PERCEPTUAL VALIDATION OF AN ACOUSTIC ROBUSTNESS OF CONTRAST MEASURE

by

Ryan Mary Sovinski

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

Masters of Science

(Communicative Disorders)

at the

UNIVERSITY OF WISCONSIN-MADISON

2011

Acknowledgement

This thesis is the culmination of my Master's degree in Speech-Language Pathology in the Department of Communication Disorders at the University of Wisconsin-Madison. I would like to briefly thank the people that have supported me throughout this enter process.

To begin, I would like to extend the utmost gratitude and appreciation to my advisor, Dr. Jan Edwards, whose patience and guidance have been invaluable. I would like to thank the ever patient Dr. Eun Jong Kong, without whom my thesis would have never been completed. I would also like to thank my committee members Dr. Marios Fourakis and Dr. Rita Kaushanskaya for their eagerness and flexibility with my "shot gun" prospectus and thesis defense. I would also like to extend the greatest thanks to Joan Kwiatkowski for her belief in me and her support throughout the past two years.

My sincerest thanks go to my mother, Teresa Markley, my sister, Christine Jaynes, my aunt and uncle, Linda and Matt Radecki, and my grandfather, Thomas McClanahan, without whom a Master's degree would be nothing but a dream. They have always encouraged my educational endeavors and faultlessly believed in me. Without their love and support I would not be where I am today.

Lastly, I owe a special thanks to my partner, Nadia Beddawi. She is a rock in a storm of doubt, insurmountable stress, and unbelievably fast approaching deadlines. It is with immense appreciation of her unwavering understanding and bottomless cups of coffee that I thank her for being a part of my life and my thesis.

To those mentioned, I dedicate this thesis and my Master's degree. Thank you for everything that you have done, continue to do, and will do. You have touched my life in a ways that words cannot express. Thank you.

Table of Contents

INTRODUCTION

Phonetic transcription is the gold standard for assessing the correctness of speech sounds. . Transcription is relied heavily upon in clinical and research settings. Transcription is used clinically when evaluating speech sound disorders in children with both organic and functional problems and it is also used to evaluate dysarthria and apraxia in adults. Phonetic transcription has the advantages of being fast, efficient, and ecologically valid. Clinically, assessments of articulation such as the *Goldman-Fristoe Test of Articulation-2nd edition* (Goldman & Fristoe, 2000), the *Photo Articulation Test* (Lippke et al., 1997) and the *Arizona Articulation Proficiency Scale, 3rd edition* (Fudala, 2003) are commonly used to determine the correctness of various speech sounds. Clinician judgments on these tests are used to determine 1) need for services, 2) continuation of services, and 3) dismissal from services. The results are often used to create therapeutic goals and objectives; in addition, continued evaluation of speech targets dictates the procession from one objective to another based. In research, trained phoneticians make judgments of the accuracy of speech sounds. Their responses are used to answer various research questions ranging from phonological development to treatment efficacy. However, research has shown that phonetic transcription is a subjective measure that is dependent upon experience, training, and expectations of the transcriber.

There is a large body of literature describing the limitations of transcription. Transcription judgments are influenced by a variety of listener expectations. For example, judgments may be influenced by gender expectations(Johnson, Strand & D'Imperio, 1999), regional dialect expectations (Niedzielski, 1999; Hay, Nolan, &

Drager, 2006a; Drager & Hay, 2006), sociolinguistic expectations such as age and social class (Hay, Warren, Drager, 2006b), by referral to a speech-language pathologist (Teitler, 1995), and lexical knowledge (Kent, 1996). Additionally, Kent (1996) warns about phonemic restoration and verbal transformation during transcription. Phonemic restoration occurs when a listener is unaware of a target sound having been replaced by a non-speech sound (e.g., a cough or a glottal stop) by a speaker. This phenomenon is seen frequently in normal speech which is not free of errors. When applied to disordered speech, the transcriber must be even more careful not to lose information from the acoustic signal. Verbal transformation results from a transcriber replaying an audio sample several times resulting in their original perception then becoming altered. Thus, clinicians may mishear the speaker in the course of phonetic transcription.

There is also high variability between transcribers. Even specialists disagree on how to optimally judge disordered speech with less reliability when audio-perceptual measures are used without accompanying visual information (Kent, 1996). Given their individual and different experiences, transcribers are likely to approach each speech dimension differently. The transcription errors and disagreements among raters may provide additional information about a speaker's phonological system. Pye, Wilcox, and Siren (1998) found that the most common differences among three transcribers were the addition or omission of certain sounds and changing features of sounds. They posit that the addition of segments may reflect the bias of the transcriber to complete words and sentences.

Another area of concern regarding phonetic transcription is the familiarity of the listener with the speaker. Nygaard, Pisoni, and Somers (1994) found that increased

familiarity with a speaker's voice may affect and, in fact, improve the listener's perception of the speaker's speech. This is observed clinically when parents report that their child's speech is intelligible to family members but unintelligible to unrelated persons (Weist & Kruppe, 1977). This phenomenon has been observed in individuals with voice disorders (Nygaard, Pisoni, & Somers, 1994), dsyarthric speech (Tjaden & Liss, 1995), and in children with speech sound delays (Flipsen, 1995).

An additional problem for transcription is the existence of covert contrast, a subphonemic contrast between two sounds that is not detectible in most listening tasks, but which can be observed when using acoustic analysis. In a longitudinal study, Macken and Barton (1980) observed that all three children that they studied produced a covert contrast between voiced and voiceless stops (a significant difference in voice-onset-time [VOT], although all VOT's were within the adult voiced range) before they produced an overt contrast. This suggests a gradual move toward an adult-like production. Because covert contrast is not always perceptible to the listener's ear, combining transcription with acoustic analysis would provide more information on a child's speech production than transcription alone (Maxwell & Weismer, 1982; Scobbie & Gibbon, 1997; Li, Edwards, & Beckman, 2009). Children presenting with a covert contrast acquire the adult production in the therapy clinic faster than those children who do not (Tyler, Figurski, & Langdale, 1993). If a child has productive knowledge of a target, the treatment may focus on articulatory movements whereas if a child is lacking the proper cues to signal differences between phonemes, treatment should focus not only on articulatory movements but on phonetic cues as well (Gibbon & Scobbie, 1997).

While acoustic analysis is very time-intensive, listeners can also perceive convert contrast with some perception tasks, Visual analog scales (VAS) offer benefits to clinicians and researchers (Urberg-Carlson, Kaiser, & Munson 2008; Johnson, Munson, & Edwards, 2010). In a VAS rating task, individuals are asked to scale a psychophysical parameter by indicating their percept on an idealized visual display. Munson and colleagues have used VAS tasks in which naïve listeners are asked to judge how well children produce a consonant contrast. They present participants with an arrow with endpoints labeled as, for example, "the s sound" and "the sh sound." Then subjects are asked to judge whether a sound is more /s/-like or more /ʃ/-like by clicking somewhere on the line. In a series of studies (Urberg-Carlson, Kaiser, & Munson 2008; Urberg Carlson, Munson, Kaiser, 2008; Johnson, Munson, & Edwards, 2010), Munson and colleagues found that VAS was valid for both naïve and trained listeners for a variety of contrasts. They found that VAS judgments were closely related to judgments of a trained transcriber and to acoustic measures that differentiated between two sounds. Johnson, Munson, and Edwards (2010) also found that clinicians' ratings on the VAS were more closely related to acoustic measures than non-clinicians, indicating that clinical experience makes them more sensitive to the phonetic detail in speech productions. A similar measure is Direct Magnitude Estimation (DME), in which listeners rate a quality of the stimulus on an equal appearing interval scale. DME has been shown to be similar to VAS in judging between two sounds (Urberg-Carlson, Kaiser, & Munson 2008) and it has also been found to be an appropriate tool for measuring speech naturalness in individuals who stutter (Metz, Schiavetti, & Sacco, 1990). DME was also found to be a

valid way of quantifying the intelligibility of speakers with hearing impairment (Schiavetti, Metz, & Sitler, 1981).

Acoustic measures to quantify the degree of contrast between two sounds have also been proposed. In a recent paper, Holliday, Beckman, and Mays (2010) proposed a method for quantifying the robustness of the contrast between /s/ and /ʃ/ using a logistic regression model with peak ERB as the dependent variable. Peak ERB is a psychophysical measure of the highest spectral peak (Moore & Glasberg, 1987). This measure is similar to the more commonly used acoustic measure of centroid, but it is less sensitive to the changes in the oral cavity than centroid, an indirect measure of resonance in the vocal tract, which masks the size of the vocal tract when there is a loose constriction of the articulators. Holliday et al. (2010) built individual step-wise logistic regression models for 20 speakers in five age groups (adults, 5-year-olds, 4-year-olds, 3-year-olds, and 2-year-olds). They found that the models for the adults had almost a perfect step function, with above 90% correct classification based on peak ERB alone for 18 out of 20 adults. The results were similar for the five-year-olds, with steep slopes differentiating /s/ and /ʃ/ for most of the children, and above 80% correct classification for 17 out of 20 children. By contrast, the individual step-wise logistic regression models for the 2- and 3-year-olds had much shallower slope values and lower classification success. For example, for the 2-year-olds, only 2 out of 20 speakers had classification values above 80%. All of the children's productions were included in the models as long as they were produced as fricatives, so the poor classification results for the 2-year-olds may be because a common substitution pattern for young English-speaking children is [s] for /ʃ/ (Li, Edwards, & Beckman, 2009).

The Holliday et al. (2010) result is interesting, but it suffers from two limitations. First, it is impractical for clinical use, as it requires a time-consuming psychophysical analysis of the children's productions. Second, Holliday and colleagues did not provide data to show that the slope differences observed were perceptible. Thus, the purpose of this study is to examine whether the slope differences observed by Holliday et al. are perceptually valid. If this is found to be true, it is important for two reasons. One, it validates a quantitative measure of robustness of contrast in speech production that may be useful for research studies. In addition, it will also provide validation for the clinical use of goodness ratings to judge the robustness of contrast in individual children.

The purpose of this study was to use goodness ratings by naïve adults to validate Holliday et al.'s robustness of contrast in production measure. We predicted that if this robustness measure is valid, then the productions of children characterized by steep slopes would have higher goodness ratings than the productions of children characterized by shallow slopes. This study differs from previous work in that it directly compares a production measure of robustness of contrast with naïve adults' ratings of the same sounds.

METHODS

Stimuli

The stimuli were word initial consonant-vowel (CV) syllables excised from real words and nonwords produced by 2- to 5-year-old English speaking children. All of the stimuli were elicited with a picture-prompted auditory word repetition task and were

taken from a large database of 2- to 5-year-old American English speakers (Edwards & Beckman, 2008a, b). In this database, there were multiple productions of real words and nonwords with word-initial /s/ and /ʃ/ before the vowels /a, e, E, I, o, u/. While Holliday et al. used all productions that were transcribed as correct or as a fricative substitution in their acoustic analysis, for the perception study, we used only those productions that had been transcribed as correct or "intermediate" by a trained native speaker/phonetician. Intermediate productions were those in which the transcriber heard the consonant as in between two sounds, for example [s:θ] meant that the consonant was transcribed as in between /s/ and /θ/, but more like /s/. We included intermediate productions only if the sound was "more like /s/" or "more like /ʃ/" and the alternative sound was also a fricative. We excluded stimuli from children whose slope was in the wrong direction. We included productions from nine 2-year-olds, nine 3-year-olds, eight 4-year-olds, and eight 5 years-olds. For each age (2-, 3-, 4-, and 5-years), we selected stimuli by child. We chose children with either steep or shallow slopes based on the results of Holliday et al. (2010). Insofar as possible, we selected children so that 50% of children in each age group would have steep slopes and 50% of children would have shallow slopes. Figure 1 shows the Holliday et al. analysis for each of the subjects whose stimuli were included in the perception experiment. As described above, a steep slope indicates a more robust contrast than a shallow slope. Because the younger children (2- and 3- years) had fewer correct productions, we used all of their correctly transcribed /s/ and /ʃ/ CV sequences if there were fewer than 6 correct fricative productions in each category. The CV sequences from the older children were randomly selected, as long as they were correctly transcribed and were from a child with either a steep or shallow slope. We also

considered the quality of the recording and the lack of background noise in choosing CV sequences. Insofar as possible, the CV sequences were matched by age (2 through 5) and gender, as well as the slope of /s/ and /ʃ/ contrast so that there were an equal amount of tokens from children with steep and shallow slopes for each age and consonant. However, as children mature, the contrast between the two targets becomes more robust. Therefore, there is not as large a difference between the steep and shallow slopes for the 5-year-olds as there is for the 2-year-olds

Tables 1a and 1b provide a list of children whose stimuli were included in the experiment, along with demographic information, slope value, and number of /s/ and /ʃ/ tokens. For subjects with a steep slope, the average age was 47 months (SD = 13 months), there were 9 males and 9 females, the average slope was -4.262 (SD = 6.313) and there were 103 correct /s/ tokens and 108 correct /ʃ/ tokens. For subjects with a shallow slope, the average age was 47 months (SD = 13 months), there were 8 females and 8 males, the average slope was - 0.1739 (SD = - 0.2476), and there were 83 correct /s/ tokens and 82 correct /ʃ/ tokens produced by the selected subjects.

Participants

The participants in the study were 20 (7 male, 13 female) young adults currently enrolled in introductory course in the Department of Communicative Disorders at the University of Wisconsin-Madison and received course credit for participation. Based on self-report, participants had no history of hearing loss or a speech and language delay or disorder.

Procedure

The adults were seated in a quiet room in front of a computer in the Waisman Center at the University of Wisconsin-Madison. They listened to all stimuli over headphones. A training session preceded the test session. For both the training and the test sessions, adults listened to two blocks consisting of either all /s/-initial or all /ʃ/ -initial CV sequences. Half of the participants began the experiment with the /s/-initial block and half began with the /ʃ/ -initial block. After hearing the stimulus item the participants were asked to make a goodness rating by clicking a mouse on a line indicating if each token they heard was a “good,” “bad,” or somewhere in between production of the target sound (see Figure 2).

The participants were provided with visual instructions before the practice items were presented. The participants read through the four slides describing the task and instructions by clicking a mouse at their own pace. The slides read:

> In this experiment, you will listen to two types of consonant-vowel sequences, namely "so" as in soap and "sho" as in show. You will hear the sequences in two different segments. One will have all "s" initial sounds and the other will have all "sh" initial sounds. After hearing each stimulus, you will indicate how good of a "s" or a "sh" it was by clicking on a line. When you hear what you think is a GOOD example of a "s" or a "sh" sound, click on the line close to where it says "good 's'" or "good 'sh'." When you hear what you think is a BAD example a "s" or a "sh" sound, click on the line close to where it says "bad 's'" or "bad 'sh'." Sometimes, you won't be sure the syllable began with a good or bad "s" sound or "sh" sound. In those cases, you should click a place on the line to show whether you thought it sounded like a better or worse example of a "s" or

> a "sh" sound. If the sound wasn't really good or bad but sounded more like a good example of a "s" or a "sh" sound than a bad example, you should click somewhere on the line closer to the text that says "good 's'" or "good 'sh'." If it sounds more like a bad example of a "s" or a "sh" sound than a good example, you should click closer to the text that says "bad 's'" or "bad 'sh'." We encourage you to use the whole line. That is, don't just click at the ends, click at the location on the line that corresponds to how good of an example you think the consonant was. We don't have any specific instructions for what to listen for when making these ratings. We want you to go with your 'gut' feeling about what you hear at the beginning of the syllables.

After the participants read the instructions, they practice items were administered. The examiner stayed in the experiment room through the practice session to ensure that all participant questions were answered.

RESULTS

We transformed the pixels of the click locations along the arrow into generalized logit values. We then examined the relationships between goodness ratings and the two robustness-of-contrast groups (steep slope vs. shallow slope) for /s/ and /ʃ/ separately. For these analyses, we ran a series of regressions predicting the goodness rating of either /s/ or /ʃ/. In the first set of logistic regression analyses, the dependent variable was the transformed logit values of the click locations from the listeners' goodness rating and the independent variables were robustness-of-contrast group (shallow slope vs. steep slope), speaker-age in months (range = 25 to 69 months), and the robustness-of-contrast group by age interaction. The results of these analyses are presented in Table 2 and plotted in

Figure 3. For both goodness ratings for /s/ and for /S/, the intercept was significant, indicating that there was a significant difference for goodness ratings for productions with steep vs. shallow slopes. None of the independent variables were significant predictors of goodness ratings for /s/. However, there was a significant effect of age on goodness ratings for /ʃ/. As expected, goodness ratings increased as age increased. A visual inspection of the data shows that little variability in goodness ratings for /ʃ/ was observed after the speakers reach 48 months. Therefore, in the second set of analyses, we analyzed data only for stimuli for which the speaker was 48 months or younger.

For this second set of analyses, again, the dependent variable was the transformed logit values of the click locations from the listeners' goodness rating and the independent variables were robustness-of-contrast group, speaker-age in months, and the robustness-of-contrast group by age interaction. The results of these analyses are presented in Table 3 and plotted in Figure 4. Again, none of the independent variables were significant predictors of goodness ratings for /s/. For /ʃ/, both age, robustness-of-contrast group, and the interaction between the two independent variables were significant predictors of goodness ratings for /ʃ/. As before, goodness ratings increased as age increased. Also, goodness ratings were higher for stimuli from children who had a steep slope in the robustness-of-contrast group as compared to children who had a shallow slope. It can be observed in Figure 4 that the interaction is due to the fact that the goodness ratings increase more sharply for the shallow-slope group than for the steep-slope group.

We also examined the data separately for goodness ratings of stimuli excised from real words only. The fricatives in the CV sequences excised from nonwords were, on average, 20 milliseconds longer for those excised from real words. Furthermore,

goodness ratings for stimuli excised from real words were, on average, 0.058 (5.8%) logit values greater than the ratings for stimuli excised from nonwords. Because of these differences between fricatives excised from real words and nonwords, we ran a third set of logistic regression analyses on goodness ratings only for stimuli excised from real words for stimuli from speakers who were 48 months or younger. For /ʃ/, 64% of the stimuli came from real words, while for /s/, 90% of the stimuli came from real words. Again, the dependent variable was the transformed logit values of the click locations from the listeners' goodness rating and the independent variables were robustness-of-contrast group, speaker-age in months, and the robustness-of-contrast group by age interaction. The results of these analyses are presented in Table 5 and plotted in Figure 6. As in the earlier two sets of analyses, there were no significant predictors of goodness ratings for /s/. For /ʃ/, age was a significant predictor and robustness-of-contrast group was marginally significant. Again, goodness ratings increased as age increased and goodness ratings were higher for stimuli from the steep-slope-group, relative to those from the shallow-slope-group. The robustness-of-contrast-group by age interaction was not significant. When the regression lines for Figures 4 and 5 are compared, it appears that the interaction observed in the analysis that included both real words and nonwords is probably due to goodness ratings increasing more sharply as age increases for nonwords relative to real words.

DISCUSSION

This study investigated the perceptual validity of a robustness of contrast in production measure proposed by Holliday et al (2010). This robustness of contrast measure aimed to quantify the difference between children with a larger or smaller acoustic contrast between two sounds. The results of this study showed that this measure is perceptually valid for the productions of /ʃ/ for children who are under 4 years of age. That is, there was a significant effect of robustness-of-contrast group on goodness ratings for /ʃ/ for stimuli from children who were less than 48 months. It should be noted, however, that neither robustness-of-contrast nor age predicted goodness ratings for /s/ and that robustness-of-contrast group was not a significant predictor of goodness ratings if stimuli from children over 48 months were included.

It is of interest that neither robustness-of-contrast group nor even speaker age were significant predictors of goodness ratings for /s/. Li and colleagues (Li et al., 2011) found that English listeners had a larger perceptual space for /s/ than for /ʃ/. Perhaps the results observed here are related to this finding – that is, perhaps the large acoustic-auditory space for /s/ for English listeners results in less sensitivity to within-category differences. Conversely, the smaller acoustic-auditory space for /ʃ/ for English listeners may result in greater sensitivity to within-category differences.

It was also observed that robustness-of-contrast was a significant predictor only when the goodness ratings were limited to those of children 48 months or younger. It

may be that by 48 months, children's productions have stabilized and are rated as equally good, regardless of their robustness-of-contrast group.

The reason that it was important to validate the robustness of contrast in production measure of Holliday et al. (2010) was not in order to recommend to clinicians that they acoustically analyze their clients' productions. Rather, these results suggest that goodness ratings, at least for /ʃ/ productions for children 48 months and younger, are a valid measure of children's acquisition of the /s/-/ʃ/ contrast. This is important, because this perceptual measure is relatively easy for clinicians to obtain.

There area number of limitations to the present study. Because this was a cross-sectional design and not a longitudinal study, we are not able to say that robustness-of-contrast slopes increase as children grow older. Additionally, this study only looked at the perceptual ratings of naïve listeners and not those of experienced speech-language pathologists. Previous research (Munson, Kaiser, & Urberg Carlson, 2008) found that experienced speech-language pathologists were more accurate raters than naïve listeners and that their ratings were more closely related to acoustic measurements. It is possible that if experienced speech-language pathologists had done the ratings, then there might have been significant effects of robustness-of-contrast group for /s/ as well as for /ʃ/ and for productions from older children as well as productions from younger children. A third limitation is that only correct productions were included. Holliday et al. (2010) included incorrect as well as correct productions, as long as the incorrect productions were fricative substitutions. It is possible that the inclusion of incorrect fricative substitutions as stimuli in this study might have resulted in significant effects for /s/ as well as for /ʃ/ and for productions from older children as well as productions from

younger children. This research is also limited in that only a single contrast (/s/ vs. /ʃ/) and a single acoustic measure (peak ERB) were examined. Future research can address these limitations.

Nevertheless, this study showed that goodness ratings for /ʃ/ are a valid measure of how well children produce this sound. These goodness ratings were correlated with the acoustic robustness of contrast measure in production proposed by Holliday and colleagues (2010) for children under age 4. If clinicians use goodness ratings by naïve listeners to evaluate productions of /ʃ/ for children under 4, they can feel confident that this is a valid measure of how well children can produce this sound.

REFERENCES

Arbisi-Kelm, T., Beckman, M.E., Kong, E.J., Edwards, J. "Psychoacoustic measures of stop production in Cantonese, Greek, English, Japanese, and Korean." Poster 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 of /s/ in children. *Journal of Phonetics, 18*, 51-63.

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 consonant acquisition. *Clinical Linguistics &Phonetics, 22*(12), 937-956.

Forrest, K. & Rockman, B.K. (1988). Acoustic and perceptual analysis of word-initial stop consonants in phonologically disordered children. *Journal of Speech and Hearing Research, 31*, 449-459.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R.N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84*(1), 115-123.

Fudala, J., (2003). *Arizona Articulation Proficiency Scale, Third Revision (Arizona-3).* Los Angeles, CA: Webster Psychological Services.

Gibbon, F. & Scobbie, J. (1997). Covert contrasts in children with phonological disorder. *The Australian Communication Quarterly (Autumn)*, 13-16.

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

Hay, J., Nolan, A., & Drager, K. (2006a). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review, 23*, 351-379.

Hay, J., Warren, P., Drager, K. (2006b). Factors influencing speech perception in the context of a merger-in-progress. *Jounral of Phonetics, 34*, 458-484.

Holliday, J.J., Beckman, M.E., & Mays, C. (2010). Did you say susi or sushi? Measuring the emergence of robust fricative contrasts in English- and Japanese-acquiring children. *Proceedings of INTERSPEECH 2010.* 1886-1889.

Johnson, K., Strand, E.A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 37*, 359-384.

Johnson, Julie M. (2010). *The role of clinical experience in listening for covert contrasts in children's speech*. M.A. thesis. Department of Speech-Language-Hearing Sciences, University of Minnesota.

Kent, R.D. (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.

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.

Li, F., Munson, B., Edwards, J., Yoneyama, K., & Hall, K. (2011). Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development. *Journal of the Acoustical Society of America*, *129*(2), 999-1011.

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

Macken, M.A. & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice onset time in word-initial stop consonants. *Journal of Child Language, 7*, 41-74.

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.

Metz, D.E., Schiavetti, N., & Sacco, P.R. (1990). Acoustic and psychophysical dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Jounral of Speech and Hearing Disorders, 55*, 516-525.

Moore, B.C.J. & Glasberg, B.R. (1987). Formulae describing frequency selectivity as a function of frequency and level and their use in excitation patterns. *Hearing Research, 28*, 209-225.

Morris, H.L., Spriestersbach, D.C., & Darley, F.L. (1961). An articulation test for assessing competency of velopharyngeal closure. *Journal of Speech and Hearing Research, 4*, 48-55.

Munson, B., Edwards, J., Schellinger, S., Beckman, M.E., & Meyer, M.K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. *Clinical Linguistics & Phonetics, 24*(4-5), 245-260.

Munson, B., Kaiser, E., & Urberg Carlson, K. (2008). Assessment of phonetic skills in children 3: Fidelity of responses under different levels of task delay. Paper presented at the 2008 ASHA Convention, Chicago, 20-22 November 2008.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*, 62-85.

Nissen, S.L. & Fox, R.A. (2005). Acoustic and spectral characteristics of young children's fricative productions: A development perspective. *Journal of Acoustical Society of America, 118*(4), 2570-2578.

Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker contingent process. *Psychological Science, 5*, 42-46.

Pye, C., Wilcox, K.A., & Siren, K.A. (1987). Refining transcriptions: The significance of transcriber 'errors.' *Journal of Child Language, 15*, 17-37.

Samar, V. J. & Metz, D.E. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing-imapired population. *Journal of Speech and Hearing Research, 31*, 307-316.

Schiavetti, N., Metz, D.E., Sitler, R.W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research, 24*, 441-445.

Scobbie, J.M., Gibbon, F., Hardcastle, W.J., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M.B. Broe & J.B. Pierrehumbert (Eds.) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, 194-207. Cambridge: Cambridge University Press.

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.

Toner, M.A. & Emanuel, F.W. (1989). Direct magnitude estimation and equal appearing interval scaling of vowel roughness. *Journal of Speech and Hearing Research*, *32*, 78-82.

Tyler, A.A., Figurski, G.R., Langsdale, 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 phonetic skills in children 1: Continuous measures of children's speech production: Visual analog scale and equal appearing interval scale measures of fricative goodness. Paper presented at the 2008 ASHA Convention, Chicago, 20-22 November 2008.

Urberg Carlson, K., Kaiser, E., & Munson, B. (2008). Assessment of phonetics skills in children 2: Testing gradient measures of children's productions. Paper presented at the 2008 ASHA Convention, Chicago, 20-22 November 2008.

Weist, R.M. & Kruppe, B. (1977). Parent and sibling comprehension of children's speech. *Journal of Psycholinguistic Research, 6*(1), 49-58.

Table 1a: Speaker identification number, age in months, gender, number of tokens and slope values for subjects with a "steep" slope.

SpeakerID	Age (months)	Sex	Number of /s/ tokens	Number of /ʃ/ tokens	Total number of tokens	Slope Value
203	26	F	1	6	7	-0.377

210	28	M	6	6	12	-0.18
219	32	M	6	6	12	-2.13
221	33	F	6	6	12	-0.29
222	34	M	6	6	12	-1.13
311	39	M	6	6	12	-19.04
317	45	M	6	6	12	-0.23
318	42	F	6	6	12	-1.032
323	45	F	6	6	12	-0.77
324	46	F	6	6	12	-0.968
402	48	M	6	6	12	-2.25
407	51	F	6	6	12	-1.048
420	58	F	6	6	12	-12.13
425	59	M	6	6	12	-1.15
505	62	M	6	6	12	-1.705
506	62	F	6	6	12	-15.28
516	66	F	6	6	12	-15.49
522	69	M	6	6	12	-1.52

Table 1b: Speaker identification number, age, gender, number of tokens and slope values for subjects with a "shallow" slope.

SpeakerID	Age (months)	Sex	Number of /s/ tokens	Number of /ʃ/ tokens	Total number of tokens	Slope Value
204	25	M	3	1	4	-0.015
218	33	M	6	2	8	-0.11
220	33	F	3	6	9	-0.075
224	35	F	6	2	8	-0.10
301	40	F	4	6	10	-0.005
309	39	M	6	6	12	-0.05
316	42	F	1	5	6	-0.01
321	43	M	6	6	12	-0.11
403	39	F	6	6	12	-0.22
406	51	M	6	6	12	-0.90
411	53	M	6	6	12	-0.017
413	54	F	6	6	12	-0.06
501	60	F	6	6	12	-0.031
508	62	F	6	6	12	-0.509
519	68	M	6	6	12	-0.48
521	69	M	6	6	12	-0.09

Table 2a. Results of the logistic regression analysis for all stimuli (speaker age range = 25 months to 69 months) for goodness ratings for /ʃ/.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**

Intercept	**- 0.856481**	**0.183578**	**- 4.665**	**< 0.001**
Robustness-of-contrast-group	0.62722	0.252442	2.485	< 0.1
Age in months	**0.025474**	**0.007293**	**3.493**	**< 0.001**
Robustness by age	- 0.0122	0.009962	-1.225	0.23017

Table 2b. Results of the logistic regression analysis for all stimuli (speaker age range = 25 months to 69 months) for goodness ratings for /s/.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**
Intercept	- 0.512554	0.194599	- 2.634	< 0.05
Robustness-of-contrast-group	0.450730	0.267598	1.684	0.1025
Age in months	0.1025	0.007731	1.871	< 0.1
Robustness by age	- 0.003741	0.010560	- 0.354	0.7256

Table 3a. Results of the logistic regression analysis for /ʃ/ for stimuli from children 48 months and younger for both real words and nonwords.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**
Intercept	**- 1.39107**	**0.15327**	**- 9.076**	**< 0.001**
Robustness-of-contrast-group	0.74555	0.19323	3.858	< 0.01
Age in months	**0.07868**	**0.01204**	**6.534**	**< 0.001**
Robustness by age	- 0.02898	0.01439	- 2.015	< 0.1

Table 3b. Results of the logistic regression analysis for /s/ for stimuli from children 48 months and younger for both real words and nonwords.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**
Intercept	- 0.479240	0.304663	- 1.573	0.135
Robustness-of-contrast-group	0.112057	0.384077	0.292	0.774
Age in months	0.004573	0.023937	0.191	0.851
Robustness by age	0.851	0.028594	1.072	0.299

Table 4a. Results of the logistic regression analysis for /ʃ/ for stimuli from children 48 months and younger for real words only.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**
Intercept	**- 1.56536**	**0.25749**	**- 6.079**	**< 0.001**
Robustness-of-contrast-group	0.92728	0.28402	3.265	< 0.01
Age in months	**0.09066**	**0.01907**	**4.753**	**< 0.001**
Robustness by age	-0.04234	0.02069	-2.047	< 0.1

Table 4b. Results of the logistic regression analysis for /s/ for stimuli from children 48 months and younger for real words only.

Independent Variable	**Estimate**	**Std. Error**	**t value**	**p value**
Intercept	-0.241023	0.327163	-0.737	0.472
Robustness-of-contrast-group	-0.099562	0.412443	-0.241	0.812
Age in months	-0.006969	0.025704	-0.271	0.790
Robustness by age	0.043200	0.030706	1.407	0.179

Figure Captions

Figure 1: Results from individual stepwise logistic regression models from Holliday et al. (2010) for 2-, 3-, 4-, and 5-year olds used in the current study.

Figure 2a: Goodness rating scale for /s/ that the participants were presented with during the experiment.

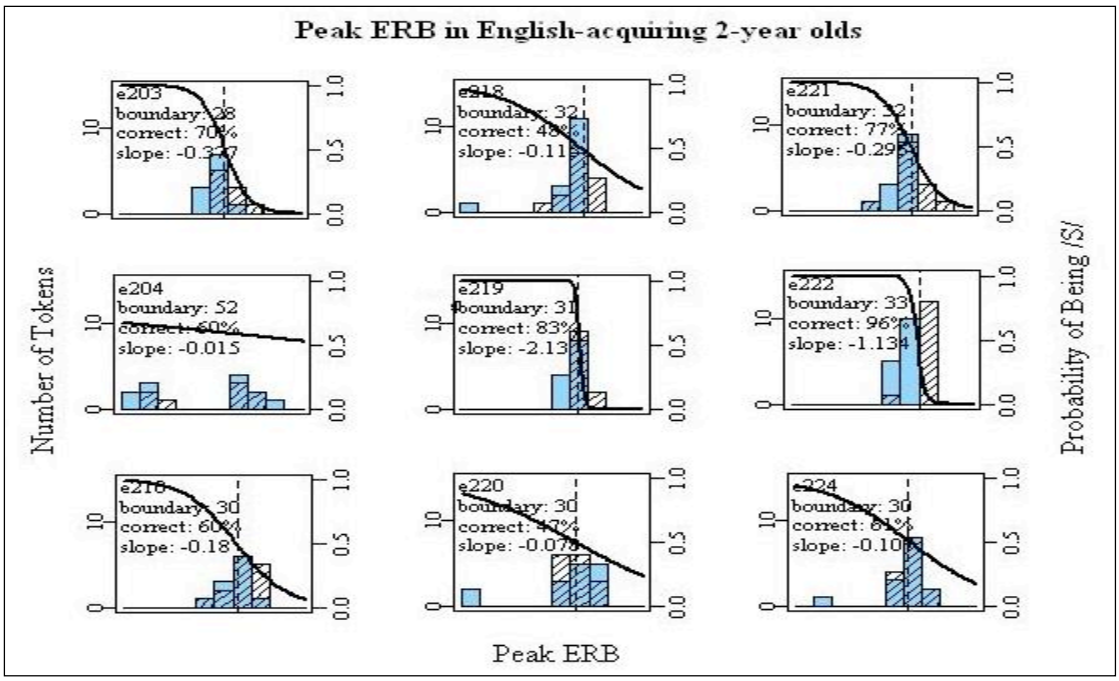
Figure 2b: Goodness rating scale for /ʃ/ that the participants were presented with during the experiment.

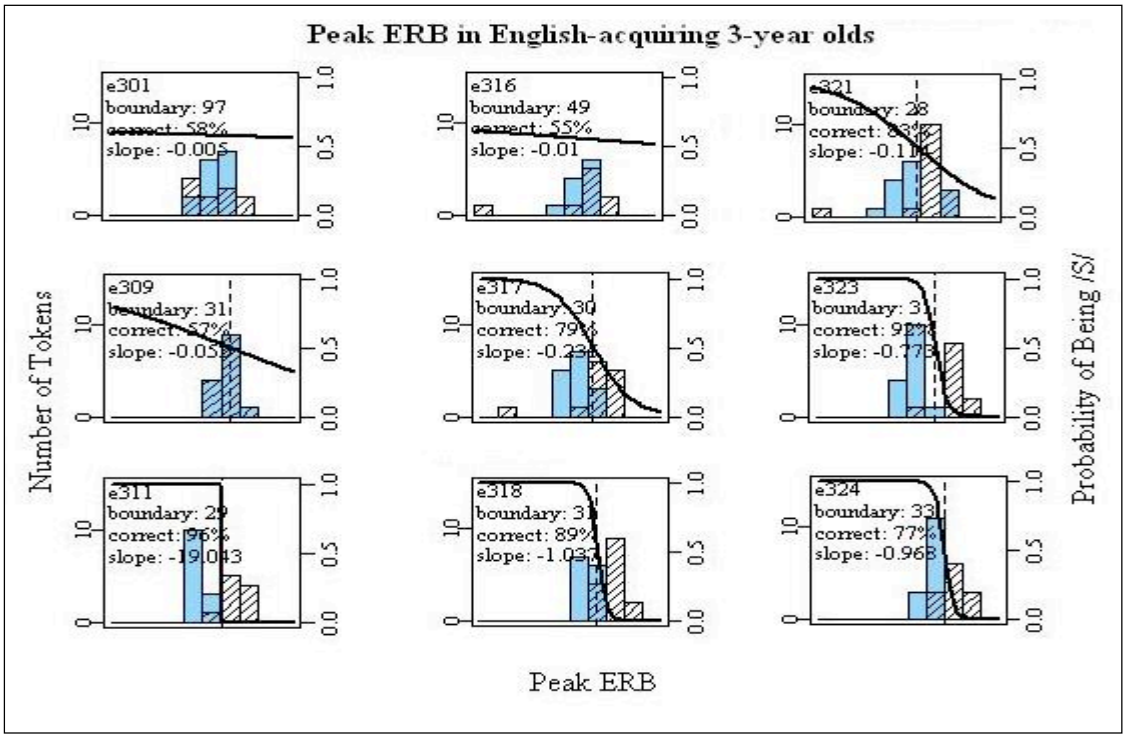
Figure 3: Goodness ratings plotted against age for both robustness-of-contrast groups for /ʃ/ (left plot) and /s/ (right plot) for all stimuli. Regression lines are plotted regardless of significance in the logistic regression analyses. Solid lines are for the shallow-slope group and dotted lines are for the steep-slope group.

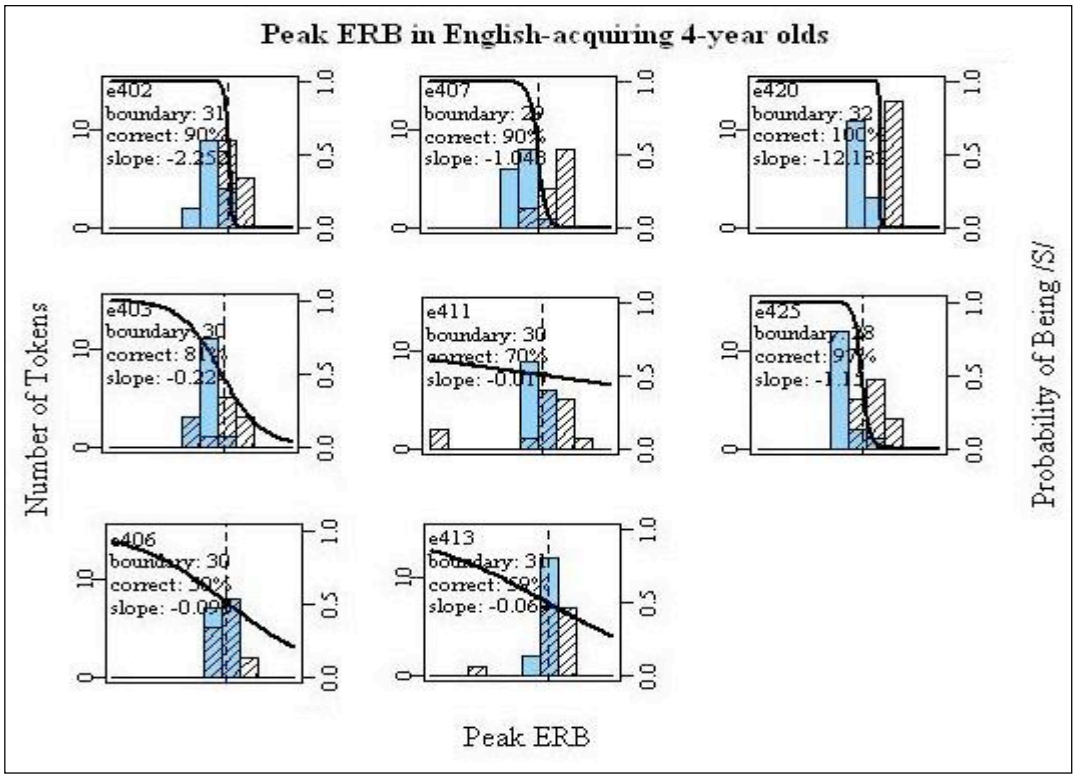
Figure 4: Goodness ratings plotted against age for both robustness-of-contrast groups for /ʃ/ (left plot) and /s/ (right plot) for stimuli from children 48 months and younger. Regression lines are plotted regardless of significance in the logistic regression analyses. Solid lines are for the shallow-slope group and dotted lines are for the steep-slope group.

Figure 5: Goodness ratings plotted against age for both robustness-of-contrast groups for /ʃ/ (left plot) and /s/ (right plot) for stimuli from children 48 months and younger for real words only. Regression lines are plotted regardless of significance in the logistic regression analyses. Solid lines are for the shallow-slope group and dotted lines are for the steep-slope group

Figure 1:


Peak ERB in English-acquiring 2-year olds
e203
boundary: 28
correct: 70%
slope: -0.317
e218
boundary: 32
correct: 48%
slope: -0.11
e221
boundary: 32
correct: 77%
slope: -0.29
e204
boundary: 52
correct: 60%
slope: -0.015
e219
boundary: 31
correct: 83%
slope: -2.13
e222
boundary: 33
correct: 96%
slope: -1.134
e216
boundary: 30
correct: 60%
slope: -0.18
e220
boundary: 30
correct: 47%
slope: -0.078
e224
boundary: 30
correct: 61%
slope: -0.10
Number of Tokens
Probability of Being /S/
Peak ERB


Peak ERB in English-acquiring 3-year olds
e301
boundary: 97
correct: 58%
slope: -0.005
e316
boundary: 49
correct: 55%
slope: -0.01
e321
boundary: 28
correct: 83%
slope: -0.11
e309
boundary: 31
correct: 57%
slope: -0.05
e317
boundary: 30
correct: 79%
slope: -0.23
e323
boundary: 31
correct: 92%
slope: -0.773
e311
boundary: 29
correct: 96%
slope: -19.043
e318
boundary: 31
correct: 89%
slope: -1.037
e324
boundary: 33
correct: 77%
slope: -0.963
Number of Tokens
Probability of Being /S/
Peak ERB


Peak ERB in English-acquiring 4-year olds
e402
boundary: 31
correct: 90%
slope: -2.25
e407
boundary: 29
correct: 90%
slope: -1.048
e420
boundary: 32
correct: 100%
slope: -12.13
e403
boundary: 30
correct: 81%
slope: -0.22
e411
boundary: 30
correct: 70%
slope: -0.01
e425
boundary: 28
correct: 97%
slope: -1.1
e406
boundary: 30
correct: 90%
slope: -0.09
e413
boundary: 31
correct: 59%
slope: -0.06
Number of Tokens
Probability of Being /S/
Peak ERB

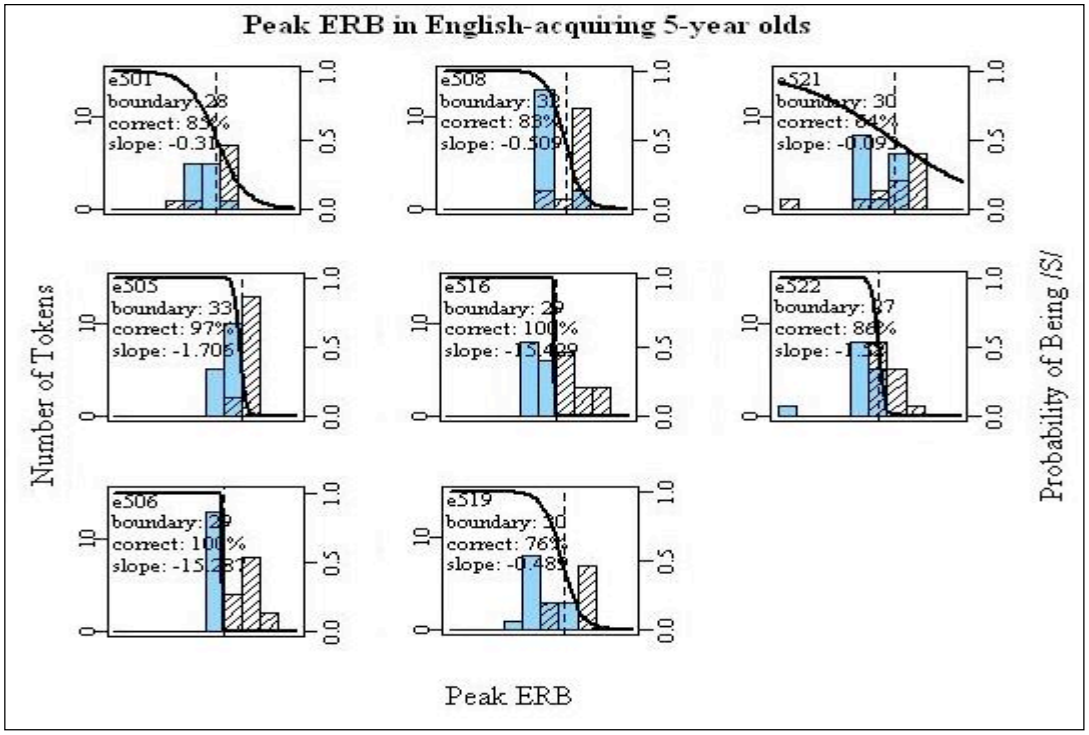

Peak ERB in English-acquiring 5-year olds
e501
boundary: 28
correct: 83%
slope: -0.31
e508
boundary: 32
correct: 83%
slope: -0.509
e521
boundary: 30
correct: 84%
slope: -0.095
e505
boundary: 33
correct: 97%
slope: -1.706
e516
boundary: 29
correct: 100%
slope: -15.499
e522
boundary: 27
correct: 86%
slope: -1.31
e506
boundary: 29
correct: 100%
slope: -15.231
e519
boundary: 30
correct: 76%
slope: -0.489
Number of Tokens
Probability of Being /S/
Peak ERB

Figure 2a:

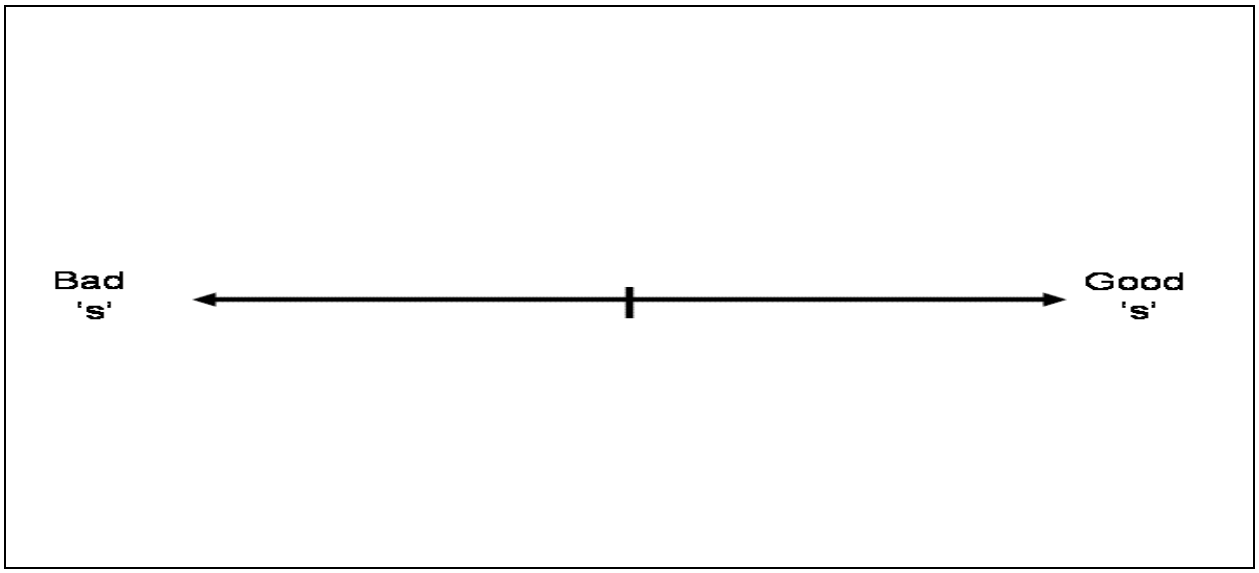


Figure 2b:

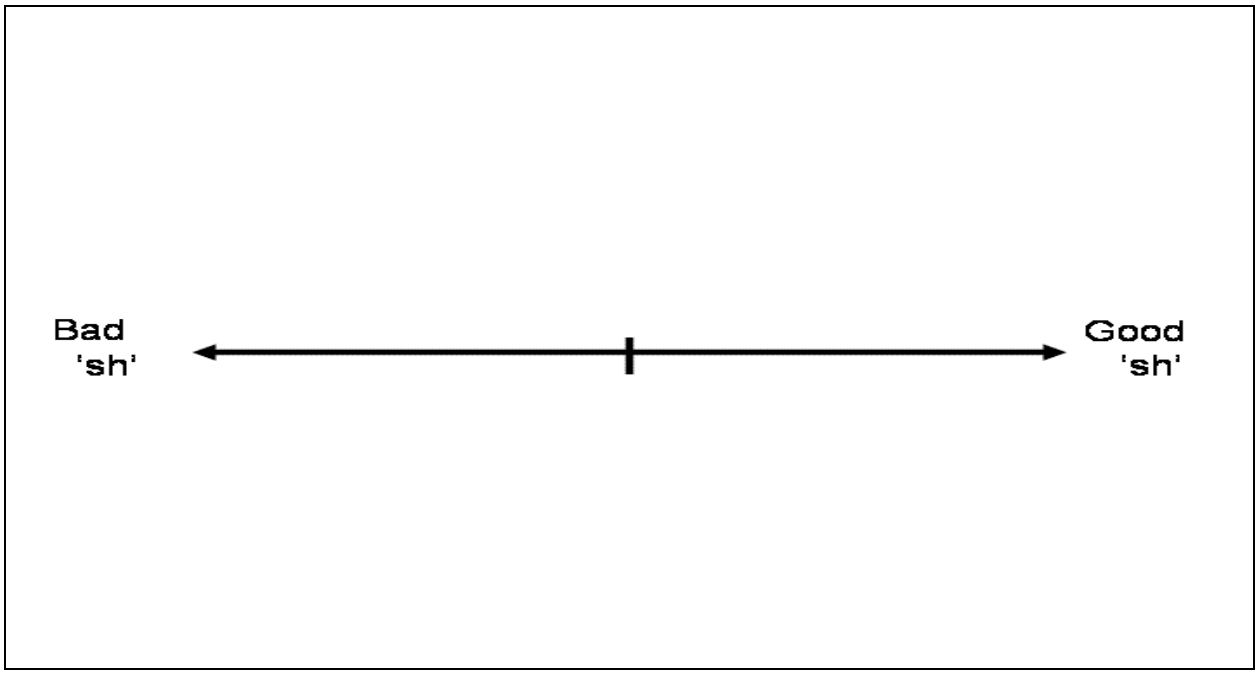


Figure 3:

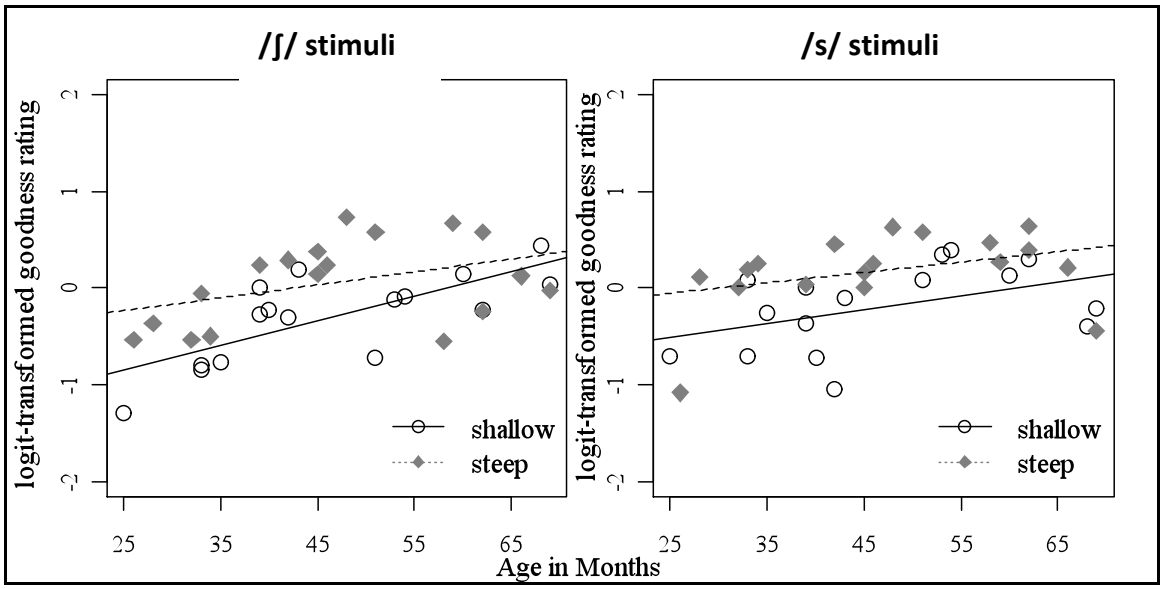


Figure 4:

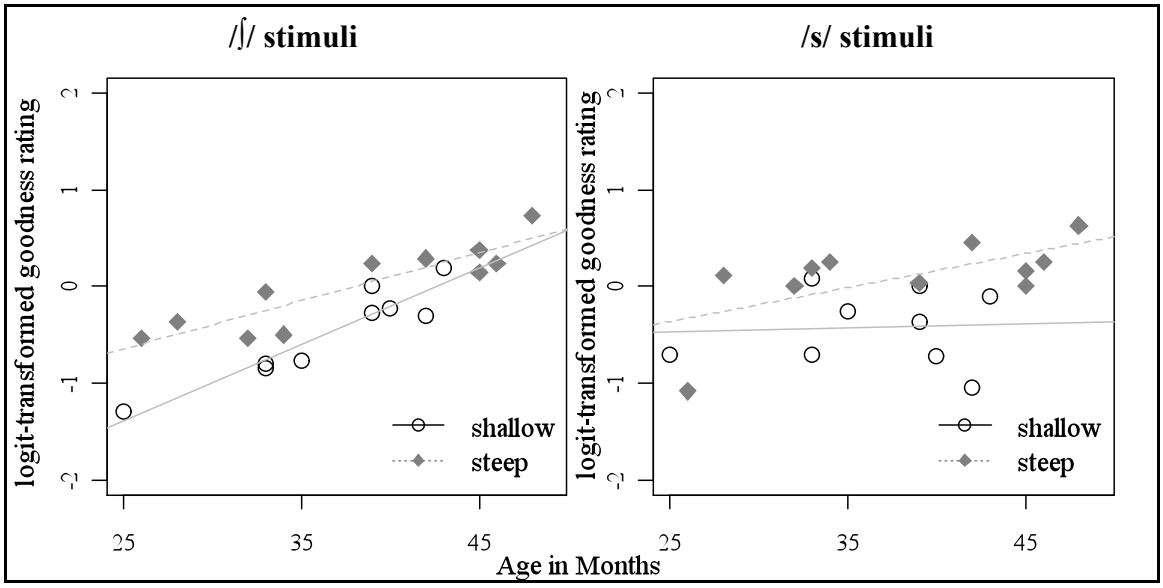
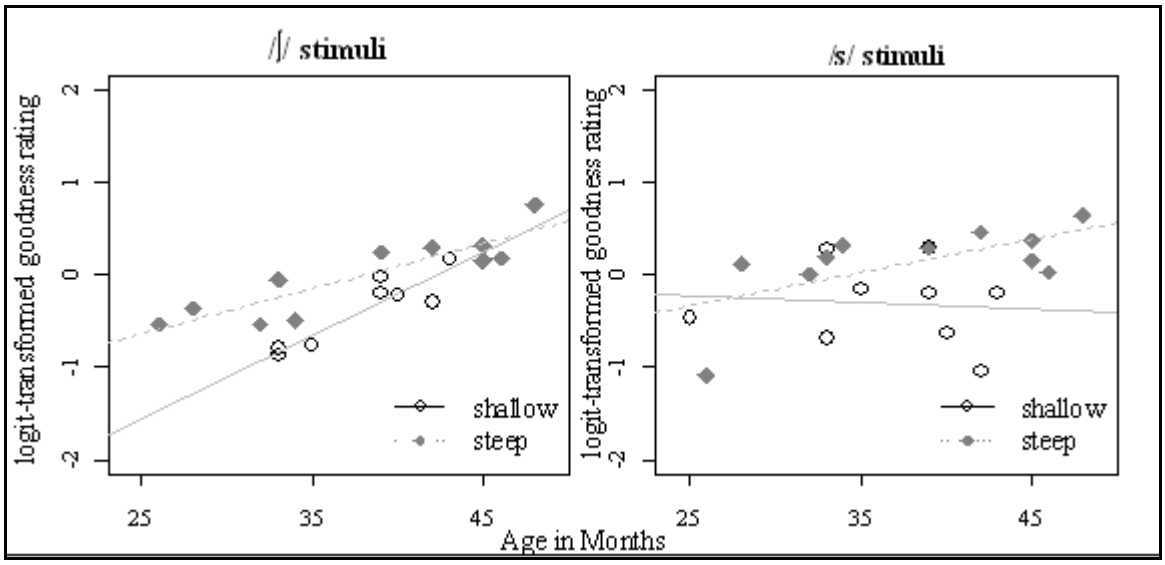


Figure 5:


/ʃ/ stimuli
/s/ stimuli
logit-transformed goodness rating
shallow
steep
25
35
45
Age in Months