

The Role of Intermediate Productions and Listener Expectations  
on the Perception of Children's Speech

Sarah K. Schellinger<sup>1,2</sup>, Jan Edwards<sup>2</sup>, and Benjamin Munson<sup>1</sup>

<sup>1</sup>University of Minnesota, <sup>2</sup>University of Wisconsin-Madison

Correspondence:

Sarah K. Schellinger

Department of Speech-Language-Hearing Sciences

University of Minnesota

115 Shevlin Hall, 164 Pillsbury Drive

Minneapolis, MN 55455

(612) 624-3322, Fax: (612) 624-7586

Schel024@umn.edu

### Abstract

*Purpose:* This paper examined whether naïve listeners could perceive phonetic detail in children's productions of s and T, and whether their perception of /s/ and /θ/ could be biased by their belief about the child's overall speech-production ability.

*Method:* In experiment 1, listeners provided judgments of children's productions of s and T using a visual analog scale (VAS). In Experiment 2, different listeners provided 'correct' and 'incorrect' judgments of the same tokens in a task in which they were led to believe that some children were older and had more-accurate speech, and others were younger and had less-accurate speech. For Experiment 1, linear regression model was used to determine the relationship between VAS responses and psychoacoustic characteristics of the stimuli. For Experiment 2, within-subjects comparisons of accuracy judgments for the two conditions were conducted.

*Results:* In Experiment 1, VAS judgments showed that listeners were able to perceive fine phonetic detail in children's productions, including differences between correct productions and clear substitutions. In Experiment 2, listener bias was found to have a small influence on listener judgments.

*Conclusions:* Naive listeners are able to perceive fine phonetic detail in children's speech. Moreover, this perception is relatively impervious to bias.

Children learn to speak like adults in a remarkably short period of time. Numerous cross-sectional and longitudinal studies have shown that by the age of only 5 or 6 years, children produce most or all of the sounds of their language correctly, as judged by phonetic transcriptions made by experienced transcribers (e.g., Smit, Hand, Freilinger, Bernthal, & Bird, 1990). From the first vocalizations to the point at which children's productions are transcribed as completely accurate, children are learning contrasts. The earliest contrasts that children learn may be as simple as the contrast between different syllable shapes. As children later produce more adult-like speech, they are able to produce even such fine-grained contrasts as the contrast between two highly similar sounds

Just as is the case with many other aspects of language acquisition, the acquisition of contrast is *gradual*. The mechanics of this gradual change, however, can be understood differently depending on how the acquisition of contrast is measured. Many large-scale studies of speech-sound acquisition (e.g., Sander, 1972; Smit et al., 1990; Templin, 1957) use phonetic transcription as a basis for assessing the development of contrast. Incorrect productions are generally classified as deletions, substitutions, or distortions of target consonants. Errors are often grouped together as a system of 'phonological processes' (e.g., Stampe, 1979) that simplify adult forms. Consider a hypothetical child who is transcribed as producing [t] for /k/ and [d] for /g/ errors. In this approach, the child would be characterized as having fronting errors, which arise because he or she has not yet developed a contrast between velar and alveolar stops. The transcribed errors are taken to be substantively equivalent to the production of the same sounds in correct words (i.e., the [t] in *key* is taken to be equivalent to the [t] in *tea*). From this perspective, the transition from consistent [t] for [k] production, to inconsistent production of the correct phonemes across different words and phonetics contexts, to correct production in all

target environments would be evidence of gradual acquisition. Thus, within this view, gradual change refers to the gradually increasing percentage of correctly produced phonemes across the lexicon, as the phonological categories of the language are mastered.

A second way of thinking about gradual acquisition is to focus on a finer-grained level of detail, specifically, the gradual acoustic or articulatory differentiation of similar sounds (see Hewlett & Waters, 2004 for a review). Consider again a child producing what is transcribed as a [t] for target /k/. Detailed articulatory and acoustic studies would examine developmental changes in the articulatory-acoustic differentiation of target /t/ from target /k/. As compared to studies that characterize phonological development from transcribed samples, articulatory-acoustic research suggests that children's development of contrast progresses gradually from productions of, for example, /t/ and /k/, that are undifferentiated from one another to productions that are robustly differentiated. As a consequence, children's developmental paths include points during which they produce "intermediate productions." These are productions with acoustic-articulatory properties that are intermediate between a target phoneme and the phoneme associated with the error. The [t] that a child with an apparent fronting error produces in the word *key* may be substantially different from the [t] in the word *tea*. In addition, some productions of the initial sound in *key* may sound as if they are in between a /t/ and a /k/. Thus, when thinking about gradient phonological learning from this perspective, the focus is on the acquisition of contrast at the level of the speech-sound category, rather than on the acquisition of contrast at the level of the lexicon.

Support for gradient change in articulatory-acoustic learning can be found in literature describing *covert contrast*. Covert contrast occurs when significant acoustic differences are present between two phonemes in a child's speech, but both phonemes are transcribed with the

same symbol. Because both variants fall within a single adult perceptual category, transcribers perceive the two variants as the same phoneme. Covert contrast has been found in the speech of typically developing children and children with phonological disorders (e.g., Baum & McNutt, 1990; Forrest, Weismer, Elbert, & Dinnsen, 1994; Forrest, Weismer, Hodge, Dinnsen, & Elbert, 1990; Hewlett, 1988; Li, Edwards & Beckman, 2009; Macken & Barton, 1980; Scobbie, Gibbon, Hardcastle, & Fletcher, 2000). One of the earliest studies of covert contrast was by Macken and Barton (1980). Macken and Barton examined three children's development of the stop consonant voicing contrast (measuring voice onset time (VOT) in word-initial stop consonants in four children). Macken and Barton found that before the children acquired an adult-like VOT contrast, they went through two stages. Initially, the children had no VOT differences between target voiced and voiceless stops. Next, the children went through a phase in which they produced target voiceless stops with longer VOTs than voiced stops. However, most of the productions had VOTs that fell into the adult range for voiced stops. As a result, the children's voiceless stops were perceived as voiced stops. That is, children were perceived as substituting voiced stops for target voiceless stops. Other studies have also found evidence for covert contrast, not only for voicing contrasts, but also for contrasts involving place of articulation for both stops and fricatives (Baum & McNutt, 1990; Forrest et al., 1990, 1994; Gierut & Dinnsen, 1986; Maxwell & Weismer, 1982). For example, Baum and McNutt (1990) compared children's correct productions of /θ/ with correct productions of /s/ and frontal misarticulations of /s/. Although frontal misarticulation of /s/ are commonly described as a substitution of /θ/, acoustic analyses revealed significant differences between frontal misarticulations and correct productions of both /s/ and /θ/.

Recently, Kong (2009) provided further evidence of the gradient nature of speech-sound acquisition in an analysis of voice onset time (VOT) in children's productions of word-initial stop consonants. In a cross-sectional study of children aged 2- through 5-years-old, Kong found that there is a large range of VOT values in children's speech, spanning values that are appropriate for adult voiced categories to those appropriate for voiceless ones. Her results demonstrate that although the VOT distributions were highly adult-like overall, there was considerable variability in the children's productions. Children's VOT values did not all fall into clearly distinguishable voiced and voiceless categories. Instead a natural continuum was formed with some productions closely approximating the prototypical adult-like VOT values and others falling intermediate between prototypical VOT values for voiced and voiceless stops.

### **Implications of gradient change for transcription**

One potential reason for differing reports of gradient versus categorical change in speech sound acquisition is the type of analysis tool used to characterize children's speech sound productions. In the field of phonological development and disorders, transcription has long been the preferred tool—and, often, the only tool—to identify and characterize children's speech sounds. Broad transcription typically involves a coarse-grained denotation of the production using phonetic symbols, which are used to make binary judgments of "correct" or "incorrect." Transcription relies on two main assumptions on how speech is perceived. First, it relies on the assumption that children's productions can be parsed into a finite set of phonetic categories. Secondly, it relies on the idea that these sounds can be denoted with a set of standard phonetic symbols. Both of these assumptions plausibly reflect the biases that occur because of categorical perception. Categorical perception refers to the observation that listeners parse continuous acoustic variation in obstruent consonants (and, to a lesser degree, other consonant manners) into

a discrete set of phonemes, and that subtle acoustic differences within a category are imperceptible (e.g., Liberman, Harris, Hoffman, & Griffith, 1957). If listeners perceive children's productions categorically, then their intermediate productions would not be reflected in transcriptions. These productions will be heard as correct if they remain in the same perceptual category as the target, even if they vary acoustically from the prototypical adult form. Productions will be heard as clear substitutions only if their acoustic characteristics are consistent with another perceptual category.

Macken and Barton's (1980) findings illustrate this point. Using transcription alone, children's productions during the second phase of development would likely be transcribed as a substitution error and denoted with the phonetic symbol [b], [d], or [g]. Labeling the production as a substitution error does provide a coarse-grained description of the child's production. However, more fine-grained information is lost—namely the fact that the children were, in fact, making a systematic contrast between voiced and voiceless stops. This highlights the limitations that are imposed when phonetic transcription is used as the sole method for denoting children's speech productions. When children's productions vary subtly in acoustic-phonetic properties from a prototypical adult form, the use of transcription may obscure potentially important information, such as that the children are capable of producing a subphonemic contrast between two phonemes. In other words, children may perceive that two phonemes are different and have begun to produce them differently at a subphonemic level, but are simply not yet able to consistently produce an adult-like contrast. This distinction has important implications for the assessment and treatment of children with disorders in speech-sound production, as it has been shown that children who exhibit covert contrast progress more quickly in therapy than children without covert contrast (Tyler, Figurski & Langdale, 1993).

For this reason, it has been suggested that acoustic analysis be used to supplement transcription (Kent, 1996). Certainly, acoustic analysis can provide a wealth of information about children's speech sounds. However, it is also extremely time-consuming, and, as many clinicians point out, it is often impractical in clinical practice for this reason. It would also likely be impossible to develop adequate norms solely using acoustic analysis, as acoustic variation relates both to the attainment of adult-like speech motor control as well as the development of an adult-sized and shaped vocal tract. In addition, transcription can be done "on the fly" as a child speaks. This allows clinicians to provide immediate feedback to children during treatment. This is not a possibility with manual acoustic analysis, at least with current technology.

The question then arises: how can we accurately perceive and differentiate intermediate productions from targets and substitutions in the clinical assessment of children's speech? Is it possible for transcription to be used in such a way that intermediate productions are reliably identified and important subtle cues distinguishing productions are not lost? The answer to this question depends in part on listeners' ability to perceive subphonemic variation. If perception were strictly categorical (i.e., if listeners were unable to perceive subtle subphonemic differences, regardless of the perception task), it would not seem possible to perceive and denote intermediate productions during transcription. However, there is significant evidence that listeners are able to perceive a wide range of subphonemic variation in speech when given the appropriate task, and that the apparent inability to do so in some earlier studies was a consequence of the method used to assess perception. In tasks that elicit categorical judgments, listeners perceive speech sounds categorically, with little attention to subtle, subphonemic acoustic differences.

Other speech perception tasks, however, do not ask listeners to respond categorically and, given these tasks, listeners are able to detect subphonemic detail. A number of studies have shown that listeners can perceive subtle within-category acoustic differences for obstruents in certain tasks (e.g., McMurray, Tanenhaus, & Aslin, 2002; Carney, Widin, & Viemeister, 1977; Pisoni & Tash, 1974). The perception of subphonemic detail appears to affect perception and recognition at multiple levels of abstraction beyond mere sensory perception. For example, MacMurry (2002) found that subphonemic acoustic differences were involved in patterns of lexical activation in a minimal pair word discrimination task. These studies have primarily relied on synthetic speech, varying along the VOT dimension, as stimuli.

There is relatively little research on adults' perception of the natural 'continua' that are present in children's speech as a consequence of their intermediate productions. The small body of research that has addressed this topic has focused on important clinical consequences of this variation. For example, research suggests that transcribers are less reliable in identifying some misarticulated speech sounds as compared with correctly produced sounds (Pye, Wilcox, and Siren, 1988). Although Pye et al. (1988) do not specifically comment on the nature of these misarticulated speech sounds, it is possible that one reason for decreased reliability is that these speech errors reflected intermediate productions that did not clearly fall into a single adult perceptual category. Other researchers have established that some adult listeners are able to distinguish between correct productions of /r/ and productions of /r/ that are intermediate between /r/ and /w/ in synthesized child speech (Sharf, Ohde, & Lehman, 1988; Wolfe, Martin, Borton, & Youngblood, 2003). To our knowledge, no research has been done to address how adults perceive naturally occurring, within-category variation in the obstruent productions of young children.

Given that listeners appear able to perceive these subphonemic differences, several suggestions have been made for how these might be denoted and incorporated into standard clinical assessments. Both Stoel-Gammon (2001) and Edwards and Beckman (2008) have suggested that one way to improve transcription reliability is to distinguish between intermediate productions and correct productions or clear substitutions. To the extent that listeners are able to perceive intermediate sounds, this would allow coding of subtle distinctions that may be lost using the standard transcription process.

Another related possibility is to use a scaling procedure in which listeners do not simply judge whether a given sound is correct or incorrect, but *how* correct or incorrect it is. One such method uses Visual Analog Scaling [VAS]. VAS is often used in the assessment of complex, multidimensional percepts, such as the perception of pain in clinical medical settings. There is considerable research on the reliability and validity of this measure in the pain literature (e.g., Price, McGrath, Rafii & Buckingham, 1983; Bijur, Silver, & Gallagher, 2001; Gallagher, Liebman & Bijur, 2001). It is also used widely in the study of voice disorders, and is part of one standardized voice assessment, the CAPE-V (Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009). VAS has also been used to study adults' perception of children's speech. In one such procedure, Urberg-Carlson, Kaiser, and Munson (2008) used a horizontal line with endpoints representing two contrasting phonemes: /s/ and /ʃ/. Listeners were asked to use a mouse to click a point on the line where they perceived that a given sound production fell along the continuum. Urberg-Carlson et al. found a strong correlation between the click location for individual stimuli and the centroid frequency of the stimuli, where centroid frequency is the principle acoustic parameter discriminating the endpoints /s/ and /ʃ/.

The basic principle behind both of these techniques (i.e., using VAS to measure phoneme production or expanding transcription to include symbols for intermediate productions) is that using a less categorical, non-binary measure in auditory perceptual tasks may help to elicit judgments that are more consistent with the gradient nature of speech development. However, few studies have systematically examined the effects of these suggestions on transcription validity and reliability. This study aims to fill this gap by examining the utility of VAS ratings to assess children's productions of /s/ and /θ/.

### **The role of listener bias in perceptual judgments**

There exists one final concern with transcription that has not yet been addressed: the potential for bias in perceptual judgments. This is a concern for transcription in general, given the ample evidence that speech-sound perception can be biased by numerous factors. In other words, a constant auditory signal may be perceived differently by the same listener solely based on his or her expectations regarding the talker. These listener expectations may stem from a variety of different sources of information about a talker. Johnson, Strand, and D'Imperio (1999) found that expectations regarding the gender of a talker influenced vowel perception, such that a given vowel sound was perceived differently depending on whether listeners believed the talker was a man or a woman. Sociolinguistic expectations related to regional dialect have also been shown to affect speech perception. Niedzielski (1999) found Detroit listeners perceived the diphthong /au/ differently depending on whether they believed the talker was from Detroit or Canada. Similarly, it has been demonstrated that listeners from New Zealand perceive diphthongs differently if they believe the talker is from New Zealand versus from Australia (Hay, Nolan, and Drager, 2006; Drager & Hay, 2006). Hay, Warren, and Drager (2006) also found that expectations regarding the perceived age and social class of a talker also impacted

listeners' perceptions the same diphthongs. Additionally, the classic McGurk effect shows that visually-based listener expectations affect how adults perceive speech sounds (McGurk and McDonald, 1976). Listener expectations have even been shown to affect whether a non-speech acoustic signal is heard as speech (Remez, Rubin, Pisoni & Carrell, 1981). Studies using adult speech and synthetic speech stimuli suggest that listener expectations play a larger role in the perception of ambiguous stimuli than in unambiguous stimuli (Diehl, Lotto, & Holt, 2004; Samuel, 2001). The effect of expectations may be particularly strong in the perception of children's speech because children's speech is more variable than that of adults (Baum & McNutt, 1990; Kong, 2009) and may contain speech sound errors (e.g., Ingram, 1976) including these intermediate (more ambiguous) productions. However, very little research has directly investigated the influence of listener expectations on the perception of children's speech, and the results are equivocal. For example, Meitus, Ringel, House, and Hotschkiss (1973) found no evidence that listener expectations based on case history information influenced adults' perception of children's speech. On the other hand, Wilson and Gasek (1975) reported that experienced speech-language pathologists' perceived children's speech differently depending on whether they were biased to think a child had a "mild-moderate" articulation disorder versus a "moderate-severe" disorder. Thus, a second focus of this study is to examine individual listeners' susceptibility to bias when perceiving children's /s/ and /θ/ productions.

### **Aims of the current study**

The present study aimed to address these concerns with transcription by exploring how adult listeners perceive the speech of young children, with a special focus on their perception of intermediate productions. Specifically, we looked at how adults perceive children's correct productions of /s/ and /θ/, clear substitutions ([s] for /θ/ and [θ] for /s/), and intermediate

productions (tokens perceived to be neither clearly /s/ nor clearly /θ/). The /s/ and /θ/ sounds were chosen for several reasons. First, both are typically mastered relatively late in development (e.g., Sander, 1972; Fudala & Reynolds, 1986; Smit et al., 1990). Additionally, children have often been observed to produce /θ/-like sound substitutions for /s/ (McGlone & Proffitt, 1973; Smit, et al., 1990). Indeed, in the speech of 100 English-speaking children recorded for a larger project (Edwards & Beckman, 2008), numerous cases of frontal misarticulations of /θ/-like sounds for /s/ were observed, and these errors were the predominant errors of /s/ in the Smit et al. (1990) study. By including correct productions, clear substitutions, and intermediate productions, we essentially created a natural "continuum" of speech sounds ranging from /s/ to /θ/.

Our first research question was to determine whether naïve listeners perceive sounds characterized as intermediate by a trained phonetician differently than those characterized as either clear substitutions or correct productions, when assessed using a non-binary choice task (namely, VAS). Secondly, we wanted to determine the relationship between listeners' VAS responses and the acoustic measures that differentiate between the endpoint /s/ and /θ/ stimuli. Specifically, we wanted to know whether the gradient responses on the VAS task would relate meaningfully to the acoustic parameters that differentiate between /s/ and /θ/. Thirdly, if these intermediate productions were perceived differently on VAS measures, we were interested in learning whether the intermediate productions would also pattern differently on a speeded binary choice task in which accuracy data and reaction times were recorded. This measure is more similar to a traditional, categorical transcription task

We predicted that naïve listeners would perceive intermediate productions differently from clear /s/ and /θ/ productions when the VAS measure was used. We also expected that the VAS click locations, both within and across subjects, would correlate with acoustic measures that differentiate between /s/ and /θ/. Furthermore, we predicted that response times collected on the speeded binary choice task would be higher for the intermediate productions than for the clear endpoints.

The second purpose of this study was to examine how susceptible listeners' binary judgments of speech-sound accuracy are to bias. Specifically we examined whether listeners' perception of /s/ and /θ/ is mediated by what they believe the age and overall level of phonological development of the talker to be. That is, we examined whether listeners adjust their perceptual criteria when they believe talkers to be older and to have more-advanced phonological development, compared to being younger and having less-advanced phonological development. As with Experiment 1, the analysis places special emphasis on whether intermediate productions are particularly susceptible to bias. Additionally, we were interested in determining whether clinical experience would be a mediating factor in understanding the effects of bias. A logical goal in clinical assessment is to decrease the likelihood that bias will influence clinical judgments. However, in a study exploring the effects of case history bias on judgments of children's speech, Meitus and Ringel (1975) concluded that experienced clinicians were actually more prone to bias than were undergraduate speech-language pathology students. On the other hand, other studies (e.g., Podol & Salvia, 1976) failed to replicate this result. Therefore, the relationship between susceptibility to bias and clinical experience is not well-understood.

### Experiment 1

The purpose of the first experiment was to determine whether naïve listeners perceive children's productions intermediate between /s/ and /θ/, as determined by skilled transcribers, differently than those characterized as either clear substitutions or correct productions. To elicit perceptual judgments, a visual analog scaling (VAS) task was used. Responses on the VAS task were then compared to psychoacoustics analyses of the same productions.

#### Method

##### Participants.

Twenty adult listeners participated in this study. All were living in Minneapolis, MN, were native speakers of North American English, and were between the ages of 18 and 45. Participants were recruited by referral or by postings at the University of Minnesota and in the surrounding community. According to self-report, none of the participants had a history of speech, language, or hearing disorders. Each participant provided informed consent and was compensated for his or her time.

##### Stimuli.

For this experiment, 200 word-initial consonant-vowel (CV) syllables beginning with /s/ and /θ/ were excised from single word productions of familiar words (such as *sofa*) and non-words (such as /sʌp<sup>h</sup>on/) which were elicited from two- to five-year-old native English speakers using a word repetition task. These words came from a larger study (Edwards & Beckman, 2008) on obstruent development across several languages. Full details of the elicitation protocol, as well as a description of the effects of lexicality, word length, and prosodic structure on consonant accuracy, can be found in Edwards and Beckman (2008). All of the words were

transcribed by a native speaker of English (the first author). During the transcription process, the transcriber first made a binary judgment as to the accuracy of the /s/ or /θ/ production. The transcriber then broadly transcribed it using IPA phonetic symbols. Additionally, using Stoel-Gammon's (2001) suggestion, the transcriber identified and coded productions that she perceived as intermediate between /s/ and /θ/, differentiating between those that were intermediate but closer to /s/, and intermediate but closer to /θ/. For the purposes of our analyses, however, these were all pooled into a single "intermediate" category. The CV syllables that were selected for this experiment were those transcribed as containing one of the following: a correct /s/, a correct /θ/, an [s] for /θ/ substitution, a [θ] for /s/ substitution, or a sound that was intermediate between /s/ and /θ/. The stimuli were balanced such that approximately half were transcribed as /s/ and half were transcribed as /θ/. In addition, for each transcription category, vowel context and the speaker's age were balanced as best as possible. Tables 1 and 2 describe the full set of the CV stimuli. Each CV syllable was normalized for amplitude. The CV syllables contained the initial fricative and a 150 ms vocalic portion. All of the vowels were monophthongs.

To further characterize the stimuli, a set of acoustic analyses was conducted. The acoustic characteristics of the stimuli are presented in Table 3, which is identical to Table 2 from another study using these stimuli (Munson, Edwards, Schellinger, Beckman, & Meyer, 2010). Specifically, acoustic analyses of spectral moments and intensity as well as psychoacoustic measures of loudness, peak ERB and compactness were conducted. Analysis of the four spectral moments (mean, variance, skewness, and kurtosis) has often been used to examine the spectral properties of fricative noise and treats the power spectrum as a probability distribution (Forrest, Weismer, Milenkovic, & Dougall, 1988 (also Forrest et al., 1990; Shadle & Mair, 1996;

Jongman, Wayman, & Wong, 2000). Psychoacoustic measures aim to incorporate the loudness and frequency scales of the human auditory system into acoustic analyses. Peak ERB (equal rectangular bandwidth) is a measure of the peak amplitude frequency at the point of highest specific loudness (measured in sones). Compactness refers to the proportion of energy (in sones) within a 3-ERB band centered at peak amplitude frequency, from a normalized spectrum. All acoustic analyses were conducted using Praat (Boersma & Weenink, 2001). An explicit justification for the use of psychoacoustic measures can be found in Kong (2009) and Arbisi-Kelm, Beckman, Kong, and Edwards (2008). Briefly, the spectral moments analyses used to examine fricative production, while quite useful for discriminating among different places of articulation for fricatives, are not based on auditory models of frequency and loudness perception. The psychoacoustic measures are based on auditory perception models, and are intended to supplement the spectral moments both in describing the stimuli, and in serving as independent measures in analyses examining the predictors of fricative perception.

A discriminant function analysis was conducted to determine which acoustic and psychoacoustic measures best differentiate among different fricative types. Peak ERB frequency, the compactness index, fricative duration, and total loudness in sones (i.e., loudness summed across all ERBs) were independent measures. The second-formant frequency of the following vowel at its onset (henceforth onset F2) was also included, as this has been shown to differentiate between another English fricative contrast, /s/ and /ʃ/, by Li, Edwards, and Beckman (2009). In this DFA, F2 was scaled in Equivalent Rectangular Bandwidths (ERBs). These were used as predictors in a discriminant function analysis predicting classification as either /s/, /θ/, or an intermediate fricative. The psychoacoustic measures correctly classified 69% of the

fricatives. Target /s/ was classified at 74% accuracy, /θ/ at 67%, and the intermediate category at 64%.

---

Insert Tables 1, 2, and 3 here.

---

### **Procedure.**

Each participant was tested individually in a sound-proof booth, seated in front of a computer monitor. Each of the 200 CV stimuli was played over headphones in random order using E-Prime software (Schneider, Eschmann, & Zuccolotto, 2002). Listeners were informed that they would hear consonant-vowel syllables taken from words that were supposed to start with "s" or "th." Instructions gave examples of words beginning with /θ/ to cue them that they were to listen for the voiceless variant, and not for /ð/. The listeners were asked to rate the consonant in each CV syllable using a visual analog scale (shown in Figure 1) that was presented on the computer monitor. Listeners were explicitly instructed to click the location along the line that corresponded with the percept of 'proximity' to "s" or "th" and were encouraged to use the entire line.

---

Insert Figure 1 here.

---

### **Data Analysis.**

The click location for each stimulus trial was analyzed in terms of the number of pixels along the x-dimension of the visual analog line. The left end of the VAS line (corresponding to "the "s" sound") was denoted as the zero point. Responses that were more than +/- 20 pixels from the line in the y-dimension, and that were more than +10 pixels greater than the right end of the line, or -10 pixels less than the left end of the line were excluded from the analysis.

Click locations for each trial were then normalized. For each CV stimulus, the minimum click location across the subjects was subtracted from the maximum click location across subjects to determine the range. The click location for each trial was then divided by this range, resulting in click location values that ranged from zero to one. Thus, click locations closer to zero correspond with percepts more like "the "s" sound" and click locations closer to one correspond with percepts of more like "the "th" sound". Average proportional VAS ratings for the six transcription categories were calculated for each subject. A second set of dependent measures was made to address our second research question of determining how listeners' VAS responses relate to the psychoacoustic parameters that vary among the 200 fricatives. These were the average proportional VAS ratings calculated across all subjects for each stimulus.

## Results

### **Mean VAS response by transcription category.**

Our first analysis focused on the effect of transcription category on VAS click location. Figure 2 shows the mean VAS click location for each of the five transcription categories. A repeated measures analysis of variance was performed with VAS response as the dependent variable and transcription category as the within subjects variable. A significant main effect ( $F[4,80] = 115.4$   $p < 0.001$ ,  $\eta^2_{\text{partial}} = .85$ ) of transcription category was observed. Bonferroni-corrected post-hoc paired comparisons revealed significant differences between all transcription categories ( $p < 0.001$  for all 10 comparisons) at the  $\alpha = .05$  level.

---

Insert Figure 2 here.

---

**Mean VAS by psychoacoustic measures.**

The next set of analyses examined the statistical relationship between the average VAS rating for individual tokens, averaged across the 20 listeners, and the measured characteristics of the stimuli. A multiple regression analysis using psychoacoustic measures as predictors of VAS click location was conducted. In this model, compactness, frication loudness, peak ERB, fricative duration, and onset F2 (in Bark) were entered into the model. The model was significant ( $F[5,194] = 42.63, p < 0.0001$ ) and accounted for approximately 52% of the variance in VAS click location ( $R^2 = 0.5235$ ). Further results are given in Table 4. When all variables were entered into the model, compactness, fricative loudness, and peak ERB were significant predictors of click location. Simple correlation plots of these three significant predictors with VAS click location are displayed in Figure 3. Regression diagnostics were carried out to assess the viability of the assumptions of normality and constant variance of the residuals. Results indicated that the assumption of normality of residuals was violated. This could result in p-values that are inaccurate. However, because the p-values for the F-test of the regression model and the t-tests for individual predictors were very small, some degree of inaccuracy is unlikely to change the general conclusions.<sup>1</sup>

In addition to the primary regression models including all predictors, an exploratory all-subsets regression analysis was performed to determine an optimal model for the psychoacoustic predictors. All-subsets regression examines the fit of all possible models, using every possible combination of predictors. In this analysis, we used the Bayesian Information Criterion (BIC) as an index of fit. The model with the highest BIC value (i.e., the best fit) was the three-predictor

---

<sup>1</sup> A separate regression analysis was also conducted using more traditional acoustic measures of the stimuli as predictors of VAS click location. In this analysis, the four spectral moments, intensity of frication (in decibels), fricative duration, and onset F2 (in Hertz) were entered into the model. The full model was significant ( $F[7,192]=29.87, p<.0001$ ) and accounted for approximately 52% of the variance in click location ( $R^2=0.5213$ ). The first and second spectral moments, intensity of frication, and onset F2 emerged as significant predictors.

model with fricative loudness, peak ERB, and compactness predicting click location. The best-fitting single-predictor model used compactness to predict click location.

In each of these analyses, the dependent variable was VAS click location, averaged across listeners. We were also interested in learning whether the results found in these analyses would hold for individual listeners. To address this question, we conducted additional regression analyses using individual listeners' VAS responses as the dependent variable. The results indicated that similar patterns emerged for the group regression and individual listener regressions. For all listeners, the full psychoacoustic model was significant in predicting VAS click locations. When averaged across listeners, compactness, fricative loudness, and peak ERB were significant predictors of click location. In the regressions for individual listeners, compactness was a significant predictor of VAS click location for 19 of the 20 listeners, fricative loudness was significant for 18 listeners, and peak ERB was significant for 14 listeners. This suggests that VAS click locations are related to the psychoacoustic properties of the stimuli for all listeners. Nevertheless, the importance of different predictors may vary to some degree from listener to listener.

---

Insert Table 4 here.

---

---

Insert Figure 3.

---

## **Discussion**

Our first research question was to identify whether listeners were able to perceive within-category differences of sounds when a VAS rating task was used, and to determine whether the naïve listener judgments patterned similarly to that of the trained phonetician/transcriber. The results indicated that naïve listeners did perceive subphonemic differences using the VAS task.

Furthermore, their responses were very similar to that of the transcriber who was able to listen to each sound as many times as she wished and to examine the waveform and spectrogram. The average VAS click location, averaged across listeners, differed for each of the transcription categories. As expected, productions originally transcribed as a correct /s/ had the highest (i.e., most "s-like") VAS ratings and productions transcribed as correct /θ / had the lowest VAS ratings. Interestingly, there was a significant difference in listeners' ratings for productions transcribed as [s] and [θ] based on target type; that is, there was a significant difference in VAS rating for correct /s/ productions and for productions originally labeled as clear substitutions of [s] for / θ/, despite the fact that both were originally transcribed using the same phonetic symbol. Likewise, VAS ratings differed between productions initially transcribed as correct /θ/ versus clear substitutions of [θ] for /s/, even though the same phonetic symbol was used to characterize both. This result may be a consequence of a covert contrast between correct productions and substitutions. The VAS results represent an average across all listeners, rather than the ratings of individual listeners. Therefore it is possible that although our original transcriber did not perceive a difference between these "clear" substitutions and correct productions, differences in our 20 listeners' perceptual sensitivities resulted in mean group differences between these transcription categories. Intermediate productions also patterned differently from both correct productions and clear substitutions. Taken together, these results indicate that the use of VAS, when averaged across multiple listeners, is sensitive to subphonemic differences.

In addition, the VAS results indicate that these listeners were biased to rate productions as more /s/-like than / θ/-like. Even those stimuli that were transcribed as correct productions of /θ/ were had an average VAS rating that was close to the midpoint of the scale. Conversely, the average VAS rating for productions transcribed as correct /s/ was much closer to the /s/ endpoint.

The reason for this bias is unclear, but may be the result of subtle, unintentional bias in task instructions, in that the words that were given as exemplars of words beginning with /s/ (e.g., see, say) had a higher frequency of usage than those for words beginning with /θ/ (i.e., think, thought). Alternatively, it may reflect a difference in the size of these listeners' perceptual spaces for these two phonemes.

Our second research question was to determine whether listeners' VAS responses would be meaningfully related to the psychoacoustic predictors that differentiate between /s/ and /θ/.

We found that the psychoacoustic model was significant in predicting VAS click location.

Results of the main regression analysis, exploratory all-subsets regression analysis, and the individual regressions all point to compactness, fricative loudness, and peak ERB as important variables in determining how listeners' perceive these sounds and make VAS judgments.

Nevertheless, there is some variability in the importance of different psychoacoustic predictors in predicting click location across individual listeners.

## **Experiment 2**

The purpose of the second experiment was to elicit perceptual judgments of the same 200 CV syllables used in the first experiment using a speeded binary choice task. Our intent was to measure accuracy data (whether the listeners judged the production as correct or incorrectly produced) and reaction times for stimuli of each transcription category. Additionally, we wanted to determine whether listeners were biased by expectations regarding the child's age and the presence (or absence) of a phonological disorder. These expectations were signaled using a carrier phrase that preceded the CV syllable. Finally, we wanted to examine whether the patterns of the results differed based on whether listeners did or did not have clinical experience in speech-language-pathology.

Experiment 2 consists of two tasks: one to determine the optimal carrier phrases to give the illusion of being older and phonologically more-advanced, and one to give the illusion of being younger and phonologically less-advanced. To accomplish this, the CV tokens from Experiment 1 were presented in a new perception experiment preceded by a carrier phrase that either contained speech-sound errors or by one that did not contain errors. Previous work by Munson and Baylis (unpublished data) showed that children who produce greater than 20% of phonemes in error are perceived by naïve listeners to be 13 months younger than their chronological age. Moreover, regression analyses by Munson and Baylis showed these errors are better predictors of discrepancies between children's actual and perceived ages than are measures of the  $f_0$  and resonant frequencies of children's voices. We reasoned, therefore, that preceding CVs by carrier phrases with or without errors would be sufficient to cue listeners to judge the children as younger or older. To verify this, a pre-experiment norming study was conducted examining the perception of the age of the carrier phrases.

### **Norming Study**

#### **Method**

**Stimuli.** The carrier phrase, "I really like," was recorded by a five-year-old boy who was a native speaker of Standard American English (from Minneapolis, MN). Nine productions of this carrier phrase were elicited. In four productions, all of the sounds were produced correctly. In five productions, the child was instructed to produce [w] for /r/ and [w] and [j]-for-/l/ substitutions, as in "I weawwy yike." The carrier phrase "I really like" was selected for several reasons. First, it does not contain either of the target sounds, /s/ or /θ/. Additionally, it contains the liquids /l/ and /r/, both of which are often produced incorrectly in the speech of young

children. Finally, the phrase consists of words familiar to young children and sounds like a natural phrase that could be produced by a child. The error patterns for the misarticulated phrase were selected based on the common substitution of /w/ for /r/ and of both /w/ and /j/ for /l/ in child speech.

Once these carrier phrases were recorded, the fundamental frequency ( $F_0$ ) and formants were altered to create the percept of a younger child and an older child. This was accomplished using the Pitch-Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2005). PSOLA includes a tool to scale the talker's apparent vocal-tract size, using Wakita's (1977) algorithm for estimating vocal-tract size from acoustic signals. To create the percept of an older child, the  $F_0$  was scaled to 90% and the formant frequencies were scaled so that the apparent vocal tract was 110%. To create the percept of a younger-sounding child, the  $F_0$  was scaled to 110% and the formant frequencies were scaled so that the apparent vocal tract was 90%. For the original (unaltered) carrier phrases, the  $F_0$  and the apparent vocal tract were each scaled to 100%. After these transformations, there were 27 unique carrier phrases: the original nine carrier phrases, the original nine carrier phrases with increased fundamental frequency and formant patterns, and the original nine carrier phrases with decreased fundamental frequency and formant patterns. The basic goal was to create six distinct conditions, as detailed in Table 5.

---

Insert Table 5 here.

---

**Participants.** Twenty women between the ages of twenty and thirty-five participated in this study. All were either undergraduate or graduate students in the Department of Communicative Disorders at the University of Wisconsin-Madison. According to self-report,

none of the participants had a history of speech, language, or hearing disorders. Additionally, all participants were native speakers of American English from the same dialect region as the child who produced the carrier phrases.

**Procedures.** Each participant was tested individually in a sound-proof booth, seated in front of a computer monitor. The stimuli were played over speakers. Each listener listened to a total of 108 presentations of the carrier phrases in random order during two separate tasks. The order of the two tasks was counter-balanced across listeners. For one task, the listeners were told that they would hear different children producing a phrase. They were asked to listen closely to the phrase and judge how old the child sounded using a five point scale, where "1" corresponded to a younger child (age three or younger) and "5" corresponded to an older child (age seven or older.) A visual display of the scale was presented both on the computer monitor and printed on a sheet of paper placed on the table in front of the listener (as shown in Figure 4, top). The listeners responded by pressing the appropriate number key on the computer keyboard. In this task, each listener heard each of the 27 phrases presented two times for a total of 54 stimuli.

In the second task, the listeners also heard all 27 phrases presented twice for a total of 54 stimuli. Listeners were again told that they would hear different children producing a phrase. However, instead of judging the child's age, they were asked to judge how adult-like the child's production was using a five point scale, where "1" corresponded to "less adult-like" (more likely to have a phonological disorder) and "5" corresponded to "very adult-like." A visual display was again presented on the computer monitor and on a sheet of paper in front of the listener (as shown in Figure 4, bottom). The listeners responded by pressing the appropriate number key on the keyboard.

---

Insert Figure 4 here.

---

## Results

We calculated mean rating by subject for each of the two rating conditions for each of the six sets of carrier phrases (higher  $F_0$  and formants/speech-sound errors, unchanged  $F_0$  and formants/speech-sound errors, lower  $F_0$  and formants/speech-sound errors, higher  $F_0$  and formants/error-free, unchanged  $F_0$  and formants/error-free, lower  $F_0$  and formants/error-free). An independent two-sample t-test found that there was no significant difference ( $t[138] = 0.154$ ,  $p = 0.88$ ) between the mean ratings for the two different orders (disorder-rating task first and age-rating task second versus age-rating task first and disorder-rating task second), so the data were combined across the two order conditions for subsequent analysis.

Figure 5 shows mean ratings for the disorder-rating task plotted against mean ratings for the age-rating task. Separate plotting symbols are used for the two sets of speech conditions (error-free versus speech-sound-errors) and for the three sets of  $F_0$ /formant values (original, raised, lowered). It can be observed that the two sets of ratings are highly correlated ( $r = 0.94$ ,  $p < 0.001$ ). Two two-way analyses of variance with speech errors (error-free vs. speech-sound-errors) and  $F_0$ /formant values (original, raised, lowered) as the independent variables were performed. The dependent variable for one of the analyses was the age ratings and the dependent variable for the other analysis was the disorder ratings. For the age ratings, the results showed that there was a significant main effect of speech errors ( $F[1,19] = 417.42$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = .956$ ) and a significant main effect of  $F_0$ /formant values ( $F[2,38] = 56.05$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = .747$ ). There was also a significant interaction between the two independent variables ( $F[2,38] = 14.54$ ,  $p < 0.001$ ,  $\eta^2_{\text{partial}} = 0.434$ ). This interaction was due to the fact that the age

rating difference for the two speech-error groups was somewhat smaller for the raised F<sub>0</sub>/formant value condition than for the other two groups. For the disorder ratings, there was a significant main effect of speech errors ( $F[1,19] = 618.413, p < 0.001, \eta^2_{\text{partial}} = 0.97$ ), but not of F<sub>0</sub>/formant values ( $p = 0.07$ ). There was also a significant interaction between the two independent variables ( $F[2,38] = 8.71, p = 0.001, \eta^2_{\text{partial}} = 0.314$ ). This interaction was due to the fact that the F<sub>0</sub>/formant manipulations only affected ratings for the correct stimuli, and not for the speech-error stimuli.

---

Insert Figure 5 here.

---

## Discussion

The results of this experiment suggest that when listeners were asked to judge the age of the child and whether or not the child has a speech error, they were influenced both by the F<sub>0</sub> and formant values of the carrier phrase and by the presence or absence of phonological errors within the phrase. Moreover, the two sets of ratings are highly correlated.

Of course, these results are highly tentative because only a single voice and a single carrier phrase were used in this norming study. For the purposes of the following experiment, however, these results suggest that we should choose only two carrier phrase conditions, one that was rated as "younger" and "phonologically disordered" and one that was rated as "older" and "typically developing" to ensure the maximal contrast in the carrier phrase conditions.

## Perceived Accuracy Study

### Method

**Stimuli.** For this experiment, the same 200 word-initial consonant-vowel (CV) syllables beginning with /s/ and /θ/ were used. Carrier phrases were chosen based on the results of the previous experiment. We created two maximally different carrier phrase conditions to use in the

current study. We will call these conditions "younger-disordered" and "older-typical." The "older-typical" carrier phrases consisted of the carrier phrases produced with no speech sound errors ("I really like") with either unchanged  $F_0$  and formant patterns or lowered  $F_0$  and formant patterns. Thus, eight different carrier phrases were included within this condition. The "younger-disordered" carrier phrases consisted of the carrier phrases produced with speech sound errors (" I weawwy yike") with either unchanged  $F_0$  and formant patterns or raised  $F_0$  and formant patterns. Because we wanted to have an equal number of carrier phrases in each of these two conditions, two carrier phrases matching this description were omitted, resulting in eight carrier phrases for the "younger-disordered" condition. By creating these two conditions, we ensured that the two carrier phrase types were maximally distinct from one another. Table 5 shows which carrier phrases were used (and those that were not used) from the complete set of carrier phrases that we constructed.

Each CV production was randomly paired with two different carrier phrases: one "younger-disordered" phrase and one "older-typical" phrase. Thus, during the experiment, each CV was presented twice (once with a carrier phrase of each type).

**Participants.** Thirty naïve listeners participated in this study. All were either undergraduate or graduate students in the Department of Communicative Disorders at the University of Wisconsin-Madison. According to self-report, none of the participants had a history of speech, language, or hearing disorders (with the exception of one graduate student who received articulation therapy for glide errors as a young child), and all were native speakers of American English. The listeners were divided into two groups. The first group consisted of fifteen undergraduate students between the ages of 19 and 21. This group had no clinical experience with children with speech disorders (except for several students who had a one-

semester undergraduate clinical practicum experience with a language and literacy focus). The second group consisted of fifteen graduate students, aged 21 to 24, who were enrolled in the Master's program in Speech-Language Pathology. They had all completed at least one semester of clinical practicum, although not necessarily with children with speech disorders.

**Procedures.** Each listener was seated in front of a computer screen, wearing headphones. Instructions were presented visually on the computer screen and were also read aloud by the researcher. Listeners were instructed that they would hear a variety of children producing sentences. They were told that each sentence would begin with the phrase, "I really like," and end with a consonant-vowel sequence beginning with "s." Listeners were informed that sometimes the "s" sound would be produced correctly and sometimes it would be produced incorrectly. Their job was to judge whether the "s" sound was produced correctly. Additionally, we told listeners that their responses would be timed and asked them to respond as quickly as possible after hearing the stimulus. Listeners responded by pressing buttons with the index finger of their dominant hand on a serial response box. The left-most button corresponded to a correct "s" and the right-most button corresponded to an incorrect "s." Carrier phrase-CV stimuli were presented in random order. Furthermore, because each of the 200 CV sequences was paired with both a "younger-disordered" carrier phrase and an "older-typical" phrase, listeners rated the accuracy of each CV twice. As a result, listeners provided accuracy judgments for a total of 400 stimuli.

## **Results**

**Accuracy.** Our first analysis focused on whether listener responses were affected by their experience, the carrier phrase condition, and the transcription categories. Figures 6 and 7 show mean percent of correct /s/ responses for all transcription categories plotted separately for

the two carrier phrases and for the two listener groups. These percentages were arcsine transformed for all statistical analyses. Arcsine transforms are commonly used on percentage data to normalize the distribution (Fazio, 1990). A three-way repeated measures analysis of variance was performed with percent correct [s] judgments as the dependent variable, transcription category and carrier phrase condition as the within-subject variables, and listener group as the between-subject variable. A significant main effect ( $F[4, 25] = 1534.36, p < 0.001, \eta^2_{\text{partial}} = .996$ ) of transcription category was observed. Post-hoc paired comparisons revealed significant differences between all transcription categories ( $p < 0.001$  for all ten comparisons). The main effect of carrier phrase was not significant ( $F[1, 28] = 0.015, p = 0.90$ ). Similarly, the main effect of listener group was also not significant ( $F[1, 28] = 0.907, p = 0.35$ ). The only significant interaction was between carrier phrase and listener group ( $F[4, 25] = 3.11, p = 0.033, \eta^2_{\text{partial}} = 0.332$ ). Visual inspection of the data shows that this interaction is due to the fact that for the undergraduates, there were a higher percentage of "correct [s]" responses for the "older-typical" carrier-phrase condition, whereas, for graduate students, there were a higher percentage of "correct [s]" responses for the younger-disordered carrier-phrase condition.

---

Insert Figures 6 and 7 here.

---

Our second analysis focused on only those stimuli for which intra-subject disagreement was observed across the two carrier phrase conditions. Intra-subject disagreement was measured by identifying any CV stimulus for which a given listener rated the consonant as correct when it followed one carrier phrase and incorrect when it followed a carrier phrase in the opposite condition. The percentage of stimuli for which this intra-subject disagreement occurred was calculated for each transcription category for each listener. Figure 8 shows the percentage of intra-subject disagreement as a function of transcription category. A two-way analysis of

variance (transcription category by listener group) showed a significant main effect of transcription category ( $F[4,25] = 118.68, p < .001, \eta^2_{\text{partial}} = 0.95$ ). The main effect of listener group and the transcription category by listener group interaction were not significant. Post-hoc paired comparisons found that there was a significant difference between the intermediate transcription category and all other categories ( $p < 0.001$ ). Other post-hoc paired comparisons were also significant, with the exception of these three: [s] for / θ / as compared to [θ] for /s/, [s] for / θ / as compared to correct / θ /, and [θ] for /s/ as compared to correct / θ /.

Table 6 gives the number of trial pairs for which there was intra-subject disagreement for a single CV syllable between the two carrier phrase conditions, divided by whether the subject said "yes" (correct [s]) for the "younger-disordered" carrier phrase context or for the "older-typical" carrier phrase context. A chi-squared analysis found that there were significantly more "yes" (correct [s]) responses for the intermediate transcription category with the "younger-disordered" carrier phrase, as compared to the "older-typical" carrier phrase context ( $\chi^2_{[n=20 \text{ df}=4]} = 12.95, p = 0.012$ ).

---

Insert Figure 8 and Table 6 here.

---

**Response Times.** Response times were also collected. Results paralleled the findings with the accuracy analysis and are otherwise uninformative. For this reason, the results are not presented here.

### Discussion

In designing this study, we had several primary questions. First, we wanted to examine how naïve adult listeners judge the accuracy of children's correct productions of /s/ and /θ/, clear

substitutions ([s] for /θ/ and [θ] for /s/), and intermediate productions. For the purposes of this analysis, intermediate productions were not divided by target type or whether they were transcribed as sounding slightly more /s/-like versus slightly more /θ/-like. Our results confirmed that naïve listeners' responses to each of these five transcription categories patterned differently. In other words, the mean percent of responses for which initial /s/ or /θ/ in consonant-vowel sequences was judged as a correct /s/ differed for each transcription category, such that productions transcribed as correct /s/ were judged by naïve listeners to be correct the highest percent of the time. Tokens transcribed as a substitution of [s] for /θ/ were judged as a correct /s/ the next highest percent of the time. Intermediate productions had the next highest percentage, followed by substitutions of [θ] for /s/. Finally, tokens transcribed as a correct /θ/ were judged to be a correct /s/ the lowest percent of the time.

A second purpose of this study was to determine whether listeners' expectations regarding age and the presence or absence of a phonological disorder, as signaled by a carrier phrase, would affect how they judged the accuracy of children's productions. Overall, we found no significant main effect of carrier phrase type on accuracy judgments. To some degree, this is not surprising. For example, unambiguous productions are less likely to be influenced by listener expectations. Thus, it is easy to understand why correct productions of /s/ and /θ/ were not affected by expectations. This could also explain why accuracy judgments of clear substitutions were not affected by expectations. On the other hand, this result is surprising because we had hypothesized that listener expectations would affect judgments for the more ambiguous, intermediate productions.

A second analysis did find an effect of carrier phrase, however. For this analysis, we examined only those productions where there was an intra-subject disagreement across the two

carrier phrase conditions. As predicted, there was a significant main effect of transcription category in this analysis. Listeners were most likely have different ratings for a single CV syllable across the two carrier phrase conditions for the intermediate transcription category. Interestingly, when listeners judged a given CV differently over the two presentations for this transcription category, they were more likely to judge it as a correct /s/ when it was preceded by the "younger-disordered" carrier phrase. One possible reason for this result is that listeners are more lenient when they think they are listening to a younger/phonologically disordered child rather than an older/typically developing child for these ambiguous productions. As a result, they accept a broader range of productions as correct.

Further research into the role of listener expectations is warranted. A strong base of evidence supports the claim that listeners' expectations affect perception, even when listeners are given only very slight cues to shape their expectations. For example, as discussed earlier, Drager and Hay (2006) found that even the presence of a stuffed animal (representing a given nationality) in the testing room was enough to affect listeners' perceptions of a talker's speech. Thus, it is somewhat surprising that we did not find a larger overall effect of carrier phrase. It is possible that the reason for this lies in our methodology. First, a single child produced all the carrier phrases, whereas the CV tokens were produced by many different children. Although the  $F_0$  and formants were altered, the fact remains that there were only sixteen different carrier phrases. Listeners may have quickly realized that the speaker was different for the carrier phrase and the CV. In addition, the design of this study required that all of the consonant-vowel sequences be paired with both a "younger-disordered" and an "older-typical" carrier phrase. As a result, CVs produced by children as young as two years old were preceded by a carrier phrase that was designed to sound like a much older child. Similarly, CVs produced by five-year-olds

were also preceded by a carrier phrase manipulated to sound like a very young child. Clearly, this unnatural condition could also cue listeners that it was not a single child producing the carrier phrase and the CV. A methodology that minimizes these problems might yield a greater effect of listener expectations. For example, it may be beneficial to elicit carrier phrases from the same children that produce the CVs. These carrier phrase productions could then be classified in more a naturalistic way that eliminates the need for synthetically altered carrier phrases from a single child. Alternatively, other methods of providing listeners with expectations about a child might prove successful. In this study, biasing information was presented to listeners implicitly through carrier phrases. It is possible that information that is presented explicitly (e.g., in task instructions or in a case history) may have a greater impact on listeners' judgments. Finally, this study only examined the effect of information about a child's age and phonological ability. Other information, such as social and medical history, academic functioning, native language, gender, etc. might also impact listeners' perceptions. Clearly, future research on this is needed.

Finally, our last question regarded whether clinical experience affected how listeners perceived these CV productions. Although previous research (e.g., Wolfe et al., 2003) indicated improvement in the ability to perceive subtle acoustic differences as a result of clinical experience, we found no significant differences between the group of undergraduate students versus the group of graduate students in terms of mean percent correct [s] judgments for any of the five transcription categories. However, we only used one factor as a measure of experience, namely, level in school. There was also some overlap between groups in that several of the undergraduate students had completed an undergraduate-level clinical practicum experience and some of the graduate students had also only completed a single clinical practicum. Furthermore, not all of the graduate students had clinical experience working specifically with children with

phonological/articulation disorders. In addition, there may be better indicators of experience than level of education and clinical experience. For example, we did not collect data on familiarity with children, including whether listeners had young children within their immediate families or had non-clinical work-related experience with children. Future research adopting a similar paradigm, but with a more thorough, controlled method to assess listener experience, might reveal differences in performance based on experience.

### **General Discussion**

Both experiments 1 and 2 aimed to assess listeners' perceptions of a range of children's productions of /s/ and /θ/ along a natural 'continuum.' This continuum was broadly achieved via inclusion of multiple transcription categories, including those that had been transcribed by a trained phonetician as "intermediate." Despite employing differing methodologies to elicit perceptual judgments (i.e., identification of the consonant using a VAS scale in Experiment 1 and binary choice accuracy judgments in experiment 2), both experiments yielded similar results. Mean VAS responses, averaged across listeners, were significantly different for each transcription category. Similarly, the mean percentage of "correct" responses in the binary choice task differed significantly for all five transcription categories. These parallel results are important for several reasons. First, the results validate our transcription categories in that as a group, naïve listeners' judgments paralleled our original transcriptions. Second, they indicate that naïve listeners are able to perceive subphonemic differences, given the appropriate task. They also provide support that "intermediate" is a valid transcription category. Productions labeled "intermediate" by a trained phonetician also had intermediate accuracy judgments by one group of listeners and intermediate VAS ratings by an additional, independent group of listeners. However, it must be noted that both of our experiments averaged the results across listeners.

Therefore, the gradient change in mean percent correct /s/ responses and in VAS response across all five transcription categories only reflects a group effect, rather than the judgments of individual listeners. Finally, our results suggest that covert contrast is far more pervasive than has been observed in previous acoustic studies of this phenomenon (e.g., Li et al., 2009). In both studies, correct productions and clear substitutions patterned differently. In other words, correct /s/ productions and [s] for /θ/ productions differed significantly in terms of both VAS results and accuracy judgments. The same is true for correct /θ/ productions and [θ] for /s/ substitutions.

However, to more definitely document the existence of covert contrast in these CV productions, a more sophisticated analysis comparing the acoustic and psychoacoustic properties of the stimuli in each transcription category is required.

These results are promising news for clinicians who are asked to make these distinctions every day. In a clinical setting, it can be challenging to decide exactly what is "good enough" to be correct. If these intermediate productions comprise a valid transcription category, clinicians may be able to use them in clinical practice to keep data, compare a child's productions with other children, and to provide feedback to the child. Further research is warranted to study the judgments of individual listeners on these intermediate productions, clear substitutions, and correct productions to learn the extent to which they are able to perceive subtle acoustic differences between productions.

Our results also support the use of a visual analog scale as a possible supplement to traditional assessment measures. Not only were VAS judgments sensitive to transcription category differences between stimuli, but they also correlated with acoustic and psychoacoustic measures believed to differentiate between /s/ and /θ/. A continuous measure such as visual

analog scaling may allow clinicians to better identify subtle, yet potentially important, subphonemic articulatory-phonetic detail in children's speech productions. This could be especially powerful in cases when it is impossible or inconvenient to incorporate acoustic analyses into an assessment protocol. Additionally, because visual analog scaling, like transcription, can be done "on the fly" as a child speaks, it allows clinicians to provide immediate feedback to children.

As discussed earlier, the lack of a significant main effect of bias in Experiment 2 was surprising. However, based on the results of our second analysis, we discovered that when intermediate productions were judged inconsistently by a given listener over the two presentations of a CV stimulus, a significant biasing effect of carrier phrase was observed. Importantly, we also found that intermediate productions were more likely to be rated inconsistently than either correct productions or clear substitutions. This has important clinical implications. In our study, only typically-developing children were included. Presumably, speech-language pathologists treating phonological/articulation disorders would encounter a greater number of incorrect productions, including intermediate productions. These intermediate productions might be especially prevalent when children are in the process of acquiring new speech sounds, but have not yet arrived at the prototypical, adult-like pronunciation. Clinicians must be cautious in how they approach these productions that sound somewhere in between /s/ and /θ/ (or any other contrast). If they must make a binary decision as to whether a production is an /s/ or a /θ/, clinicians should be aware that their own biases may impact their decision. It is also for this very reason that the "intermediate" category may prove especially useful in clinical practice. If these ambiguous productions are more difficult to judge and are susceptible to bias, it

may make more sense to simply consider them "intermediate" rather than force them into a phoneme category in which they do not clearly fit.

Alternatively, use of a visual analog scale may prove more beneficial in tracking subtle changes in the articulatory-acoustic differentiation of speech sounds as children gradually acquire the fine phonetic detail of the sound system. Because this study did not study the effects of bias on VAS judgments, it is not clear whether the same pattern of results seen in the binary choice tasks will also apply to VAS judgments. Follow-up research to address the susceptibility of VAS responses to bias is currently underway.

It is clear that there is much future research that remains to be done on how adults perceive children's correct and incorrect consonant productions, especially those with articulatory-acoustic features that do not clearly map onto that of a target, adult production. Nevertheless, the results of this study strongly support the use of additional procedures, such as use of a visual analog scale or an intermediate transcription category, to supplement traditional transcription. Use of these additional procedures may result in assessment results that are more sensitive to subtle acoustic properties of children's speech. It is also possible that these procedures may result in judgments that are less prone to listener bias, although further research is needed to support this claim.

## References

- Arbisi-Kelm, T., Beckman, M. E., Kong, E-J., & Edwards, J. (2008). Psychoacoustic measures of stop production in Cantonese, Greek, English, Japanese, and Korean. Paper presented at the 156th Meeting of the Acoustical Society of America, Miami, 10-14 November 2008.
- Baum, S. R. & McNutt, J. C. (1990). An acoustic analysis of frontal misarticulation on /s/ in children. *Journal of Phonetics*, 18, 51-63.
- Bijur, P.E., Silver, W. & Gallagher, E.J. (2001). Reliability of the visual analog scale for measurement of acute pain. *Academic Emergency Medicine*, 8, 1153-7
- Boersma, P. & Weenink, D. (2005). Praat: doing phonetics by computer (Version 4.3.28) [Computer program]. Retrieved from <http://www.praat.org/>.
- Carney, A., Widin, G., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America*, 62, 961-970.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, 55, 149-179.
- Drager K. & Hay J. (2006). Can you really believe your ears? The effect of stuffed toys on speech perception. Presented at *New Zealand Language and Society Conference*, Christchurch.
- Edwards J. & Beckman, M. E. (2008). Methodological questions in studying phonological acquisition. *Clinical Linguistics and Phonetics*, 22, 937-956.
- Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick, M. S. Clark, et al. (Eds.), *Research methods in personality and*

*social psychology: Review of personality and social psychology* (Vol. 11, pp. 74-97).

Newbury Park, CA: Sage.

Forrest, K., Weismer, G., Elbert, M., & Dinnsen, D. A. (1994). Spectral analysis of target-appropriate /t/ and /k/ produced by phonologically disordered and normally articulating children. *Clinical Linguistics & Phonetics*, 8, 267-281.

Forrest, K., Weismer, G., Hodge, M., Dinnsen, D. A., & Elbert, M. (1990). Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics & Phonetics*, 4, 327-340.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, 84, 115-123.

Fudala, J.B. & Reynolds, W.M. (1986). Arizona Articulation Proficiency Scale: Second Edition. Los Angeles: Western Psychological Services.

Gallagher, E. J., Liebman, M., & Bijur, P. E. (2001). Prospective validation of clinically important changes in pain severity measured on a visual analog scale. *Annals of Emergency Medicine*, 38, 633-638.

Gierut, J. & Dinnsen, D. A. (1986). On word-initial voicing: converging sources of evidence in phonologically disordered speech. *Language and Speech*, 29, 29-114.

Hay, J., Nolan, A. & Drager, K. (2006). From Fush to Feesh: Exemplar Priming in Speech Perception. *The Linguistic Review*, 23, 351-379.

- Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34, 458-484.
- Hewlett, N. (1988) Acoustic properties of /k/ and /t/ in normal and phonologically disordered speech. *Clinical Linguistics and Phonetics*, 2, 29-45.
- Hewlett, N. & Waters, D. (2004). Gradient change in the acquisition of phonology. *Clinical Linguistics & Phonetics*, 18, 523-533.
- Ingram, D. (1976). *Phonological Disability in Children*. New York: Elsevier North Holland, Inc.
- Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, 27, 359-384.
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108, 1252-1263.
- Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18, 124-132.
- Kent, R. (1996). Hearing and Believing: Some Limits to the Auditory-Perceptual Assessment of Speech and Voice Disorders. *American Journal of Speech-Language Pathology*, 5, 7-23.
- Kong, E. (2009). The development of phonation-type contrasts in plosives: Cross-linguistic perspectives. Unpublished Ph.D. Dissertation. Columbus, OH: Department of Linguistics, Ohio State University.

- Li, F., Edwards, J., & Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, 37, 111-124.
- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54, 358–368.
- Macken, M., & 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.
- McMurray, B., Tanenhaus, M., and Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access, *Cognition*, 86, B33-B42.
- Maxwell, E. M., & Weismer, G. (1982). The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics*, 3, 29-43.
- McGlone, R. & Proffitt, W. R. (1973). Patterns of tongue contact in normal and lisping speakers. *Journal of Speech and Hearing Research*, 16, 456–476.
- McGurk, H. & MacDonald, J.W. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- Meitus, I.J., Ringel, R.L., & House, A.S. (1973). Clinician bias in evaluating speech proficiency. *International Journal of Language and Communication Disorders*, 8, 146-151.
- Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., & Meyer, M.K. (2010). Deconstructing phonetic transcription: covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics and Phonetics*, 24, 245-260.
- Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, 18, 62-85.

- Pisoni, D., & Tash, J. (1974). Reaction times to comparisons within and across phoneme categories. *Perception and Psychophysics*, *15*(2), 285-290.
- Podol, J. & Salvia, J. (1976). Effects of visibility of a prepalatal cleft on the evaluation of speech. *Cleft Palate Journal*, *13*, 361-366.
- Price, D., McGrath, P., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*, 45-56.
- Pye, C., Wilcox, K. A. & Siren, K. A. (1988). Refining transcriptions: the significance of transcriber 'errors.' *Journal of Child Language*, *15*, 17-37.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception without Traditional Speech Cues. *Science*, *212*, 947-950.
- Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*, 348-351.
- Sander, E. K. (1972). When are Speech Sounds Learned? *Journal of Speech and Hearing Disorders*, *37*, 55-63.
- Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc..
- Scobbie, J., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (2000). Covert contrasts as a stage in the acquisition of phonetics and phonology. In M. Broe & J. Pierrehumbert (Eds.), *Papers in Laboratory phonology V: Language acquisition and the lexicon* (pp. 194-207). Cambridge, U.K.: Cambridge University Press.

- Shadle, C. H., & Mair, S. J. (1996). *Quantifying spectral characteristics of fricatives*. Paper presented at the International Conference on Spoken Language Processing [ICSLP], Philadelphia, PA.
- Sharf, D., Ohde, R., & Lejman, M. (1988). Relationship Between the Discrimination of /w-r/ and /t-d/ Continua and the Identification of Distorted /r/. *Journal of Speech and Hearing Research, 31*, 193-206.
- Smit, A., Hand, L., Freilinger, J. J., Bernthal, J. & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska Replication. *Journal of Speech and Hearing Disorders, 55*, 779-798.
- Stampe, D. (1979). *A Dissertation on Natural Phonology*. Bloomington, Indiana: Indiana University Linguistics Club. .
- Stoel-Gammon, C. (2001). Transcribing the Speech of Young Children. *Topics in language disorders, 21*, 12-21.
- Tyler, A. A., Figurski G. R., & Langdale, T. (1993). Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech and Hearing Research, 36*, 746–759.
- Urberg-Carlson, K., Kaiser, E., & Munson, B. (2008). Assessment of children's speech production 2: Testing gradient measures of children's productions. Poster presented at the 2008 ASHA Convention, Chicago, 20 November.
- Wakita, H. (1977). Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 25*, 183-192.

Wilson, W.R. & Gasek, G. (1975). The influence of pre-information on the rating of articulation.

*Journal of Communication Disorders, 8*, 15-22.

Wolfe, V., Martin, D., Borton, T., & Youngblood, H.C. (2003). The Effect of Clinical

Experience on Cue Trading for the /r-w/ Contrast. *American Journal of Speech-Language*

*Pathology, 12*, 221-228.

Table 1

*Stimuli Inventory: Total number of Consonant-Vowel syllables by age, vowel context, and transcription category*

Following Vowel	[θ] substitutions for /s/				Correct /θ/				Intermediate Tokens (but slightly closer to /θ/)				Total
	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	
(/i/ and /ɪ/)	1	1	2	0	0	4	13	14	2	2	4	1	<b>44</b>
(/e/ and /ɛ/)	0	4	1	0	0	0	0	0	1	2	1	1	<b>10</b>
(/ɑ/)	4	4	0	1	0	1	1	2	1	2	2	2	<b>20</b>
(/o/)	2	1	0	0	0	0	0	0	1	1	2	0	<b>7</b>
(/u/ and /ʊ/)	0	3	0	0	0	2	5	4	1	1	2	1	<b>19</b>
	<b>Total: 24</b>				<b>Total: 46</b>				<b>Total: 30</b>				<b>100</b>

Note: This table displays the half of the consonant-vowel syllables initially transcribed as sounding more /θ/-like.

Table 2

*Stimuli Inventory: Total number of Consonant-Vowel syllables by age, vowel context, and transcription category*

Following Vowel	[s] substitutions for /θ/				Correct /s/				Intermediate Tokens (but slightly closer to /s/)				Total
	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	2;0- 2;11	3;0- 3;11	4;0- 4;11	5;0- 5;11	
(/i/ and /ɪ/)	2	3	4	3	3	2	4	5	1	2	4	1	<b>34</b>
(/e/ and /ɛ/)	0	0	0	0	2	2	4	2	1	1	1	2	<b>15</b>
(/ɑ/)	1	2	1	0	0	2	2	4	0	2	2	0	<b>16</b>
(/o/)	0	0	1	0	2	1	3	2	2	2	2	0	<b>15</b>
(/u/ and /ʊ/)	1	2	4	0	1	2	4	3	1	2	0	0	<b>20</b>
	<b>Total: 24</b>				<b>Total: 50</b>				<b>Total: 26</b>				<b>100</b>

Note: This table displays the half of the consonant-vowel syllables initially transcribed as sounding more /s/-like.

Table 3

*Acoustic Characteristics of Consonant-Vowel Stimuli*

Measure	[s] for /s/		[s] for /θ/		s:θ		θ:s		[θ] for /s/		[θ] for /θ/	
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
<i>N</i>	50		24		26		30		24		46	
Peak ERB <sup>a</sup>	34.6	1.1	34.2	1.6	34.4	1.5	32.9	1.4	26.9	1.6	25.5	1.1
Compactness												
Index <sup>a</sup>	0.32	0.01	0.30	0.01	0.23	0.01	0.23	0.01	0.20	0.01	0.20	0.01
Total												
Loudness												
(sones) <sup>a</sup>	0.81	0.04	0.86	0.05	0.82	0.05	0.83	0.05	0.69	0.05	0.55	0.04
Duration												
(ms) <sup>b</sup>	209	9	210	13	214	12	223	11	187	13	174	9
Vowel F2 at onset (ERB)	21.9	0.2	22.1	0.3	22.0	.03	21.7	0.3	21.6	0.3	22.0	0.2
Vowel f0 at midpoint												
(ERB)	7.5	0.1	7.3	0.2	7.5	0.2	7.6	0.2	7.3	0.2	7.4	0.2

Table 4

*Results from the linear regression model predicting VAS click location from various psychoacoustic predictors*

	Estimate	Std. Error	<i>t</i>	<i>p</i>
(Intercept)	-.0667	.110	-.609	.543
Compactness	.942	.106	8.892	<.001
Frication		.028	5.813	<.001
Loudness	.164			
Peak ERB	.004	.001	3.941	<.001
Fricative Dur.	.071	.161	.442	.659
Bark F2	.008	.005	1.758	.080

Table 5

*Carrier phrase conditions used in the perceived accuracy study in Experiment 2*

	Younger Child	Intermediate-Aged Child	Older Child
Typically Developing	<p>“I really like” [arrililaik]</p> <p>F<sub>0</sub> and Formants raised.</p> <p>Total Number: 0</p>	<p>CONDITION 1: “older-typical”</p> <p>“I really like” [arrililaik]</p> <p>F<sub>0</sub> and Formants unchanged.</p> <p>Total Number: 4</p>	<p>“I really like” [arrililaik]</p> <p>F<sub>0</sub> and Formants lowered.</p> <p>Total Number: 4</p>
Phonologically Disordered	<p>CONDITION 2: “younger-disordered”</p> <p>“I weawwy yike” [arwiwijaik]</p> <p>F<sub>0</sub> and Formants raised.</p> <p>Total Number: 4</p>		<p>“I weawwy yike” [arwiwijaik]</p> <p>F<sub>0</sub> and Formants lowered.</p> <p>Total Number: 0</p>

Table 6

*Number of trial pairs where there was intra-subject disagreement between the two carrier phrase conditions, divided by whether the subject said "yes" (correct /s/) for the "younger-disordered" carrier phrase context or for the "older-typical" carrier phrase context*

	correct /θ/	[θ] for /s/	intermediate	[s ]for /θ/	correct /s/	total
younger-disordered	95	61	249	53	26	484
older-typical	103	68	203	39	47	460

**Figure Captions**

Figure 1. Visual Analog Scale (VAS) used in Experiment 1.

Figure 2. Mean VAS click location, plotted for each transcription category. Note: "T" refers to /θ/.

Figure 3. Mean VAS click location for each consonant-vowel stimulus plotted against various psychoacoustic measures (i.e., Leftmost plot shows mean VAS plotted by fricative loudness. Middle plot shows mean VAS by compactness. Rightmost plot shows mean VAS by peak ERB.).

Figure 4. Visual display listeners used in Experiment 2 to judge how old a child sounded (top plot) and to judge how adult-like a child's production sounded (bottom plot).

Figure 5. Mean disorder ratings (where 1= "less adult-like/more likely to have a phonological disorder," and 5= "more adult-like/excellent child speech") and mean age ratings (where 1= "younger/ three or less" and 5= "older/seven or greater") plotted for each carrier phrase condition, with a fitted regression line.

Figures 6 and 7. Mean percent correct [s] responses for each transcription category plotted separately for carrier phrase (Fig. 7) and for listener group (Fig. 8) (i.e., Figure 7 shows the mean percent of trials in which all listeners judged a consonant-vowel (CV) productions to be a "correct s" for each transcription type. Mean percents are shown separately for each carrier phrase condition. Figure 8 shows the mean percent of trials in which all listeners judged a CV production to be a "correct s" for each transcription type. Mean percents are shown separately for each listener group.) Note: "T" refers to /θ/ and "\$" refers to "substitution."

Figure 8. Percent of consonant-vowel (CV) trial pairs where there was intra-subject disagreement between the two carrier phrase conditions (i.e., a listener judged the CV production

as a correct /s/ with one carrier phrase condition and an incorrect /s/ with the other carrier phrase condition), divided by stimulus transcription category. Note: "T" refers to /θ/ and "\$" refers to "substitution."

Figure 1.

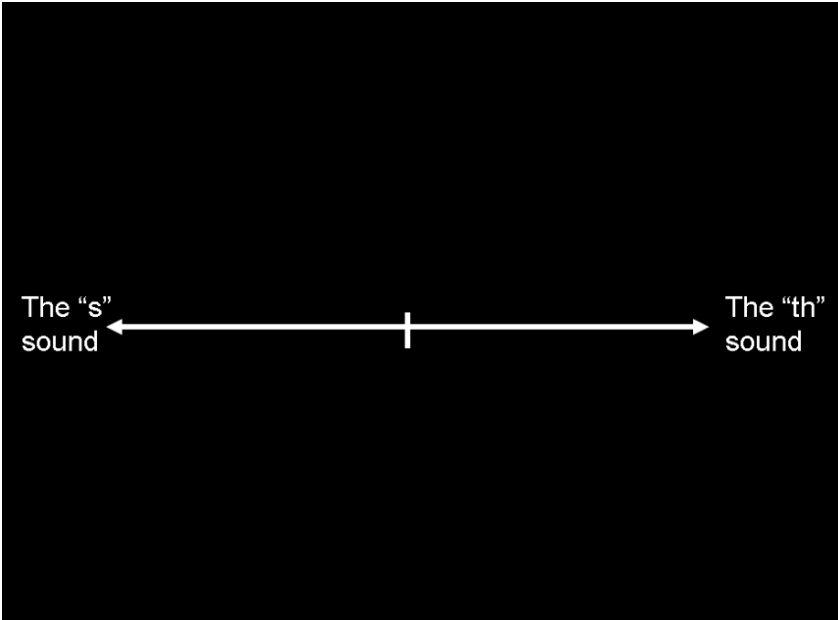


Figure 2.

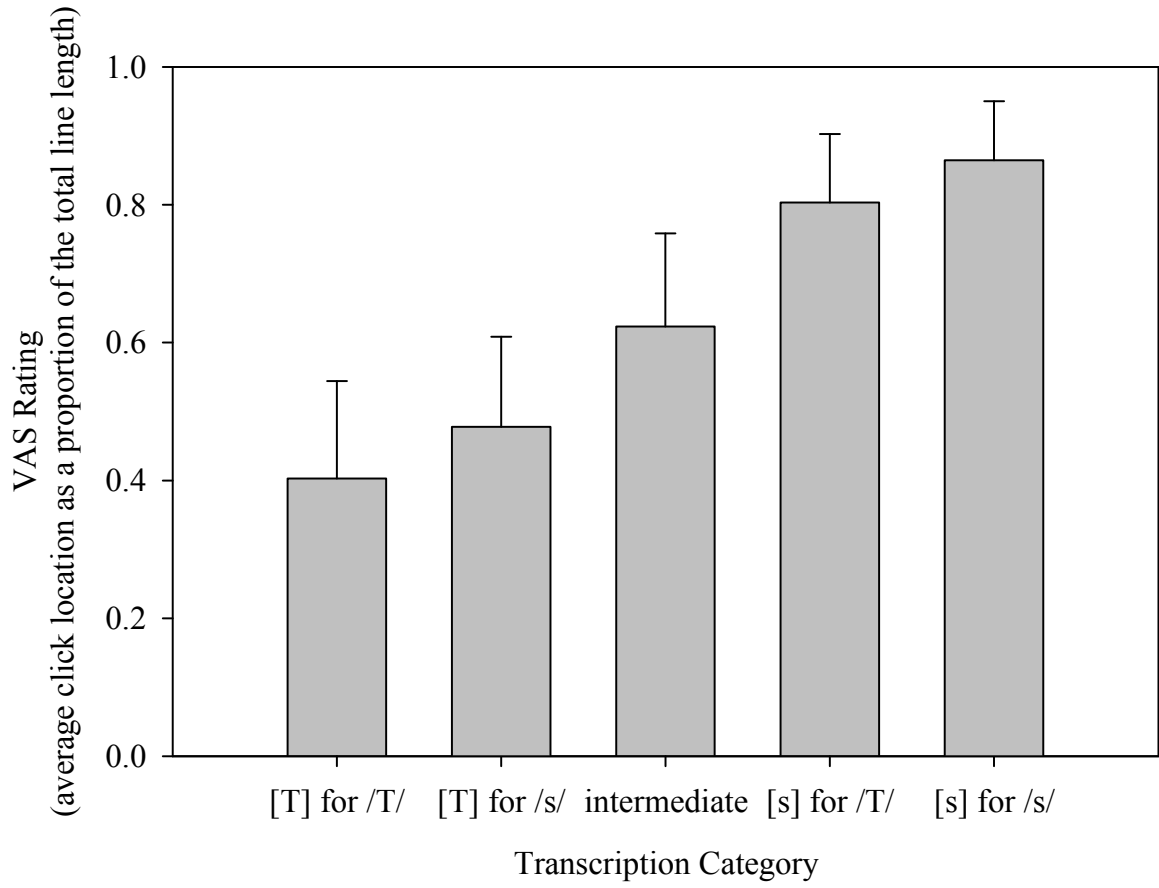


Figure 3.

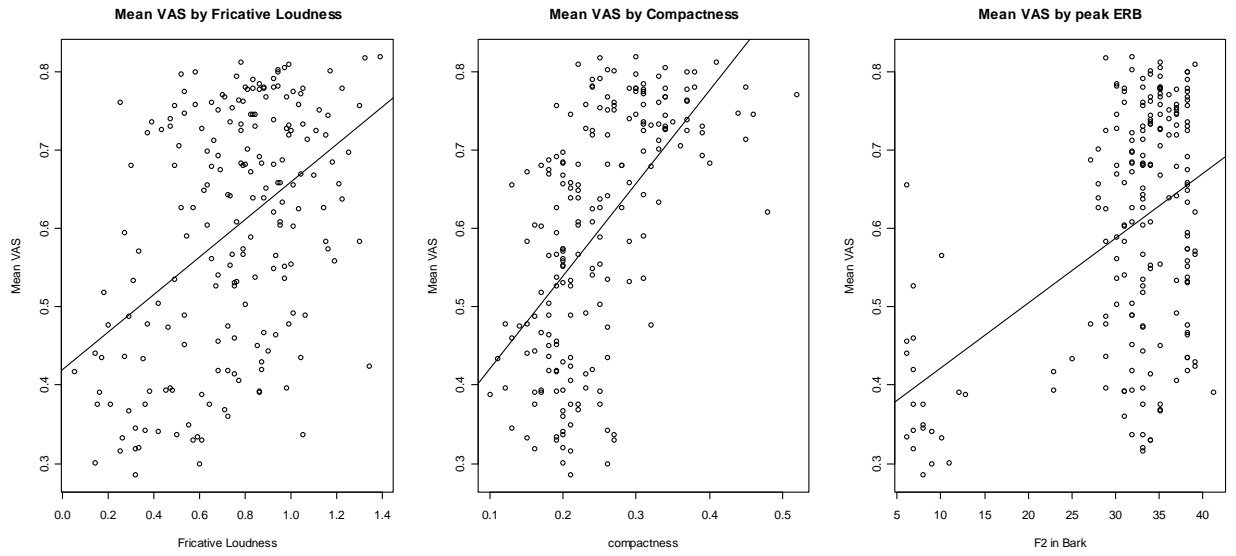


Figure 4.

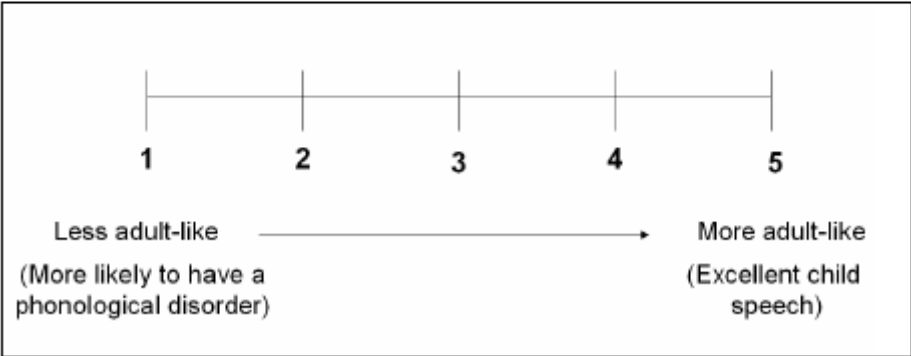
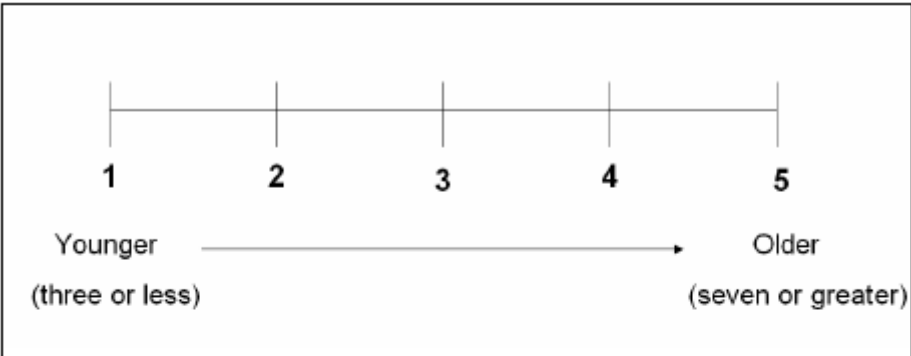


Figure 5.

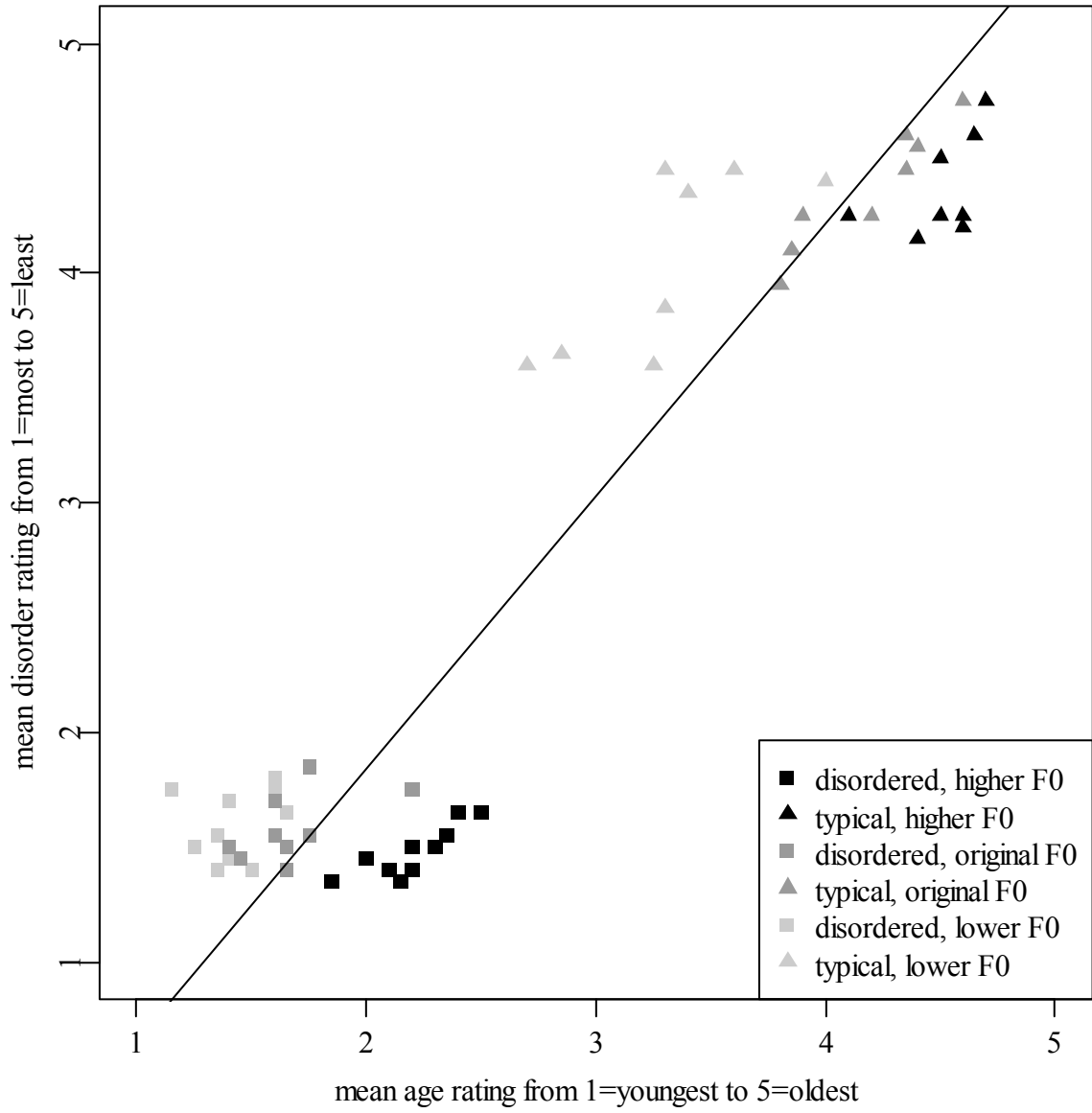


Figure 6.

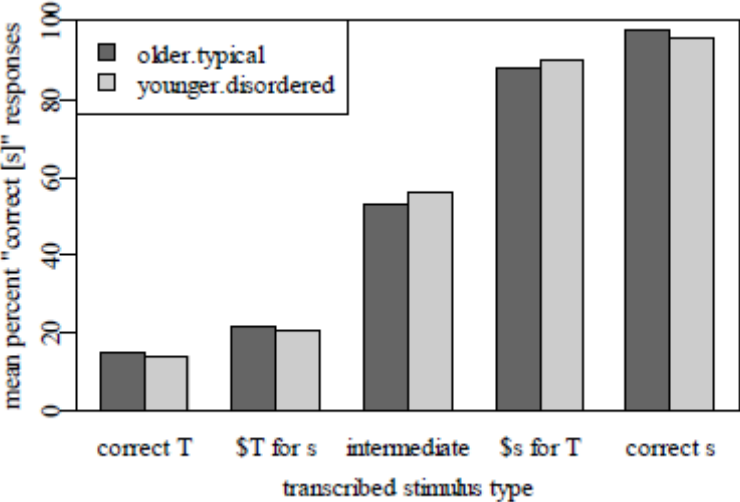


Figure 7.

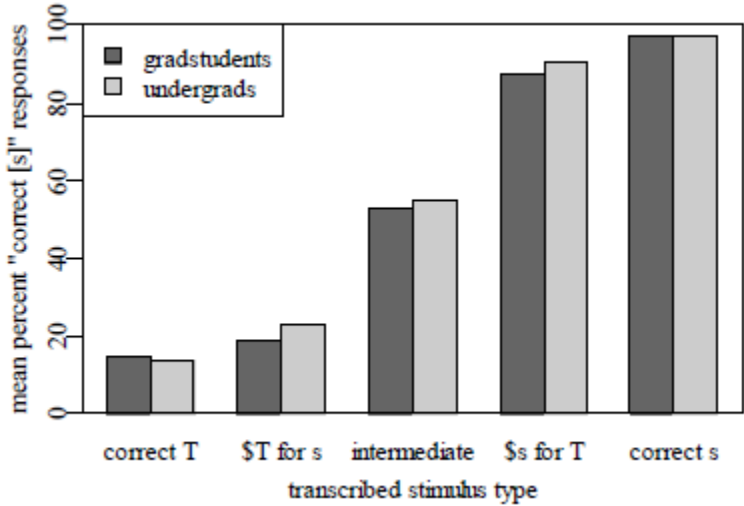


Figure 8.

