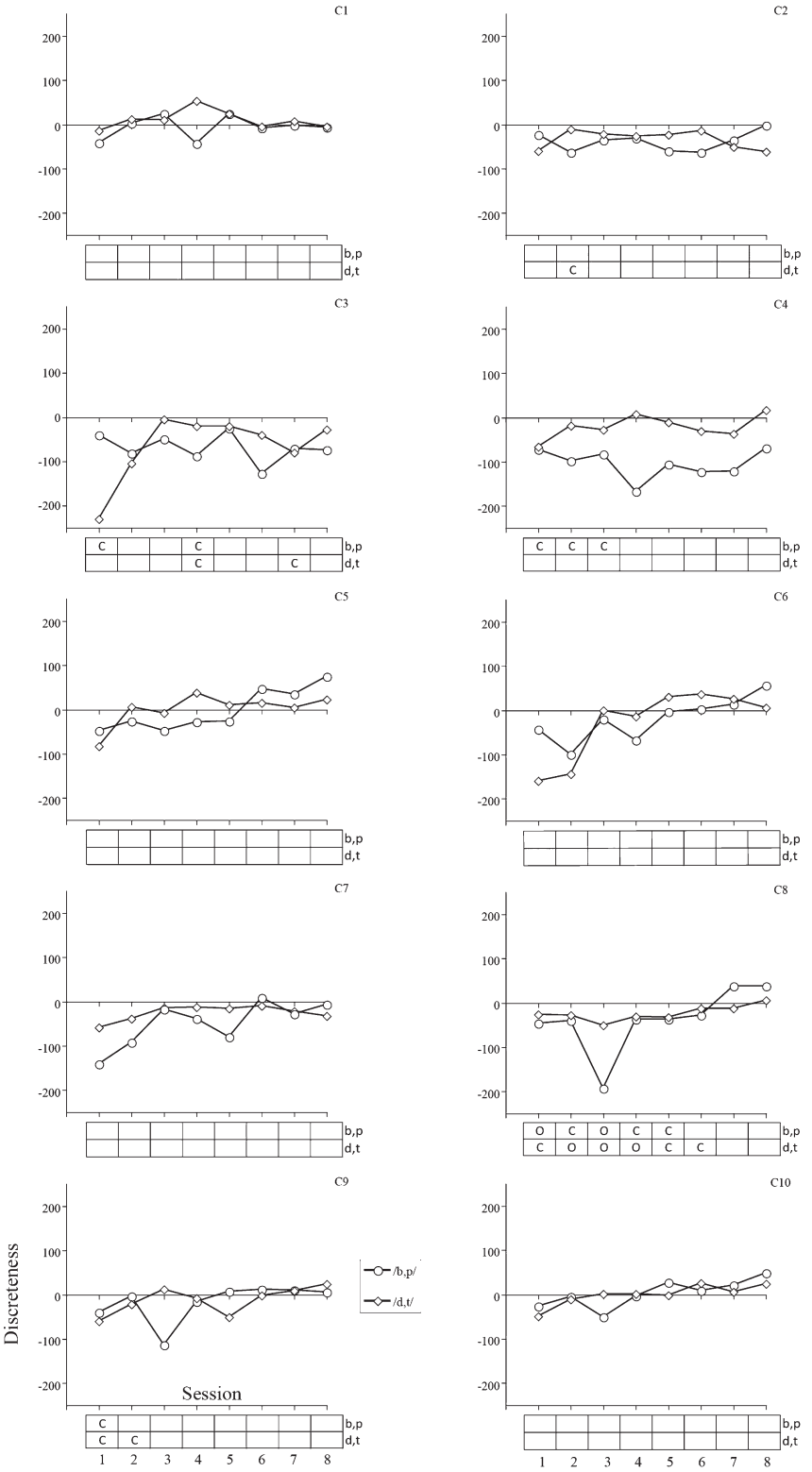
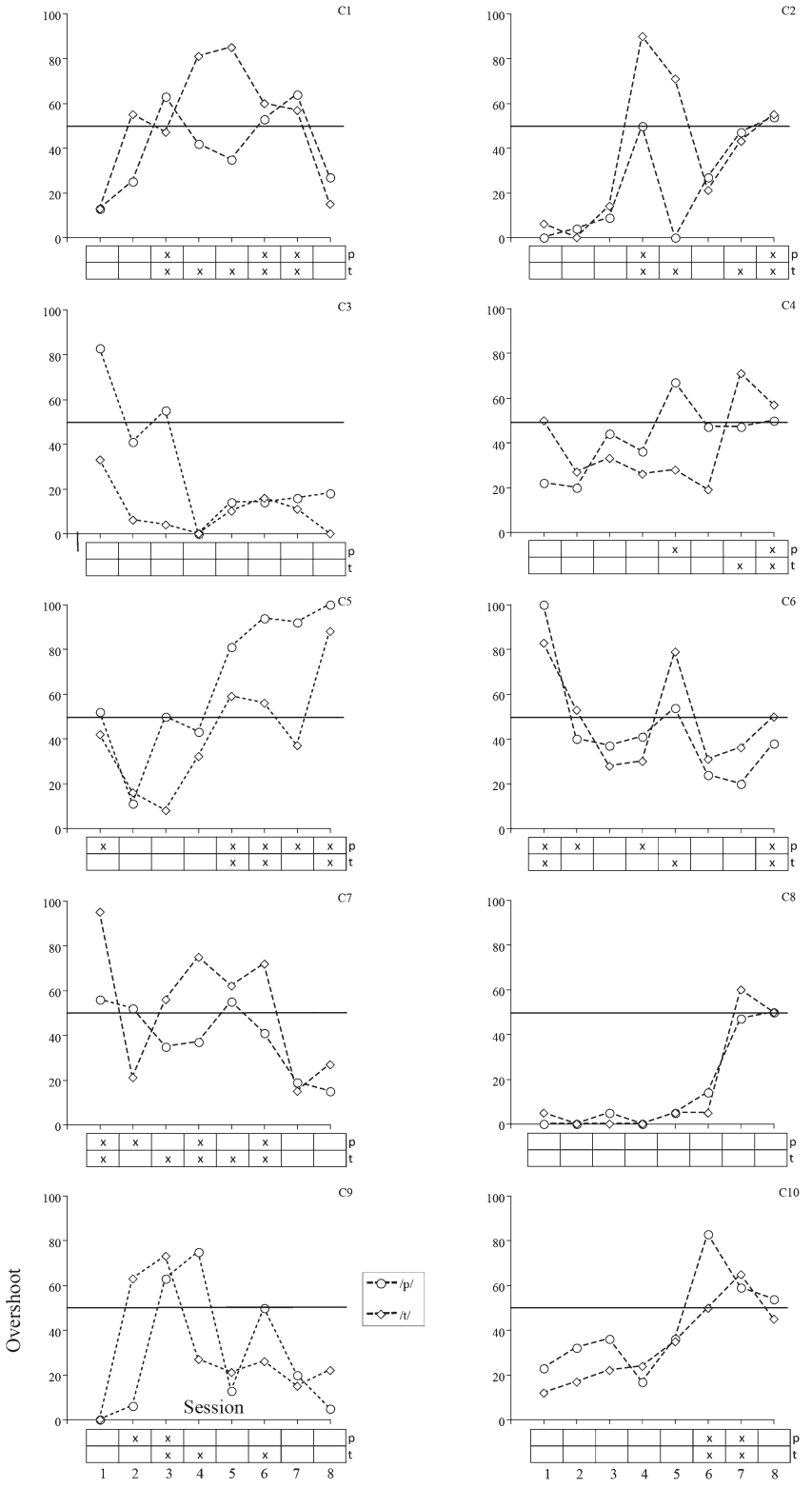


Figure 5. Overshoot of voiceless aspirated stop consonants /p/ and /t/ for each child over eight sessions. The line plots in each panel show the percentages of /p/ and /t/productions >100 ms. The grids below indicate whether session means for /p/ and /t/ were > 100 ms (with overshoot marked by an ''x'').



proportion of overshoot was similar for the analyses based on individual tokens and based on means. Averaging the overshoot percentages over sessions and children in the token-by-token analysis showed 33% of /p/ and 38% of /t/ distributions in overshoot. The analysis of means showed 30% of /p/ session means (24/80 sessions) and 33% of /t/ session means (26/80) with values > 100 ms.

However, there are some differences between analyses of individual tokens and of means. Based on the means, two children (C3 and C8) demonstrated no overshoot over the study period, but the token-by-token analysis reveals that they, like the other children, did evidence overshoot at some point for at least one phoneme. Further, the mean overshoot analysis does not reveal gradual changes toward longer VOT values over sessions. In the token-by-token analysis, some children do appear to enter overshoot quite suddenly (e.g., C2), but others show gradual lengthening of their /p t/ VOTs (e.g., C5). In the analysis of means, both C2 and C5 appear to have entered the overshoot phase abruptly.

As noted above, the findings from both the analysis of means and of individual tokens showed surprisingly low accuracy for /b/, reflecting VOT values that were longer than the 20-ms cutoff value established from the literature. For example, C3's accuracy for /b/ was below 50% at the beginning of the study (cf. Figure 3 line plot). At that time, his /p/ values were in overshoot (see Figure 5 line plot). During Sessions 1–4, as /b/ became more accurate, /p/ values became substantially shorter (even falling into the short-lag range). In Session 4, with no overshoot of /p/, and fairly high accuracy for /b/, the discreteness data (see Figure 4) showed increased overlap. Thus, during the first half of the study, productions of /b/ and /p/ were changing simultaneously. Another example of an inverse relationship between the development of /p/ versus /b/ and/or /t/ versus /d/ can be seen in Figure 3 for C8 (Sessions 4–5), with voiced targets becoming less accurate as the voiceless targets improved.

Discussion

Four Measures of Assessment for Voicing Acquisition

The purpose of this study was to evaluate the process of voicing acquisition in English-speaking children with typically developing phonology, using a battery of measures to characterize VOT patterns over time. A longitudinal design with frequent sampling intervals was used to capture rapid developmental changes. Some examples of dramatic change over 2-week periods were in fact observed for all measures. In other cases, development proceeded gradually over many sessions. Developmental changes included regression (e.g., reduced accuracy) as well as improvement.

The results show that the use of multiple measures increases our understanding of the pattern(s) of voicing acquisition (cf. Bailey & Haggard, 1980; Koenig, 2001). Measures of range, accuracy, discreteness, and overshoot provide an organizational scheme for analyzing data on a token-by-token basis. The four measures also complement each other and can be used to assess where a child is regarding VOT development. For example, short-lag VOT values, a small range, good accuracy for /b d/, poor accuracy for /p t/, extensive overlap, and absent or minimal overshoot indicate early stages in the acquisition of contrastive voicing. The developmental progression for the four measures assessed here is summarized in Table 4. The staging follows Macken and Barton's (1980) voicing acquisition schedule which, crucially, included an overshoot phase. The current data appear to support the universality of an overshoot phase; the added information on range, accuracy, and discreteness add detail to the developmental progression. An important caveat is that we do not intend to suggest that all four measures will necessarily move in unison; development is a complex and multifaceted process, and movement from one stage to another may vary somewhat across measures.

Table 4. Revised stages of English-voicing acquisition, following Macken and Barton (1980), adding the four additional measures suggested in the text.

Stage	Approximate age range (years;months)	VOT values	Range size	Accuracy	Discreteness	Overshoot
I	Typically < 2;0	Short lag	Small (≤ 100 ms)	/b d/ good; /p t/ poor	Low: Extensive overlap between /b p/ and /d t/ pairs	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 into the long-lag range; /p t/ good	Moderate–High: Overlap still present but decreased from Stages I–II	Present for /p/ and/or /t/
IIIB	2;6–3;0+	VOT values in the short-lag and long-lag range	Decreased from Stage IIIA but larger than adults	/b d/ good; /p t/ good	High: No overlap present between the categories; separation between voicing categories is < 30 ms as seen in adults	Absent

The large ranges observed here reflect widespread token-to-token variability in VOT, consistent with past studies (e.g., Eguchi & Hirsh, 1969; Gilbert, 1977; Lowenstein & Nittrouer, 2008; Zlatin & Koenigsknecht, 1976). It is instructive to evaluate the current range values in light of other measures. For example, C4 showed an increase in the range for /t/ in Session 2 (see Figure 2). This corresponded to a decrease in accuracy in the token-by-token analysis (see Figure 3). In this case, the expanded /t/ range reflected productions falling into the short-lag range (see also C8's Session 3 for /p/). On the other hand, C2 showed a large increase in range for /t/ in Session 4 that was not accompanied by decreased token-by-token accuracy; rather, it reflected overshoot (see Figure 5). These examples show how the measures complement each other.

The current discreteness data show extensive overlap between voicing categories in early stages, in accord with Macken and Barton's (1980) descriptions of Stages I and II. In later stages, overlap subsides. The token-by-token analysis showed that 50% of the children produced discrete VOT distributions at study closure. Nevertheless, many children with discrete distributions did not demonstrate a separation of > 30 ms between the highest voiced target and the lowest voiceless target; in other words, they were not yet adultlike. The group data of Zlatin and Koenigsknecht (1976) suggested that, on average, discreteness in 6-year-olds still did not approach adultlike levels. It would be interesting, in future work, to trace the development of discreteness for individual children over a larger age range.

Measures of accuracy and overshoot may also need to be considered in relationship to one another in order to be correctly interpreted. The data suggest that, in some children, reduced accuracy for /b d/ paralleled the appearance of an overshoot phase for /p t/. Token-by-token accuracy for /b/ was unexpectedly poor (see Figure 3), indicating that many VOTs were longer than adult target values. We note first that the slow speech rate typically observed in children (Kent, 1976; Kent & Forner, 1980; Smith, 1978) does not seem to provide a straightforward explanation for this finding. Data from adult speakers show that slower speech rate affects VOT mainly by lengthening voicing lag durations for voiceless aspirated stops; the voiceless unaspirated category varies little (e.g., Kessinger & Blumstein, 1997; Magloire & Green, 1999; Miller & Volaitis, 1989). Thus, slow speech rate, per se, does not seem to lead to longer VOT durations of voiceless unaspirated stops. Moreover, the longer VOT values are observed only for /b/ (not /d/), and appear to be a transient phenomenon. Macken and Barton (1980) noted that Stage IIIA could be characterized by longer-than-adult VOTs for voiced as well as voiceless targets, and the data of Lowenstein and Nittrouer (2008) also show a slight increase in the mean values of voiceless unaspirated stops as the means of voiceless aspirated stops increase. The current data are consistent with those earlier findings. The children who produced the least accuracy for /b/ at the study's onset (C3, C6, and C7), were also those who were in the overshoot phase for /p/ and/or /t/ (cf. Figure 5); C7 in particular showed "overshoot" of /b d/ in parallel with that of /p t/. Further evidence is seen in the VOT values of C2 and C5, who, over the study's duration, moved into an overshoot phase for /p/, which corresponded closely in time to longer (and less accurate) VOT values in /b/. In the current data, the /d/ distributions do not show the same upward drift as seen for /b/. VOT values for /d/ indicated only four instances out of 80 where 30% or more of the distribution was at or above 30 ms. A post hoc analysis using a more lenient criterion value for /b/ (30 ms, equivalent to that for /d/) did not seem to fully account for the low accuracy obtained for /b/. Interestingly, Okalidou, Petinou, Theodorou, and Karasimou (2010) recently observed that young learners (ages 2;0–2;5) of Standard Greek, which contrasts voiced and voiceless unaspirated stops, produced long-lag values of VOT for target /p/ (i.e., [p]) but not /t/ or /k/. More work is needed to understand whether low accuracy in voiceless unaspirated bilabial stops is an accidental result of these two studies, or represents a genuine developmental phenomenon.

The current study followed the classic work of Kewley-Port and Preston (1974) and Macken and Barton (1980) in using adult VOT production data and perceptual boundaries to evaluate a child's production accuracy. Those authors proposed that adults would perceive a child's contrast following the nature of the adult contrast. It is possible that this assumption needs to be evaluated more fully. Although Macken and Barton (1980) stated that they did not believe they could perceive a voicing contrast in Stage II, they acknowledged that secondary cues such as fundamental frequency or burst amplitude could contribute to adult categorization, and Kewley-Port and Preston (1974) observed that some child productions in the short-lag VOT range could still be categorized as /t/. As such, future work should evaluate adult perception of child stop voicing in more detail.

Token-by-Token Compared to Mean Analyses

As outlined in the introduction, many researchers have sought to determine the schedule of voicing acquisition relying largely on parametric statistics and mean values to compare typically developing children to adult norms. Past authors have also pointed out that child productions may remain nonadultlike after means approach adult norms and statistically significant differences are observed between contrastive categories (e.g., Kewley-Port & Preston, 1974; Lowenstein & Nittrouer, 2008; Macken & Barton, 1980; Zlatin & Koenigsknecht, 1976). The current study sought to compare results of token-by-token analyses with analyses based on mean values to evaluate the extent to which reliance on mean values might mask or eliminate some relevant information and potentially lead researchers to arrive at different conclusions about early voicing development. The results show that (a) some developmental stages may not be evident from mean data and/or patterns of statistical significance, and (b) using a variety of measures demonstrates how children's productions continue to differ from adults' after a statistically significant VOT contrast is achieved. One example of missing a developmental stage by relying on an analysis of means comes from Child C8. Mean VOTs

did not show evidence for overshoot (Stage IIIA), whereas the token-by-token analysis did (compare her line plots and grids in Figure 5). Another example comes from C1. Following the staging in Table 4, in the first session of this study this child appeared to be beyond Stage II according to the analysis of means, because his means were accurate and discrete. The token-by-token analysis, however, indicated that his distributions were not discrete, and accuracy was still low for /t/.

Snow (1997), in a study of 10 children, reported that children achieved bimodal distributions by age 24 months. In the current work, all 10 children produced a statistically significant difference between the voicing categories by age 30 months. Over the study as a whole, a very high proportion (91%) of the ANOVA results were significant, and effect sizes for this difference were usually large. However, the token-by-token analysis based on the four proposed measures shows clearly that statistically significant VOT differences, on their own, do not indicate adultlike production. After achieving statistically significant differences between voicing categories (see Table 3), all children in the current study showed continued VOT development in the form of diminishing ranges, increasing accuracy, and increasing discreteness. Such changes were particularly evident in the token-by-token analyses. For example, fully 30% of contrasts with statistically significant VOT differences were characterized by inaccuracy rates of 25% or more.

The presence of significant differences in voiced–voiceless VOT values in combination with nondiscrete distributions provides evidence for the existence of a covert contrast or subperceptual distinction (i.e., Macken & Barton's Stage II) during voicing acquisition. For example, C8 did not demonstrate significant VOT differences for /b p/ at the study outset (see Table 3), and was therefore in Stage I. In Session 5, she produced significant VOT differences, but token-by-token accuracy for /p/ did not yet approach 100% (see Figure 3), and her distributions were not discrete on a token-by-token basis (see Figure 4), putting her in Stage II. The production of significant VOT differences suggests that in these sessions, the child did not have errors in target awareness, but rather errors due to developing levels of proficiency in motor learning (Grigos, 2009; Whiteside, Dobbin, & Henry, 2003). Young children are constantly adapting and refining their speech production to adjust to anatomical and physiological maturation (Kent, 1976); therefore, some voicing errors may be explained by phonetic exploration and/or an inability to maintain previously learned motor speech behaviors as new patterns are added (see also Macken & Barton, 1980).

The token-by-token analysis revealed overshoot, or Macken and Barton's Stage IIIA, in all 10 children (compared to eight in the analysis of means). Zlatin and Koenigsknecht (1976), on the basis of a cross-sectional analysis, did not report evidence of overshoot and concluded that children demonstrate a consistent progression towards adultlike VOT values. Lowenstein and Nittrouer (2008), evaluating mean values for seven children 15–28 months of age recorded bimonthly, came to a similar conclusion (cf. Nittrouer, 1993). It appears that overshoot may be missed by analyses of mean data or group data. This shows how researchers using different data collection and analysis methods could come to differing conclusions regarding stages of voicing acquisition.

In sum, the current study carried out frequent longitudinal sampling, obtained many repetitions of each consonant, and implemented multiple assessment measures to evaluate how the identification of developmental patterns could vary as a function of methodological differences, and in particular, differences related to data reduction. Our data suggest that developmental studies using mean values and/or cross-sectional sampling may fail to reveal some aspects of acquisition. These include nondiscrete categories, overshoot of the voiceless aspirates (Macken & Barton's Stage IIIA), and some subtle trends (e.g., inverse accuracy patterns between /b d/ and /p t/ for C8). Finally, in longitudinal studies, the use of mean data may not differentiate between gradual and abrupt changes. Thus, token-by-token analyses provide a more complete picture of development.

Although the sample size in the current study was larger than that in most previous studies, this was still not a large-scale study. Further, children were assessed over a rather short time frame. Factors of sample size and study duration may also affect how well researchers can determine the details of voicing acquisition. Following a larger number of children would improve understanding of the range of child-specific patterns, and what is universal in development. Recording children over a longer time frame (e.g., a year as opposed to 4 months) would also lend greater insight into how children progress through acquisitional stages.

Conclusions

Consistent with past research, the 2-year-olds in this study largely showed statistically significant differences between voicing categories. Of more interest here was to characterize how young children's productions continue to differ from adults after statistically significant VOT values are achieved. Such differences were observed in the form of larger ranges and lower rates of accuracy and discreteness. The data, and particularly the token-by-token analyses, support Macken and Barton's (1980) staging, including Stage II (a subperceptual distinction within the short-lag range) and Stage IIIA (an overshoot phase for voiceless aspirates). A comparison of analyses based on session means versus token-by-token measures indicates that analyses relying entirely on session means mask some developmental patterns. Studies seeking to characterize VOT in development should ideally carry out multiple measures on a token-by-token basis.

Acknowledgments

Preliminary portions of this work were presented at the 147th meeting of the Acoustical Society of America (New York City, 2004). This research was completed as part of a doctoral dissertation by the first author at New York University in the Department of Speech Pathology and Audiology directed by Harriet Klein. We gratefully acknowledge Harriet Klein, Nassima Abdelli Beruh,

Joseph Attanasio, Jonathan Preston, Julia Irwin, and Tara McAllister Byun for their comments on earlier versions of this work. We would also like to thank Bruce Diamante, Paul Arcell, and Tom Deutsch of Kay Pentax for their technical support; Stephanie Klopfer for her assistance with statistical analysis; and Heidi Fuld, Kathryn Galbraith, Elina Izbinsky, and Lindsey Syvertsen for assistance with data management. Finally, this work would have not been possible without the 10 children and their families who volunteered to be part of the study.

References

Abramson, A. S., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In B. Hála, M. Romportl, & P. Janota (Eds.), *Proceedings of the 6th International Congress of Phonetic Sciences, Prague 1967* (pp. 569–573). Prague, Czech Republic: Academia Publishing House, Czechoslovak Academy of Sciences.

Bailey, P. J., & Haggard, M. P. (1973). Perception and production: Some correlations on voicing of an initial stop. *Language and Speech, 16,* 377–396.

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.

Barton, D., & Macken, M. A. (1980). An instrumental analysis of the voicing contrast in word-initial stops in the speech of four-year-old English speaking children. *Language and Speech, 23,* 159–169.

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.

Bond, Z. S., & Wilson, H. F. (1980). Acquisition of the voicing contrast by language-delayed and normal-speaking children. *Journal of Speech and Hearing Research, 23,* 152–161.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngologica, 57,* 1–51.

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.

Enstrom, D. H. (1982). Infant labial, apical and velar stop productions: A voice onset time analysis. *Phonetica, 39,* 47–60.

Francis, A. L., Ciocca, V., & Yu, J. M. C. (2003). Accuracy and variability of acoustic measures of voicing onset. *The Journal of the Acoustical Society of America, 113,* 1025–1032.

Gandour, J., Petty, S. H., Dardarananda, R., Dechongkit, S., & Mukngoen, S. (1986). The acquisition of the voicing contrast in Thai: A study of voice onset time in word-initial stop consonants. *Journal of Child Language, 13,* 561–572.

Gilbert, J. H. V. (1977). A voice onset time analysis of apical stop production in 3-year-olds. *Journal of Child Language, 4,* 103–113.

Goldman, R., & Fristoe, M. (2000). *Goldman Fristoe Test of Articulation* (2nd ed.). Circle Pines, MN: AGS.

Grigos, M. I. (2009). Changes in articulator movement variability during phonemic development: A longitudinal study. *Journal of Speech, Language, and Hearing Research, 52,* 164–177.

Hodson, B., & Paden, E. (1991). *Targeting intelligible speech* (2nd ed.). Austin, TX: Pro-Ed.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research, 19,* 421–447.

Kent, R. D., & Forner, L. L. (1980). Speech segment duration in sentence recitations by children and adults. *Journal of Phonetics, 8,* 157–168.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics, 25,* 143–168.

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. (2001). Distributional characteristics of VOT in children's voiceless aspirated stops and interpretation of developmental trends. *Journal of Speech, Language, and Hearing Research, 44,* 1058–1068.

Lisker, L., & Abramson, A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–442.

Lisker, L., & Abramson, A. (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. *The Journal of the Acoustical Society of America, 124,* 1180–1191.

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.

Magloire, J., & Green, K. P. (1999). A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish. *Phonetica, 56,* 158–185.

Menyuk, P., & Klatt, M. (1975). Voice onset time in consonant cluster production by children and adults. *Journal of Child Language, 2,* 223–231.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46,* 505–512.

Morrison, J. A., & Shriberg, L. D. (1992). Articulation testing versus conversational speech sampling. *Journal of Speech and Hearing Research, 35,* 259–273.

Nittrouer, S. (1993). The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech and Hearing Research, 36,* 959–972.

Okalidou, A., Petinou, K., Theodorou, E., & Karasimou, E. (2010). Development of voice onset time in standard-Greek and Cypriot-Greek-speaking preschoolers. *Clinical Linguistics & Phonetics, 24,* 503–519.

Preston, M. S., & Yeni-Komshian, G. (1967). Studies on the development of stop consonants in children. *Haskins Laboratories Status Reports on Speech Research, SR11,* 49–53.

Preston, M. S., Yeni-Komshian, G., Stark, R. E., & Port, D. K. (1968). Developmental studies of voicing in stops. *Haskins Laboratories Status Reports on Speech Research, SR13/14,* 181–184.

Smith, B. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics, 6,* 37–67.

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.

Tyler, A. A., & Edwards, M. L. (1993). Lexical acquisition and acquisition of initial voiceless stops. *Journal of Child Language, 20,* 253–273.

Tyler, A. A., & Saxman, J. H. (1991). Initial voicing contrast acquisition in normal and phonologically disordered children. *Applied Psycholinguistics, 12,* 453–479.

Whalen, D. H., Levitt, A. G., & Goldstein, L. M. (2007). VOT in the babbling of French- and English-learning infants. *Journal of Phonetics, 35,* 341–352.

Whiteside, S. P., Dobbin, R., & Henry, L. (2003). Patterns of variability in voice onset time: A developmental study of motor speech skills in humans. *Neuroscience Letters, 347,* 29–32.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale—3*. 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.

Zlatin, M. A. (1974). Voicing contrast: Perceptual and productive voice onset time characteristics of adults. *The Journal of the Acoustical Society of America, 56,* 981–994.

Zlatin, M. A., & Koenigsknecht, 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.