APPROVED

Jan. R. Edwards, Ph.D.	Date

THE ROLE OF INTERMEDIATE PRODUCTIONS AND LISTENER EXPECTATIONS ON THE PERCEPTION OF CHILDREN'S SPEECH

by

Sarah K. Schellinger

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

(Communicative Disorders)

at the

UNIVERSITY OF WISCONSIN-MADISON

2008

Abstract

Transcription has long been the tool of choice of clinicians and researchers studying phonological development and disorder. Although transcription is a useful analysis tool, longstanding concerns about transcription include the existence of covert contrast and the influence of listener expectations. This study modified the traditional transcription procedure; children's productions of word-initial /s/ and /θ/ were transcribed as correct /s/, correct /θ/, [θ] for /s/ substitutions, [s] for /θ/ substitutions, or intermediate (productions that were in between the [s] and [θ]). 30 naïve listeners (15 with clinical experience and 15 with limited or no clinical experience) heard consonant-vowel syllables, excised from words produced by children ages two through five. Each syllable was preceded by a carrier phrase that had been digitally altered to sound like either a younger child with a phonological disorder or an older, typically developing child. Listeners were asked to judge whether they heard a correct or incorrect /s/. Results indicated a concordance between the trained transcriber and the naïve listeners for all five transcription categories. The results also suggested that listeners were less reliable in judging intermediate productions than any of the other four transcription categories. Furthermore, an influence of carrier phrase on listener judgments was observed for these intermediate productions. No influence of clinical experience on listener's judgments was observed.

Acknowledgements

Thanks to my advisor, Jan Edwards, for all her invaluable support, encouragement, and help along every step of the way. Thanks also to Mary Beckman for her great ideas and her eagerness to help me learn about data analysis and statistics and to Benjamin Munson for offering his expertise and assistance on the design of this study. I also thank my committee members, Marios Fourakis, who let me use his lab time and time again, and Gary Weismer, who gave me the opportunity to get involved with research from the start.

I would also like to give my heartfelt thanks to everyone on the Paidologos Project. Working with all of you has been a wonderful opportunity, and I am constantly amazed by how much I have learned simply by being around you. Also, I cannot thank the "pdlg girls," Mina, Hyunju, and Emilie, enough for their advice and for listening to all my ideas, frustrations, and general ramblings on research and life.

I also want to acknowledge the National Science Foundation (grant BCS-0729140) and the National Institute of Health (NIDCD grant 02932) for providing the funding to make this project possible. In addition, a big thank you to all the participants who shared their time.

Last, but certainly not least, thanks to Patti, Mom, Dad, Eli, and Matthew for being the absolute best family and friends in the world. I couldn't have done any of this without you.

Introduction

Children learn to talk in a remarkably short period of time. A large body of research, including large cross-sectional studies as well as single-subject longitudinal studies (e.g., Smit, Hand, Freilinger, Bernthal, & Bird, 1990), have found that by the age of only five or six years old, children correctly produce most or all of the sounds of their language. The vast majority of these studies use transcription as the measure of the "correctness" or "incorrectness" of children's speech sounds. In fact, in the field of phonological development and disorders, transcription has long been the preferred tool (and often the only tool) to identify and characterize speech sounds.

This is true for both clinicians and researchers. In clinical practice, speech-language pathologists rely heavily on transcription throughout all stages of management. During assessment, clinicians use transcription when scoring standardized tests such as the Goldman Fristoe Test of Articulation-2 (Goldman & Fristoe, 2000) or the Photo Articulation Test (Lippke, Dickey, Selmar, & Soder, 1997). Frequently, a child's eligibility for services depends, at least in part, on the results of these tests. Additionally, much of the normative data on typical phonological development, to which clinicians compare the children with whom they work, was obtained using transcription. During intervention, clinicians use online judgments of correctness in order to give instructions and provide children with feedback on their performance. Transcription is also used to select treatment goals, to document change and eventually, to serve as criteria for dismissal from treatment. Research on phonological development and disorders has also relied heavily on transcription, often using trained phoneticians to make binary judgments on the accuracy of speech sounds. These judgments

are then used to study a wide variety of research questions, including those related to order and rate of acquisition, analysis of error patterns, efficacy of treatment programs, and many others.

Without doubt, transcription is a very useful analysis tool. First, it is an ecologically valid evaluation measure. Each time a talker speaks, the listener must be able to identify speech sounds in order to understand words and identify the linguistic content of the message. In this regard, transcribing a child's speech sounds parallels the process listeners must do each time the child speaks. Furthermore, transcription is an efficient measure. While speech can be recorded for later analysis, it is also possible for clinicians and researchers to make "on the fly" judgments while a child speaks. In clinical practice, this allows for timely feedback and efficient use of treatment time. In addition, transcription is a tool that can be easily understood by parents, teachers, and other members of an interdisciplinary team with whom clinicians work when providing treatment to children.

On the other hand, there also exist serious limitations with the use of transcription. First, there is significant evidence that listeners' expectations affect how they perceive speech. A second limitation is that when children are learning to speak, they do not necessarily progress directly from incorrect productions to correct productions. Instead, many children demonstrate a process of gradual change. In other words, as children's speech development progresses, speech sounds that were initially produced incorrectly gradually become closer to the adult form. Transcription, which requires a listener to map acoustic information present in the speech signal to corresponding phonemes and make binary judgments of the accuracy of speech sounds, may not be able to fully capture subtle

differences between a child's production and the target production. (See Kent, 1996 for a review of the limits of auditory-perceptual judgments in the assessment of speech disorders.)

Early theories of how we perceive speech, such as auditory theories of speech perception made it easy to justify transcription as an analysis tool. These theories posited that everything a listener needs to perceive speech is present in the acoustic signal (Diehl, Lotto, & Holt, 2004). According to these accounts, certain acoustic cues in the auditory signal are "decoded" by listeners to arrive at the corresponding phonemes. Extreme versions of this theory posited that there are invariant acoustic cues present in the speech of all speakers that map onto speech sounds with one-to-one correspondence (Blumstein & Stevens, 1981). In recent years, however, such theories have been called into question.

While undoubtedly, the acoustic signal is critical in the perception of speech, there also exists a substantial body of evidence suggesting that what listeners hear is not necessarily a direct correlate of the acoustic signal produced by the talker. For example, numerous studies have shown that what listeners perceive is significantly influenced by what they *expect* to perceive. 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.

For example, one line of research investigates how sociolinguistic expectations affect listeners' perceptions. Johnson, Strand, and D'Imperio (1999) found that expectations regarding the gender of a talker influenced vowel perception. Four talkers, two women and two men, were recorded reading the word "hud." One of the men was judged by a group of listeners to sound "stereotypically" male and one of the women was judged to sound

"stereotypically" female. For each speaker, the authors then synthesized a seven-step continuum of stimuli ranging from "hud" to "hood" by lowering the F1 values. The result was four separate continua that maintained the voice source characteristics of each of the original four talkers, but had identical formant trajectories. The audio productions were synced to short video clips of either a woman or a man speaking. A separate group of listeners were asked to watch the video clips and identify whether they heard the word "hud" or "hood." The authors found that both the gender of the voice and the gender of the talker on the video affected the phoneme boundary (on the F1 continuum) between /ʌ/ and /ʊ/. The average F1 value at the phoneme boundary was higher when listeners heard female voices and when they saw a female speaking in the video clip. Furthermore, voices that sounded stereotypically female had an even higher average F1 frequency at the phoneme boundary when compared to the male voices and the female voice that did not sound stereotypically female. Because the formant trajectories were identical for all the stimuli, this effect is the result of the listeners' expectations regarding gender.

Johnson et al. (1999) also found a similar effect when an ambiguous voice was used. The authors again created a seven-step F1 continuum from "hud" to "hood." Instead of pairing the audio stimuli with a video image, the authors told one group of listeners that the talker was female and told the other half that the talker was male. As in the first experiment, the listeners were again asked to identify whether they heard "hud" or "hood." When the listeners believed the talker was female, the average F1 value at the phoneme boundary was higher than when the listeners believed the talker was male. Again, because the stimuli were identical, the difference in perception was a result of the listeners' expectations about gender.

Sociolinguistic expectations related to regional dialect have also been shown to affect speech perception. For example, Niedzielski (1999) found that perception of the diphthong /aʊ/ depended on whether a listener believed the talker was from Detroit or Canada. In Canadian English, when /a/ occurs in this diphthong, it is produced with a higher, more forward tongue position that when it occurs alone. This phenomenon is called Canadian Raising. Canadian Raising is also present in the dialect of English spoken in Detroit, but Detroit speakers notice it only when they hear it in the speech of Canadians. In Niedzielski (1999), 41 adult Detroit-residents were asked to listen to sentences produced by a Detroit resident, whose speech contained the raised variant of /aʊ/. Half the listeners were told the talker was Canadian and half were told that the talker was from Detroit. Each sentence contained a word with the diphthong /aʊ/ and listeners were instructed to pay attention to this word. After hearing the sentence, listeners listened to six computer-synthesized variants of the diphthong /aʊ/ and were instructed to select the variant that best matched the diphthong produced by the talker. Each of the six variants differed in the values of the first and second formants. One of these variants corresponded to the actual diphthong (a raised variant) produced by the talker. One variant was a canonical /aʊ/ (without raising) and one was a lowered variant. Listeners who believed that the talker was from Detroit, matched the talker's /aʊ/ to the correct (raised) variant only 11 percent of the time. They matched the talker's /aʊ/ to either the canonical /aʊ/ or the lowered /aʊ/ 89 percent of the time. On the other hand, listeners who believed the speaker was Canadian matched the talker's /aʊ/ to the correct (raised) variant sixty percent of the time. Because all of the listeners heard the same Detroit

talker, the only thing that could account for the difference in perception was their belief about the regional dialect of the talker.

Hay, Nolan, and Drager (2006), found further evidence that expectations based on regional dialect affect listeners' perceptions. In this study, listeners from New Zealand perceived diphthongs differently depending on whether they believed the talker was from New Zealand or Australia. Furthermore, in a related study (Drager & Hay, 2006), the authors found that even the presence of a stuffed animal representing regional dialect (a stuffed kangaroo for Australia and a stuffed kiwi for New Zealand) affected how listeners perceived a constant stimulus. Additionally, Hay and colleagues examined the role of other types of sociolinguistic expectations, and found that the perceived age and social class of a talker also impacted listeners' perceptions of the unchanging stimulus (Hay, Warren, & Drager, 2006).

Expectations based on visual information have also been shown to have strong effects on the perception of speech. In a well-known study on the effect of visual information on perception, McGurk and McDonald (1976) showed that listeners use both the acoustic signal and visual cues to determine place of articulation of consonants in a CVCV (consonant-vowel-consonant-vowel) sequence. Audio and video recordings were made of a woman saying the sequences /baba/, /gaga/, /papa/, and /kaka/. The videos were then dubbed with the audio clips such that the videos of the woman producing velar stops were paired with the audio recordings of the woman producing labial stops, and videos of the woman producing labial stops were paired with audio recordings of the woman producing velar stops. Listeners were then asked to identify the consonants they heard. When adult listeners heard the recordings but could not see the television screen, they were very accurate in identifying the consonant. However, when listeners heard the recordings while watching the screen, they

were much less accurate. This was especially true when listeners heard /baba/, but saw the woman produce /gaga/. For this condition, adult listeners reported hearing /dada/ 98 percent of the time, leading the authors to conclude that the listeners "fused" the cues from the audio stimulus with the cues from the visual stimulus. Listener expectations can even play a role in whether an acoustic signal is heard as speech. Remez, Rubin, Pisoni, and Carrell (1981) synthesized three-tone synthetic sinusoidal replicas of an English sentence. These replicas preserved frequency and amplitude variation of the natural speech formants, but otherwise differed greatly from natural speech. For example, the synthetic replicas lacked the acoustic cues, such as fundamental frequency changes, steady-state formants, and formant transitions, that typically help listeners perceive stress, voicing, and place and manner of articulation. Two groups of listeners were asked to listen to the stimuli. One group was given no information about the stimuli and the listeners were asked to give their impressions of the stimuli. The majority of listeners did not identify the sounds as human speech. Instead they reported hearing sounds such as computer bleeps, science fiction sounds, music, etc. The listeners in the other group were told they would hear computer generated speech and were asked to transcribe it. The majority of these listeners not only heard the sentence as speech, but also accurately transcribed some or all of the words.

These studies on listener expectations make a strong case against the notion that there is a direct correspondence between the acoustic signal produced by the talker and the resulting perception of the listener. Further evidence can be found in studies exploring how previous experience with a talker's speech affects a listener's perception. For example, Nygaard, Somers, and Pisoni (1994) found that listener familiarity with a talker's voice

appears to improve speech perception. The authors trained 38 listeners to identify a speaker's voice by listening to a set of words spoken by ten different talkers. They found that the listeners were able to identify the talker both when they heard the words used in training and when they heard novel words. The listeners were then asked to identify novel words that were presented with noise. One group of listeners listened to the ten familiar voices and another group of listeners listened to ten unfamiliar voices. The authors found that the listeners who were asked to identify words spoken by the familiar voices were significantly better at identifying the novel words. Similarly, it has been documented that listener familiarity also improves perception of children's speech. For example, parents and siblings are better able to understand the speech of their child/sibling than other listeners who are not related (Weist & Kruppe, 1977). Listener familiarity has also shown to improve speech perception of foreign-accented speech. Bradlow and Bent (2003) provided native English-speaking listeners with experience listening to Chinese-accented English by asking them to transcribe English sentences spoken by a talker with a Chinese accent. Following two training sessions in which the English-speaking listeners transcribed these sentences, the authors administered a post-test in which listeners were asked to transcribe a new set of sentences spoken by a talker with a Chinese accent. They found that the perception of sentences significantly improved when listeners completed the training task for the same talker on which they were tested or when they completed the training task for multiple talkers with Chinese-accented English. Thus, it appears that listener familiarity resulted in improved speech perception, independent of any significant differences in the acoustic signal. Similar effects have been found from research in the field of speech disorders. For example, listener

familiarity has been shown to improve the perception of dysarthric speech (e.g., Tjaden & Liss, 1995) and of the speech of children with speech delays (Flipsen, 1995).

Even when listeners are not familiar with a talker, information about the talker's speech can still affect speech perception. For example, Ladefoged and Broadbent (1957) showed that listeners used acoustic information from a previously heard carrier phrase to help aid identifications in a single-word identification task. In this study, six versions of the carrier phrase, "Please say what this word is," were synthesized. All six versions were identical except that the range of formant frequency values differed for each version. The authors also synthesized four test words: *bit, bet, bat*, and *but*. These tests words were presented immediately following a version of the carrier phrase, and listeners were asked to identify the test word. The authors found that perception of the test word was greatly influenced by the carrier phrase it followed. For example, the stimulus word "bit" was perceived as "bit" 87 percent of the time when it was paired with one version of the carrier phrase, but was perceived as "bet" 90 percent of the time when it was paired with a different version of the carrier phrase. This strongly suggests that the perception of speech sounds is not solely based on the acoustic properties of the given speech sound, alone. Instead, listeners also use information about how the talker speaks, even if such knowledge is acquired from a single carrier phrase.

Together, the findings from these previous studies make several important points relevant to the use of transcription. First, it seems unlikely that there is a direct, one-to-one mapping of the acoustic signal of a given speech sound to its phoneme correlate. Secondly, listeners use a variety of information besides the acoustic signal of a given speech sound to perceive that sound. Listeners may use other acoustic information within the talker's speech,

such as information from previously heard speech (including carrier phrases), and familiarity with a talker's voice to decode what they hear. They may also use their own expectations based on sociolinguistic, visual, and other information to perceive speech sounds. Because transcription, in principle, requires objective identification of a given speech sound based purely on the acoustic signal, these findings suggest that we may want to rethink transcription as the sole tool in the analysis of children's speech or, at the very least, try to understand what other factors influence listeners' judgments about children's speech.

Besides the factors discussed above, there may be other factors that influence perception of children's speech. For example, speech pathologists are often asked to perform speech assessments because parents or other professionals are concerned about a child's speech. Simply as a result of this referral, the speech pathologist may have certain expectations regarding the child's speech that could influence how she/he perceives the child's speech. Such an effect was found in the field of voice disorders in which clinicians' expectations based on information from a client's history affected their perceptual judgments on a videostroboscopic examination of the vocal folds (Teitler, 1995). Podol and Salvia (1976) also found that visual appearance affected graduate students' judgments regarding the hypernasality of the speech of a child with a repaired a cleft lip. The authors created four conditions by pairing photographs of a child with a repaired cleft lip with speech samples of a child with typical speech and of a child with mildly hypernasal speech. One photograph showed the child with residual shortening of the lip and nares distortion following the repair. The other photograph was retouched such that no evidence of the repaired cleft lip was visible. The authors found that when the un-retouched photograph was paired with the mildly nasal speech condition, the clinicians rated the nasality higher than for the condition with the

retouched photograph and mildly nasal speech. In addition, the clinicians were more likely to recommend that the child receive speech therapy with the un-retouched photograph/mildly nasal speech condition.

Apart from the role of listener expectations, a second major limitation in the use of transcription is the fact that children do not necessarily progress directly from incorrect productions of speech sounds to correct productions. Although children's errors are commonly viewed as clear substitutions of another speech sound, research on gradient change in phonological acquisition (see Hewlett & Waters, 2004 for a review) indicates that this may not always be the case. Instead, this research suggests that that children's speech sounds gradually progress from immature forms to adult-like forms. As a result, children at times may produce forms that are intermediate to the target phoneme and another phoneme. Studies on covert contrast support this view. Covert contrast, which has been found in the speech of typically developing children and children with phonological disorders (e.g., Baum & McNutt, 1990; Scobbie, Gibbon, Hardcastle, & Fletcher, 2000) occurs when significant acoustic differences are present between two phoneme categories in children's speech. However, because both variants fall within a single adult perceptual category, transcribers perceive the two variants as the same phoneme.

For example, Baum and McNutt (1990) compared productions of /s/ and /θ/ produced by twenty children between the ages of five and eight years old. Ten of these children correctly produced both /s/ and /θ/, and ten of the children had frontal misarticulations of /s/. The authors performed acoustic analysis on these /s/ and /θ/ productions for children of both groups, focusing on measures of duration, amplitude, and spectral characteristics. Although

frontal misarticulations of /s/ are often described as substitution of /θ/ for /s/, the authors found that these misarticulated productions of /s/ differed significantly from correct productions of /θ/. In other words, these misarticulated productions were distinct from both correct productions of /s/ and correct productions of /θ/.

Some researchers (e.g., Stoel-Gammon, 2001) have suggested that one way to improve transcription reliability is to distinguish between intermediate productions (which are in between two sounds) and correct productions or clear substitutions. Unfortunately, researchers and clinicians rarely use an intermediate category. This may also have implications for transcription reliability because certain research suggests that transcribers are less reliable in identifying some misarticulated speech sounds. For example, Pye, Wilcox, and Siren (1988) found that when three different listeners transcribed a single child, they disagreed with each other significantly more often on some (but not all) misarticulated sounds as opposed to sounds the child produced correctly. Although Pye et al. (1988) do not specifically comment on the nature of these errors, it is possible that one reason that the transcribers disagreed with each other for these misarticulated sounds was that they were not always clear substitutions of another sound. Instead, some errors may have been intermediate between different phonemes.

Although substantial evidence supports the existence of intermediate productions (and that listeners are less reliable in perceiving misarticulated sounds), few studies specifically address how adults perceive these sounds in children's speech. For example, one might hypothesize that intermediate productions are more difficult to perceive, take longer to identify, or have lower inter- or intra-transcriber reliability than productions that more

closely approximate the prototypical, adult-like model. Unfortunately, little research has been conducted to address these possibilities. Of the few studies that do address this topic, the majority use synthesized child speech designed to simulate children's errors rather than natural productions. Furthermore, a large number of these studies focus exclusively on perception of productions along a continuum from /r/ to /w/. For example, Sharf, Ohde, and Lehman (1988) examined whether adult listeners were able to perceive tokens considered "distorted [r]." These were productions that had formant values intermediate to /w/ and /r/. The authors found that some listeners were able to reliably differentiate between /r/ and distorted /r/. They also concluded that perception of /w/ and /r/ does not appear to be categorical, as is the case for obstruent perception. Wolfe, Martin, Borton, and Youngblood (2003) also explored adult's perception of a synthetic child speech continuum from /r/ to /w/. Similar to the Sharf et al. (1988) study, they found that adult listeners were able to differentiate between /r/ and distorted [r]. Furthermore, they observed that experienced SLP graduate students with clinical experience were better able to identify subtle acoustic cues that signal whether a sound is closer to /r/ or /w/ (Wolfe et al., 2003). Interestingly, however, both studies found that training was not sufficient to increase the ability to make these distinctions. Sharf et al. (1988) found that a laboratory training session did not improve the ability to differentiate between /r/ and distorted [r] and Wolfe et al. (2003) found that simply completing a course in phonetics was not sufficient to improve these distinctions. Instead, the authors concluded that clinical experience is more useful in improving the ability to perceive subtle acoustic differences.

Little similar research has been conducted on other sounds, such as obstruents. This may be due to the fact that obstruents are perceived more categorically than sonorants (Fry, Abramson, Eimas, & Liberman, 1962, as cited in Sharf, et al., 1988). Nevertheless, just as clinicians are asked to perceive subtle differences between /r/, /w/, and distorted [r] in children's speech, they also must be able to perceive fine-grained differences between children's correct productions, clear substitutions, and intermediate productions of obstruents in order to provide effective treatment. Thus, it is critical to explore how adult listeners perceive children's intermediate productions of obstruents.

A similar paucity of evidence exists with regard to the role of listener expectations on the perception of children's speech, especially the perception of these intermediate productions. 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). Because children's speech is more variable than that of adults (Baum & McNutt, 1990) and may contain speech sound errors (e.g., Ingram, 1976) including ambiguous, intermediate productions, it seems likely that listener expectations may also have a large influence on the perception of children's speech.

The purpose of the present study was to explore how adults perceive children's correct productions of /s/ and /θ/, clear substitutions (/s/ for /θ/ and /θ/ for /s/), and intermediate productions (between /s/ and /θ/). In addition, we wanted to investigate the role of listener expectations in the perception of these productions. Specifically, we were interested in whether expectations about a child's age and the presence (or absence) of a phonological disorder might influence whether listeners identified children's productions as

correct or incorrect. Finally, we wanted to examine whether listeners with clinical experience would perceive these productions any differently than listeners without clinical experience. The /s/ and /θ/ sounds were chosen for several reasons. First, both are typically mastered relatively late in development (e.g., Sander, 1972 and Fudala & Reynolds, 1986, as cited in Peña-Brooks & Hedge, 2000; Smit et al., 1990). Additionally, children have often been observed to produce /θ/-like sound substitutions for /s/ (McGlone & Proffitt, 1973). 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. By including correct productions, clear substitutions, and intermediate productions, we essentially created a natural "continuum" of speech sounds ranging from /s/ to /θ/. Sounds at the center of the continuum were less easily classified as either /s/ or /θ/ by a trained listener, and we predicted that naïve listeners would also perceive these tokens differently than either correct productions or clear substitutions. Furthermore, for these difficult-to-classify productions, we hypothesized that listeners would rely more on other cues, namely their expectations about the children, to inform their perceptions of the sounds.

These hypotheses have several important implications for research and clinical practice. First, if listeners perceive intermediate productions differently from clear substitutions and correct productions, this may indicate that intermediate productions are a valid category to use during transcription. Secondly, if there is a significant effect of listener expectations on adults' accuracy judgments for children's speech (especially for adults with clinical experience), then this would suggest that we need to reconsider the use of

transcription as the only tool for judging accuracy in the assessment and treatment of children with speech disorders.

Experiment 1

Experiment 1 was used to choose the carrier phrases for experiment 2. We wanted carrier phrases that would convey to listeners information about the child's age and the presence or absence of a phonological disorder.

Methods

Stimuli:

The carrier phrase "I really like" (/aɪrililaik/) 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 [aɪwiwijaik]. 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 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 1.

Insert Table 1 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 1a). 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 1b). The listeners responded by pressing the appropriate number key on the keyboard.

Insert Figures 1a and 1b 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] = .154, p = .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 2 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 = .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 < .001, partial-eta-squared = .956) and a significant main effect of

F_0/formant values (F[2,38] = 56.05, p < .001, partial-eta-squared = .747). There was also a significant interaction between the two independent variables (F[2,38] = 14.54, p < .001, partial-eta-squared = .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_0/formant value condition. For the disorder ratings, there was a significant main effect of speech errors (F[1,19] = 618.413, p < .001, partial-eta-squared = .97), but not of F_0/formant values (p = .07). There was also a significant interaction between the two independent variables (F[2,38] = 8.71, p = .001, partial-eta-squared = .314). Again, this interaction was due to the fact that the disorder rating difference for the two speech-error groups was somewhat smaller for the raised F_0/formant value condition.

Insert Figure 2 here.

Discussion

The results of this experiment suggest that when listeners were asked to judge the age of the child, they were influenced both by the F_0 and formant values of the carrier phrase and by the presence or absence of phonological errors within the phrase. On the other hand, when listeners judged how adult-like the child's speech sounded, they were influenced only by the presence or absence of phonological errors. Nevertheless, 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 experiment 2, 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.

Experiment 2

Methods

Stimuli:

For this experiment, 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^{h}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. All of the words were transcribed by a native speaker of English (the author). The CV syllables that were selected either contained 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/. Table 2 shows the full inventory of the CV stimuli. Each CV syllable was normalized for amplitude.

Insert Table 2 here.

Carrier phrases were chosen based on the results of experiment 1. 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 ([aɪrililaik]) 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 ([aɹwiwijaik]) 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 3 shows which carrier phrases were used (and those that were not used) from the complete set of carrier phrases that we constructed.

Insert Table 3 here.

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. (Response times were not analyzed for the present study, but will be addressed in an upcoming paper.) Listeners responded by pressing buttons 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

Our first analysis focused on whether listener responses were affected by their experience, the carrier phrase condition, and the transcription categories. Figures 3a and 3b 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, $p < .001$, partial-eta-squared = .996) of transcription category was observed. Post-hoc paired comparisons revealed significant differences between all transcription categories ($p < .001$ for all ten comparisons). The main effect of carrier phrase was not significant (F[1, 28] = .015, $p = .90$). Similarly, the main effect of listener group was also not significant (F[1, 28] = .907, $p = .35$). The only significant interaction was between carrier phrase and listener group (F[4, 25] = 3.11, $p = .033$, partial-eta-squared = .332). Visual inspection of the data shows that this interaction is due to the fact that for the undergraduates, there was a higher percentage of "correct [s]" responses for the "older-typical" carrier phrase condition, whereas, for graduate students, there was a higher percentage of "correct [s]" responses for the younger-disordered carrier phrase condition.

Insert Figures 3a and 3b here.

Our second analysis focused on only those stimuli for which intra-subject disagreement was observed across the two carrier phrase conditions. Figure 4 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, $p < .001$, partial-eta-squared = .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 ($p < .001$). 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 4 gives the number of trial pairs for which 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. 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 (X^2 = 12.95, p = 0.012).

Insert Figure 4 and Table 4 here.

Discussion

In designing this study, we had several primary questions. First, we wanted to examine how naïve adult listeners perceive children's correct productions of /s/ and /θ/, clear substitutions (/s/ for /θ/ and /θ/ for /s/), and intermediate productions (between /s/ and /θ/). Our results confirmed that naïve listeners' responses to each of these five transcription categories patterned differently. In other words, the mean percent of time that the 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. This result is important in several ways. First it validates our original transcription categories in that as a group, naïve listeners' judgments paralleled our original transcriptions. Secondly, it provides support for the existence of covert contrast because the average percent correct /s/ responses were significantly lower for [s] substitutions for /θ/ than for correct /s/ productions.

Likewise, correct /θ/ productions were significantly less likely to be judged as a correct /s/ than [θ] for /s/ substitutions. Finally, we also found that productions transcribed as "intermediate" had an intermediate level of mean percent correct /s/ responses. The correct [s] judgments for this category were less than [s] for /θ/ substitutions and higher than the [θ] for /s/ substitutions. Thus, it appears that "intermediate" may be a valid transcription category. However, it must be noted that although we found a gradient change in mean percent correct /s/ responses across all five transcription categories, this only reflects a group effect, rather than the judgments of individual listeners. Nevertheless, it is 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. For example, asking listeners to rate each CV production using a rating scale, such as a visual analog scale or direct magnitude estimation, might provide insight into how well individual listeners are able to differentiate tokens of each transcription category. In fact, ongoing research by Munson and colleagues (Uberg-Carlson & Munson, 2008) are currently underway at the University of Minnesota to explore these questions.

Based on our results, we have evidence to support that dividing our CV productions into five transcription categories was an effective way to classify children's productions of /s/ and /θ/. The next step is to perform an acoustic analysis on productions in each category to determine whether acoustic evidence also supports the use of these categories. Because listeners, as a group, judged clear substitutions differently from correct productions, an acoustic study designed to compare substitutions of [s] for /θ/ with correct productions of /s/ would provide information on how these tokens differ. Likewise, an acoustic comparison of [θ] for /s/ substitutions and correct /θ/ is needed. In essence, these analyses would be a replication study of Baum and McNutt (1990) study on covert contrast. In addition, acoustic analysis of intermediate productions is needed. Our results suggest that these intermediate productions really do exist. However, we do not know what distinguishes them acoustically from correct productions and clear substitutions. Although errors on the /s/ -/θ/ continuum are commonly viewed as "fronted" (as in [θ] substitutions for /s/) or "backed" (as in [s] substitutions for /θ/), errors may actually vary on a variety of dimensions (such as intensity, spectral cues, and onset bursts). This may be especially true for intermediate productions. For example, it is possible that productions that are classified as intermediate contain certain acoustic cues that are more /s/-like and others that are more /θ/-like. This might also explain our finding that intermediate productions were more likely to be rated differently by individual listeners each time they were presented.

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 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 for these most ambiguous productions, when listeners don't expect the child to be able to produce a correct /s/ (because the child is young or has difficulty producing speech sounds correctly), they are more lenient in what they consider correct. On the other hand, when listeners expect to hear a correct /s/ (because the child is older and more capable of producing speech sounds correctly), they have a stricter guidelines for what constitutes a correct /s/.

Clinically, this is an important finding. First, we found that intermediate productions are more likely to be rated inconsistently, as compared to correct productions or clear substitutions. Next, we discovered that when intermediate productions were judged inconsistently, these productions were also the most likely to be subject to listener bias. 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/. 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.

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. Some listeners also commented that in the beginning of the task they paid more attention to the carrier phrase and later paid less attention to it, possibly because there was not enough variability within the 16 phrases. (Future analyses are planned to investigate the effect of order of presentation on listeners' judgments.) Also, 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. Finally, to create the entire stimulus, we merged the carrier phrase sound file with the CV sound file. In some cases, there was a short pause between the two, which could signal listeners to the fact that the carrier phrase and CV sequence was not a cohesive unit produced by a single child.

A methodology that minimizes these problems might yield a greater effect of listener expectations. For example, future studies might collect 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. For example, a sample of a child's narrative or conversational speech could be played before the CV judgment task. Listeners would develop expectations about the child based on speech patterns in the sample. Listeners (especially those with clinical experience) could also be provided with a case history for each child to determine how a clinician's knowledge of history and potential risk factors affect perception. Finally, information about the children could be provided to listeners explicitly. For example, one group of listeners might be told that all of the children are suspected of having a phonological disorder. Another group of listeners could be told that the children are believed to have typically developing speech. This method has the additional strength of being ecologically valid in that clinicians are often asked to assess children about whom parents or teachers are concerned. Thus, the clinician expects that the child may have difficulty producing speech sounds before even beginning the assessment process.

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.

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. Nevertheless, the results of this study strongly support the use of an additional transcription category for intermediate productions. Such a category may be particularly useful for better understanding listener bias and within listener variability.

References

Baum, S. R. & McNutt, J. C. (1990). An acoustic analysis of frontal misarticulation on /s/ in children. *Journal of Phonetics,* 18, 51-63.

Blumstein, S. E. & Stevens, K. N. (1981). Phonetic Features and acoustic invariance in speech. *Cognition,* 10, 25-32.

Boersma, P. & David Weenink (2005). Praat: doing phonetics by computer (Version 4.3.28) [Computer program].Retrieved from http://www.praat.org/.

Bradlow, A. R., & Bent, T. (2003). Listener adaptation to foreign-accented speech. In M. J. Sole, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2881-2884). Barcelona: Futurgraphic.

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. (in review, 2008). Methodological questions in studying phonological acquisition. Submitted to *Clinical Linguistics and Phonetics*.

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.

Flipsen, P. (1995). Speaker-Listener Familiarity: Parents as Judges of Delayed Speech Intelligibility. *Journal of Communication Disorders,* 28, 3-19.

Goldman, R. & Fristoe, M. (1986). *Goldman-Fristoe Test of Articulation.* Circle Pines, MN: American Guidance Service.

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. & 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.

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.

Ladefoged, P. & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America,* 24, 629-637.

Lippke, B. Z., Dickey, S. E., Selmar, J. W., & Soder, A. L. (1997). *Photo Articulation Test–3*. Austin, Tx: Pro-Ed.

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.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology,* 18, 62-85.

Nygaard, L. C., Sommers, M. S. & Pisoni, D. B. (1994). Speech Perception as a Talker Contingent Process. *Psychological Science,* 5, 42-46.

Peña-Brooks, A. & Hedge, M. N. (2000). Assessment and Treatment of Articulation and Phonological Disorders in Children. Austin, Texas: Pro-Ed Inc.

Podol, J. & Salvia, J. (1976). Effects of visibility of a prepalatal cleft on the evaluation of speech. *The Cleft Palate Journal,* 13, 361-366.

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.

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.

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.

Stoel-Gammon, C. (2001). Transcribing the Speech of Young Children. *Topics in language disorders,* 21, 12-21.

Teitler, N. (1995). Examiner bias: influence of patient history on perceptual ratings of videostroboscopy. *Journal of Voice*, 9, 95-105.

Tjaden. K. & Liss. J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics and Phonetics*, 9, 139-154.

Uberg-Carlson, K., & Munson, B. (2008, Nov) *Assessment of Phonetic Skills in Children 2: A comparative study of methods to elicit gradient judgments of children's accuracy.* Poster to be presented at the American Speech-Language-Hearing Association (ASHA) Convention, Chicago, IL.

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.

Weist, R. & Kruppe, B. (1977). Parent and sibling comprehension in children's speech. *Journal of Psycholinguistic Research,* 6, 49-58.

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. Carrier Phrase Types and Total Number Used in Experiment 1.

	Higher F_0 and formants	Unchanged F_0 and formants	Lower F_0 and formants
Error-free	**"I really like"** [aɪrililaik] Total Number: 4	**"I really like"** [aɪrililaik] Total Number: 4	**"I really like"** [aɪrililaik] Total Number: 4
Speech sound errors	**"I weawwy yike"** [aɪwiwijaik] Total Number: 5	**"I weawwy yike"** [aɪwiwijaik] Total Number: 5	**"I weawwy yike"** [aɪwiwijaik] Total Number: 5

Table 2. Stimuli inventory: Total number of all consonant-vowel sequences organized by age, vowel context, and transcription category.

Following Vowel	[θ] substitutions for /s/				Correct /θ/				Intermediate Productions (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	1	1	2	0	0	4	13	14	2	2	4	1	**44**
E	0	4	1	0	0	0	0	0	1	2	1	1	**10**
A	4	4	0	1	0	1	1	2	1	2	2	2	**20**
O	2	1	0	0	0	0	0	0	1	1	2	0	**7**
U	0	3	0	0	0	2	5	4	1	1	2	1	**19**
	Total: 24				**Total: 46**				**Total: 30**				**100**

Following Vowel	[s] substitutions for /θ/				Correct /s/				Intermediate Productions (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	2	3	4	3	3	2	4	5	1	2	4	1	**34**
E	0	0	0	0	2	2	4	2	1	1	1	2	**15**
A	1	2	1	0	0	2	2	4	0	2	2	0	**16**
O	0	0	1	0	2	1	3	2	2	2	2	0	**15**
U	1	2	4	0	1	2	4	3	1	2	0	0	**20**
	Total: 24				**Total: 50**				**Total: 26**				**100**

Table 3. Carrier Phrase Conditions used in Experiment 2.

	Younger Child	Intermediate-Aged Child	Older Child
Typically Developing		CONDITION 1: "older-typical"	
	"I really like" [aɪrililaik]	**"I really like"** [aɪrililaik]	**"I really like"** [aɪrililaik]
	F_0 and Formants raised.	F_0 and Formants unchanged.	F_0 and Formants lowered.
	Total Number: 0	Total Number: 4	Total Number: 4
Phonologically Disordered	CONDITION 2: "younger-disordered"		
	"I weawwy yike" [aɪwiwijaik]	**"I weawwy yike"** [aɪwiwijaik]	**"I weawwy yike"** [aɪwiwijaik]
	F_0 and Formants raised.	F_0 and Formants unchanged.	F_0 and Formants lowered.
	Total Number: 4	Total Number: 4	Total Number: 0

Table 4. 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

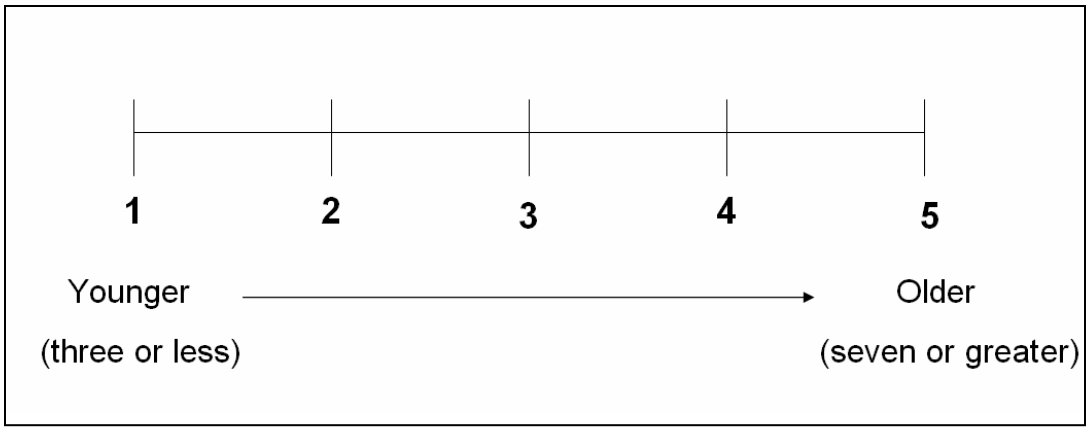
Figure Captions

Figures 1a and 1b. Visual display listeners used in experiment 1 to judge how old a child sounded (top plot) and to judge how adult-like a child's production sounded (bottom plot).

Figure 2. 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.

Figure 3a and 3b. Mean percent correct [s] responses for each transcription category plotted separately for carrier phrase (Fig. 3a) and for listener group (Fig. 3b) (i.e., Figure 3a 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 3b 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 4. 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."



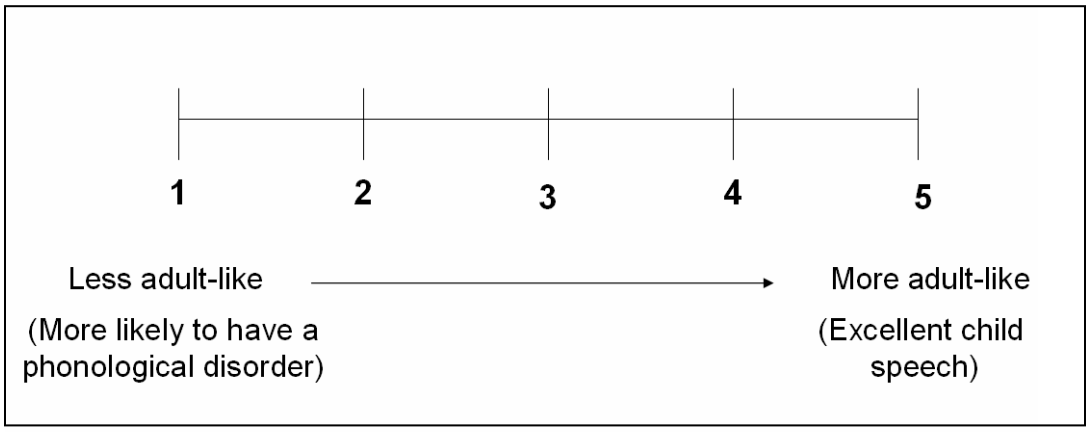


Figure 1a (top) and 1b (bottom).

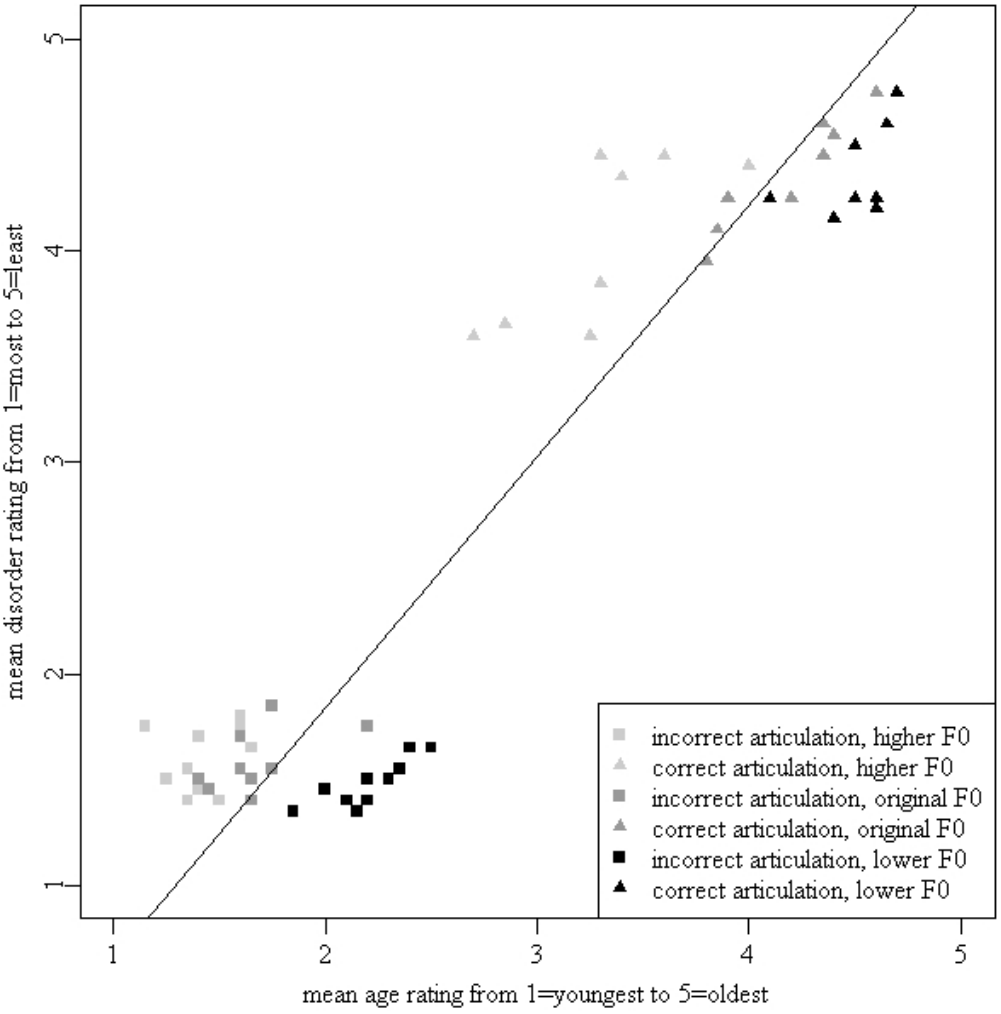
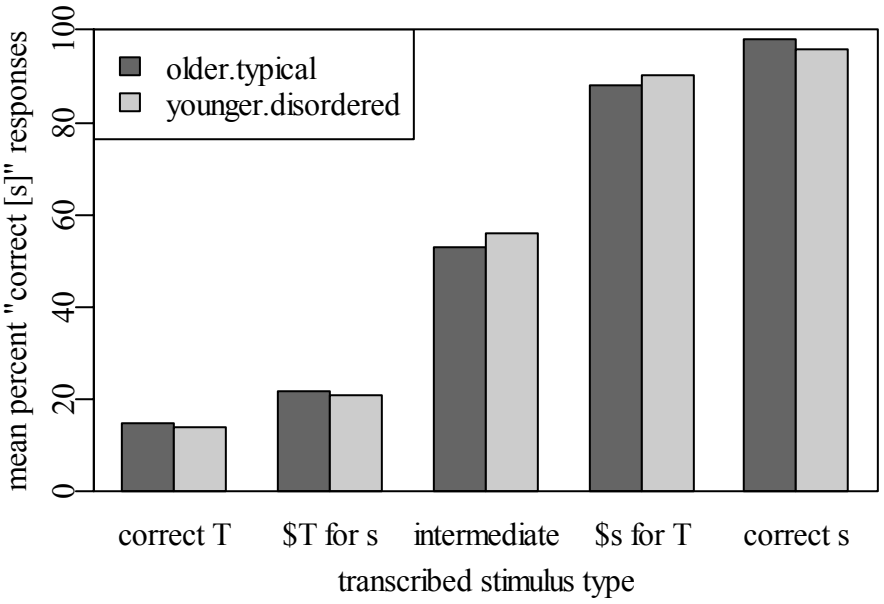

mean disorder rating from 1=most to 5=least
mean age rating from 1=youngest to 5=oldest
1
2
3
4
5
incorrect articulation, higher F0
correct articulation, higher F0
incorrect articulation, original F0
correct articulation, original F0
incorrect articulation, lower F0
correct articulation, lower F0

Figure 2.



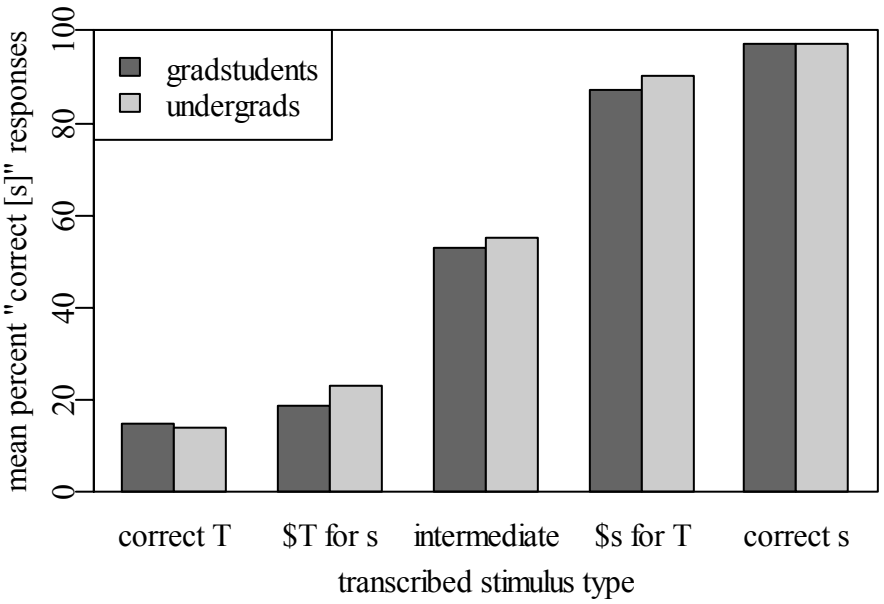


Figure 3a (top) and 3b (bottom).

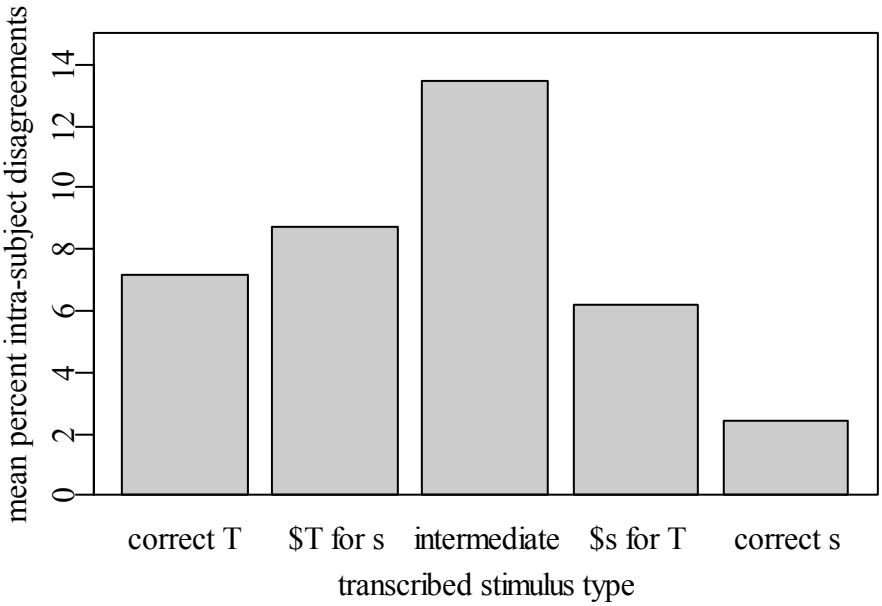

mean percent intra-subject disagreements
0
2
4
6
8
10
12
14
correct T
$T for s
intermediate
$s for T
correct s
transcribed stimulus type

Figure 4.