In Press in *Speech, Language, and Hearing* (http://www.maneyonline.com/loi/slh) 10/27/2015

An Exploration of Methods for Rating Children's Productions of Sibilant Fricatives

Benjamin Munson

Kari Urberg Carlson

Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis

Correspondence regarding this article should be sent to

Benjamin Munson
Department of Speech-Language-Hearing Sciences
University of Minnesota
115 Shevlin Hall
164 Pillsbury Dr.
Minneapolis, MN 55455
munso005@umn.edu
+1 612 624 0304
Fax: +1 612 624 7586.

Abstract

This paper examines three methods for providing ratings of within-category detail in children's productions of /s/ and /ʃ/. A group of listeners (n=61) participated in a rating task in which a forced-choice phoneme identification task was followed by one of three measures of phoneme goodness: visual analog scaling, direct magnitude estimation, or a Likert scale judgment. All three types of ratings were similarly correlated with sounds' acoustic characteristics. Visual analog scaling and Likert scale judgments had higher intra-rater reliability than did direct magnitude estimation. Moreover, both of them elicited a wider range of judgments than did direct magnitude estimation. Based on our evaluation, Likert scale judgments and visual analog scaling are equally useful tasks for eliciting within-category judgments. Of these two, visual analog scaling may be preferable because it allows for more distinct levels of response.

Studies of speech-sound development using phonetic transcription often appear to show that speech sounds are acquired discontinuously. For example, Berg's (1995) longitudinal study of the acquisition of velar obstruents in a single child shows a very sharp increase in the accuracy of her production of /k/ and /g/, coinciding with a sharp decrease in her production of 'fronting' errors.

In contrast, many acoustic and articulatory studies suggest that children's speech sound acquisition involves the *gradual* and *continuous* acquisition of phonetic contrasts. This continuous acquisition is illustrated well by Li's (2012) cross-sectional study of children acquiring /s/ and /ʃ/ in English and Japanese. Li found that two-year-old children’s productions of target /s/ and /ʃ/ did not differ in the acoustic parameters that differentiate adults’ /s/ from /ʃ/ in either language, centroid frequency of the fricative, spectral dispersion, and the F2 frequency of the following vowel at its onset. Across the two- to five-year-old age range, children’s productions of /s/ and /ʃ/ became gradually more different from one another in the relevant, language-specific acoustic dimensions (centroid for English, all three parameters for Japanese). The relationship between age and acoustic differentiation was continuous and linear. One interpretation of these acoustic results is that they represent children’s gradual articulatory development. Children begin development with articulations for /s/ and /ʃ/ that do not differentiate between the adult targets. As development progresses, children refine their articulations so that these targets become gradually more different from one another. This, in turn, leads to greater acoustic differentiation. This interpretation is consistent with studies of children’s articulation, such as Gibbon’s (1999) finding that children with speech sound errors have tongue-palate contact patterns that are merged between two target productions. The process of differentiation can have a very protracted time course. Romeo, Hazan, and Pettiano (2013) recently showed that the acoustic differentiation between target /s/ and /ʃ/ in children as old as 13 years was not yet adult-like.

Further evidence for the gradual acquisition of speech sounds comes from studies of the development of stop consonant voicing. Macken and Barton (1980) found that children go through three stages as they learn to differentiate the voice onset times (VOT) of English voiced and voiceless initial stops in production. In the first stage, children make no VOT distinction between voiced and voiceless stops. In the second stage, they produce voiced and voiceless stops with reliably different VOTs, but the mean VOT for voiceless stops falls within the range of adult voiced stops. In the third stage, children make an adult-like contrast. This second stage, when the child makes a contrast between two sounds, even though both sounds fall within the perceptual category of a single adult phoneme, is sometimes referred to as a *covert contrast*.

By definition, the gradual development noted by Li and Romeo et al., and the covert contrasts noted by Macken and Barton cannot be captured by phonetic transcription. In both of these cases, there is acoustic differentiation between sounds that are transcribed with the same symbol. Both findings suggest that phonetic transcription must be supplemented by other measures. Currently, in most clinical assessments and experimental research protocols, adults listen to a child and transcribe their productions using a phonetic symbol and an optional series of diacritic markings. These transcriptions cannot document children's gradual attainment of speech sounds or their production of covert contrasts.

As argued by Gibbon (1999), Kent (1996), Ladd (2014), Munson, Schellinger, Edwards, Meyer, and Beckman (2010), and others, the widespread use of phonetic transcription skews our understanding of speech-sound development and learning. In clinical settings, this skewed view may have negative consequences for the assessment of progress and for treatment planning.

First, a child who is making steady progress toward the adult-like production of speech sounds could be assessed as making no progress if this steady change were within a single transcription category. This could lead to a clinician abandoning a successful intervention program. Second, as illustrated by Munson, Schellinger, and Urberg Carlson (2012, Figures 1 and 2), a clinician using phonetic transcription would be unable to distinguish between two very different underlying patterns of errors and phonetic differentiation, one in which a child was making a merger between two sounds like /s/ and /ʃ/ (in the sense of Gibbon, 1999 and Li, 2012) and one where the child was making a true substitution of /s/ for /ʃ/. This would make it more difficult to tailor treatment to the child's specific error pattern.

One potential solution to this problem is to use instrumental techniques like acoustic analysis. Acoustic analysis has proven to be a useful laboratory technique for measuring phonetic detail within categories. While acoustic studies of covert contrast are extremely useful in determining the extent to which children can produce differences between sounds that are perceived to be the same, their utility in assessments of speech development is limited. The time constraints (Williams, 2002) and unfavorable recording environments (Nelson & Soli, 2000) in which many clinical assessments of speech and language take place limit the feasibility of acoustic analysis.

The focus of this report is to examine finer-grained *perceptual* methods of rating children's productions. Such methods would offer clinicians a way to document progress in children who have not yet mastered a phoneme being targeted in speech-language therapy, and would give researchers the ability to examine children's production in naturalistic settings in greater detail than is allowed by phonetic transcription. Such a measure would ideally give information that is as fine grained and continuous as that of an acoustic analysis. An ideal perceptual measure of children's production would meet at least four criteria. First, it would elicit a continuous response for a continuously varying signal. Second, it would capture meaningful phonetic differences between sounds that would be transcribed identically. Third, it would have high intra-rater reliability. Finally, it would correlate well with relevant acoustic characteristics of the sounds being rated. This final criterion is particularly important because it indicates the predictive validity of the ratings. Such a measure would not replace phonetic transcription, but would supplement it in cases where there is reason to believe that a child is producing systematic differences within a transcribed category.

This study examines three tasks for eliciting judgments of phonetic variation in children's speech. In particular, we focus on three of the criteria in the previous paragraph: the extent to which the ratings are continuous, the strength of their correlation with the acoustic characteristics of the stimuli being rated, and their intra-rater reliability. The focus of this study is on judgments of children's productions of /s/ and /ʃ/. This contrast is particularly suitable for this type of investigation. There is acoustic evidence that the acquisition of /s/ and /ʃ/ is continuous, as shown by Li (2012), Holliday, Reidy, Beckman, and Edwards (2015), and Nicholson, Munson, Reidy, and Edwards (2015). This makes these sounds particularly appropriate for a study of continuous rating scales. The contrast between these two sounds is acquired relatively late. Smit, Hand, Freilinger, Bernthal, and Bird (1990) report that it is not until children are aged 4;6 (years;months)t that trained judges rate 75% or more of their productions of /ʃ/ to be acceptable. Moreover, Smit (1993) reports that [s] productions are among the most common errors when children attempt to produce /ʃ/. For these reasons laboratory studies of children's productions of these sounds have substantial external validity, as speech-language clinicians and phonological development researchers are likely to encounter them frequently. Further, the articulatory and acoustic characteristics of these sounds are well understood. There already exist a well-established set of candidate acoustic measures to compare to the perceptual ratings, such as those described in Jongman, Wayland, and Wong (2000). The sounds /s/ and /ʃ/ have a high functional load in English: in addition to occurring in many early-acquired words, /s/ is an allomorph of the English plural and third-person singular verb form. Finally, there exists a public-access corpus of preschool children's productions of these sounds, collected from the παιδολογος (*paidologos*, from the Greek παιδο [*paido*, child] + λόγος [*logos*, speech]) project, a large-scale study of consonant acquisition across languages (Edwards & Beckman, 2008). This corpus was used to develop the stimuli in this study, which have been used in a previous cross-linguistic study of the perception of children’s fricatives (Li, Munson, Edwards, Yoneyama, & Hall, 2011) and in previous studies of fricative development by Li (2012), Holliday et al. (2015), and Nicholson et al. (2015).

This paper examines three perceptual rating tasks, each of which examines the extent to which a particular token is a prototypical example of the target sound. All three tasks requires participants to identify each token as /s/ or /ʃ/, then make a second judgment of how good an example of /s/ or /ʃ/ it is. The first method is forced choice judgments followed by Likert-scale judgments of category goodness. Likert scales require judges to select among a finite number of values on an equally appearing interval (EAI) scale, which may have labels attached to them. The second method is forced-choice judgments followed by visual analog scale (VAS) judgments of category goodness. In the VAS task in this study, listeners are presented with a line and are asked to make a mark on the line at the location that best represents where each stimulus falls with respect to category goodness. VAS and Likert scales are similar in that they are measures of category goodness. The main difference is that Likert scales are more granular, and that VAS lacks labels for non-endpoint values. The number of scale values in a VAS is limited by the measurement precision of the instrument. VAS measurements have the potential for better correlation with independent variables because there are more available values. For the same reason, they also have the potential for poorer reliability. In this paper, we refer to these tasks as *Categorization plus Likert Goodness Rating* and *Categorization plus Continuous Goodness Rating*, respectively.

The third method is forced-choice judgments followed by direct magnitude estimates of category goodness. Direct magnitude estimation (DME) requires judges to assign a number to a stimulus based on their judgment of how much of a given characteristic is present in that stimulus. In the task in this report, listeners judged the extent to which each token exemplified the attributes of a good /s/ or /ʃ/. Listeners judged each sample with respect to the first sample that is presented. This is called a modulus-free DME task, and is argued to have superior reliability and less bias than tasks in which the first sample is the same across listeners (Engen, 1971). Judges are asked to assign their own choice of number to the first sample, which then became the reference for the rest of the samples. Because each judge hears a different sample first, each judge is using a different scale, so the responses need to be scaled before analysis so that the responses of different judges can be directly compared. In this paper we refer to this task as *Categorization plus Magnitude Estimation*.

Relatively little work has used scales like these to examine people's perception of phonetic detail within categories. Massaro and Cohen (1983) examined the perception of a vowel contrast, a stop place contrast, and a stop voicing contrast using VAS. The goal of that study was to assess whether the observed distribution of responses was better predicted by a categorical model of speech perception, or by one in which listeners access continuous acoustic variation during perception. The results of that study supported the latter model. Miller (1994) reviews studies showing that individuals can provide Likert scale goodness ratings for within-category differences in stop consonant voicing. Munson et al. (2010) presented listeners with children's productions of CV sequences beginning with either target /s/ or target /θ/ and asked them to rate along a VAS, specifically, a double-headed arrow anchored by the text "the 's' sound" at one end and the "the 'th' sound" at the other. They found that listeners' ratings differentiated between pairs of sounds that had been transcribed identically, such that instances of [s] for target /s/ was rated closer to the end of the line anchored by the text "the 's' sound" than were instances of [s] for target /θ/. Miller, Massaro and Cohen, and Munson et al.'s results are consistent with a variety of other behavioral and neurophysiological investigations showing that listeners can access within-category phonetic detail during speech perception (Carney, Widin, & Viemeister, 1977; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Toscano, McMurray, Dennhardt, & Luck, 2010). These results further motivate the use of continuous rating scales.

This is an exploratory study of the utility of different tasks for assessing within-category phonetic detail in children's speech. Our goal is to determine which of these tasks is most useful for experimental and observational studies of speech sound development. The specific tasks that are evaluated were chosen either because they have been used in a small number of studies previously (Categorization plus Likert Goodness Rating), or because they are used widely in studies rating other speech variables like voice quality and nasality (i.e., Baylis, Munson, & Moller, 2011; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009). Despite the exploratory character of this study, we can make three a priori predictions. The first is that the differences among the three tasks should be in the nature of the ratings of category goodness and not in the categorizations themselves. That is, the proportion of stimuli that are judged to be /s/ or /ʃ/ should not differ as a function of whether the subsequent goodness rating is a Likert-scale judgment, a VAS rating, or a direct magnitude estimate of category goodness. Second, we predict that the cognitive demands of the direct magnitude estimation task will lead to it having lower intra-rater reliability than the other two measures. Third, we predict that there will be a lower correlation between ratings and the acoustic characteristics of the stimuli being rated for the Likert-scale judgments than for the VAS and direct magnitude estimates of category goodness, owing to the restricted range in the EAI scores.

Methods

Participants

Participants in this study were recruited from the University of Minnesota community via printed fliers and a recruitment database. They were between the ages of 18 and 50 years, were native speakers of a North American dialect of English, and self-reported no past or current speech, language, or hearing impairments. They were compensated $10 for their participation. A different group of listeners participated in each of the three tasks described below. There were 19 listeners in the Categorization plus Likert Goodness Rating task, and 21 each in the Categorization plus Continuous Goodness Rating and Categorization plus Magnitude Estimation tasks.

Stimuli

The same 400 stimuli were used for all three tasks. The stimuli were produced by monolingual native speakers of a Midwestern dialect of American English, or of Japanese. Further details on the talkers can be found in Li, Edwards, and Beckman (2009). The stimuli of greatest interest were 320 fricative-vowel sequences that were produced by nine English-acquiring two-year-old children, ten Japanese-acquiring two-year-old children, 13 English-acquiring three-year-old children and eight Japanese-acquiring three-year-old children. The remaining productions were accurate productions of /s/ and /ʃ/ by adults, as judged by native-speaker transcribers. These were included as perceptual anchors in this experiment. Ratings of these stimuli are not included in the analysis, as the focus of this paper was to assess people's perception of phonetic detail in children's productions. The inclusion of both Japanese- and English-speakers' stimuli was intended to increase the range of phonetic variation present in the stimuli, as well as to accommodate the design of a parallel study that explicitly compared the judgments of English- and Japanese-acquiring children's speech (Li et al., 2011). The stimuli were taken from the παιδολογος database. The stimuli consisted of CV syllables that were excised from a variety of single words produced in a word-repetition task in which the child produced a target word after seeing a picture of it and hearing an auditory model. The words were chosen so that they would be familiar to young children and would elicit consonants in a variety of vowel contexts that were comparable across the languages being examined. CVs were used instead of real words to remove all lexical support, which was critical given that some of the stimuli were excised from Japanese words. All of the fricatives were produced in word-initial position. The target words included the phonemes /s/ or /ʃ/ followed by a monophthongal vowel. As described in Li et al. (2011), the stimuli included fricatives that had been transcribed by native-speaker phoneticians to be correct /s/, correct /ʃ/, [s] for /ʃ/ substitutions, and /ʃ/ for [s] substitutions. Because the canonical error patterns differ for English and for Japanese, there were more [s] for /ʃ/ productions in English-acquiring children, and more [ʃ] for /s/ substitutions in Japanese-acquiring children. Within each transcription category, the stimuli were as balanced as possible for vowel context and speaker gender. The recordings were made in a quiet location. All of the tokens used as stimuli were judged to be free from extraneous noise.

A 40 ms selection was taken from the middle of each fricative, and measurements were made of the first four spectral moments: the centroid frequency (M1), the variance (M2), the skewness (M3), the kurtosis (M4), and the second-formant frequency of the following vowel at onset (onsetF2). These measures were chosen for three reasons. First, previous work has shown that the first four spectral moments discriminate between target /s/ and target /ʃ/ in adults' speech very effectively (Jongman, Wayland, & Wong, 2000). Second, Li, Edwards, and Beckman (2009) used these measures to characterize /s/-/ʃ/ contrasts, mergers, and covert contrasts in the same set of children who produced the stimuli in this study, including the specific tokens used as stimuli in this study. Li et al. argued that these measures are needed to fully characterize the /s/-/ʃ/ distinction, as well as common misarticulations of these sounds. Both centroid (M1) and skewness (M3) should be correlates of the location of the front edge of the constriction, provided the degree of constriction doesn't differ between /s/ and /ʃ/. The more /s/-like sounds should have higher M1 and M3 values. The onset F2 value should be a correlate of the back edge of the constriction, so that more /ʃ/-like sounds should have higher F2 values. In Li et al.'s work, this measure was particularly useful in characterizing the broad constriction for Japanese /ʃ/ (sometimes narrowly transcribed as /ɕ/) versus the relatively narrower constriction for English /ʃ/. Higher values for the variance (M2) can indicate a more dental, less sibilant production of [s], so we expect this to be higher in a frontal (i.e., more /θ/-like) production of target /s/ (Jongman et al., 2000). Higher values for kurtosis (M4) might indicate more than one peak in the spectrum. This might suggest the existence of multiple resonant cavities that are present when one of the constrictions in the fricative is incomplete and leaks, as in the side cavity leakage of lateral misarticulations of English /s/. Finally, these measures were chosen because they were found by Li et al. (2011) to predict listeners' categorical judgments of this same set of stimuli used in this study.

Figures 1 through 4 show selected acoustic characteristics of the children’s productions used in this study in the three acoustic dimensions that we predict will be most closely related to individuals’ perception of these sounds, based on the findings of Li et al. (2011). Figures 1 and 2 show the two-dimensional m1 by onset F2 frequency space. Figures 3 and 4 show the two-dimensional m1 by m2 space. Figures 1 and 3 show the productions from the two-year-olds and Figures 2 and 4 show the productions from the three-year-olds. These figures illustrate three important features of the stimulus set. First, they show that there is considerable variation in the acoustic characteristics of the stimuli. As such, they are appropriate stimuli for eliciting a continuous goodness rating. Second, they show considerable overlap in all three acoustic parameters between the two languages. Thus, the effect of including stimuli from both languages is simply to increase the variation in the stimuli, rather than to introducing acoustic variation in a new parameter. Finally, they show considerable variation within each age, and only modest differences between the two ages, limited mostly to the onset F2 measure. The spectral variation is therefore unlikely to be due solely to the inclusion of children who vary in age and therefore in overall vocal-tract scaling. Rather, this greater spectral variation is more likely due to the inclusion of fricatives produced with different degrees of accuracy.

Procedures

The three tasks all used the same protocol for eliciting a forced-choice judgment: the stimulus was presented and the participant was asked to respond by pressing either 1 (for ‘s’) or 5 (for ‘sh’) on a computer keyboard. In the Categorization plus Likert Goodness Rating and Categorization plus Continuous Goodness Rating tasks, this was followed by the question "You said the syllable begins with the ["s" or "sh", depending on the listener's categorization] sound. Listen to it again. How good of an ["s" or "sh"] sound do you think this is?" For the Categorization plus Likert Goodness Rating experiment, participants were presented with a vertical line with the top labeled "Perfect" and the bottom labeled "Bad". The line was divided into seven equal sections and each was labeled with a number from 1-7 starting at the bottom.

The Categorization plus Continuous Goodness Rating experiment presented participants with a vertical double-headed arrow on the screen after they made their category judgment. The top arrow was labeled "Perfect" and the bottom was labeled "Bad". Participants were asked to click somewhere along the line to indicate the goodness of the sample. Clicks that fell outside of the line in either dimension were removed from analysis. The click location in pixels was noted. These ranged from 100 to 500, for 400 possible levels of response.

For the Categorization plus Magnitude Estimation experiment, participants were presented with the stimulus first and were asked to categorize it. Following that they were presented with the same stimulus a second time. After each stimulus was presented for the second time, the participants were prompted with a text box to give a number that represented the rating of that sample relative to the first one with respect to the amount of whatever quality they associated with being a good /s/ or a good /ʃ/. Given the complexity of this task, we have included the full set of instructions and scoring procedures in Appendix A. After the data collection, the responses were separated into two groups based on whether they were identified by the listener as /s/ or /ʃ/. The responses in each group were then modulus-equalized using the transformation proposed by Engen (1971) described in Appendix A. Critically, this normalization allows us to compare the performance of individuals who used very different initial ratings and ranges of ratings.

For all three experiments, a subset of 10% of the stimuli were played a second time to assess the consistency of ratings within individual listeners.

Results

Categorical Responses across Tasks

The first analysis examined the proportion of times that stimuli were judged to be /s/ and /ʃ/ as a function of the task, to ensure that the different tasks didn't bias listeners to provide a specific label. The dependent measure was simply the average proportion of /s/ responses across listeners. The proportion of sounds judged to be /s/ was similar across the three tasks (Categorization plus Magnitude Estimation: Mean=0.55, SD=0.11; Categorization plus Likert Goodness Rating: Mean=0.53, SD=0.08; Categorization plus Continuous Goodness Rating: Mean=0.52, SD=0.10). These differences were not significant in a Kruskall-Wallis nonparametric difference test, $\chi^2_{[df=2]} = 0.914$, p=0.633. This null result indicates that any differences in the properties of the Likert goodness ratings, continuous goodness ratings, or category estimates are not a consequence of task-related differences in the fricative categorizations. A standard statistic for inter-rater reliability for the categorization judgments was also calculated. Fleiss's k was found to be 0.56 for the Categorization plus Magnitude Estimate, 0.46 for Categorization plus Likert Goodness Rating, and 0.49 for Categorization plus Continuous Goodness Rating. Following the guidelines of Landis and Koch (1977), this indicates only moderate agreement among listeners. This moderate agreement further motivates the development of finer-grained rating scales.

Continuity of Responses

One desirable property of a rating scale is that it should elicit ratings that are continuous when the stimuli vary continuously in perceptually salient acoustic dimensions. Visual inspection of probability-density distributions of individual listeners' ratings suggested that there was variation in how continuous the ratings were. Figures 5 through 7 show three distributions of ratings made by three listeners in the Categorization plus Continuous Goodness Rating task. Figure 5 shows a listener whose ratings were well distributed across the continuum. Figure 6 shows a listener whose ratings were much less distributed. Figure 7 shows a listener whose dispersion is intermediate between the other two. To compare individual listeners, measures of the dispersion of ratings were calculated separately for the 61 listeners in the three tasks. These in turn were calculated separately for the sounds judged to be /s/ and the sounds judged to be /ʃ/ (referred to as 'category choice' for the remainder of this section). Prior to calculating the measure of dispersion, the scales used in the three tasks were normalized, so that the lowest rating given by an individual was set to zero and the highest rating was set to one. This ensured that any observed differences in dispersion of ratings across tasks were not due to the different scales used. The nonparametric statistic Interquartile Range (IQR) was used as the measure of dispersion. This was to account for the fact that the response distributions were often non-normally distributed, as seen in Figures 5 through 7.

The IQRs for the three tasks were analyzed using a two-factor, mixed-model ANOVA, with category choice as the within-subjects factor and task as the between-subjects factor. There was no main effect of category choice. There was, however, a significant main effect of task, F[2,58]=32.7, $p < 0.001$, $\eta^2_{partial} = 0.53$. This was qualified by a significant interaction between category choice and task, F[2,58]=13.1, $p < 0.001$, $\eta^2_{partial} = 0.31$. The highest dispersion was found for the Categorization plus Likert Goodness Rating and Categorization plus Continuous Goodness Rating tasks for both sounds judged to be /s/ and sounds judged to be /ʃ/. The interaction arose because the restricted range of responses in the Categorization plus Magnitude Estimation task was particularly pronounced for sounds judged to be /ʃ/.

Intra-Rater Reliability.

The next analysis examined how consistent listeners' ratings were. Two measures of intra-rater reliability were calculated for the subset of stimuli presented twice during each task. The first measure of reliability was the proportion of reliability trials on which the categorization of the stimulus was the same as the first presentation. The mean proportions of agreement were similarly high across tasks: 0.87 for Categorization plus Magnitude Estimation (SD = 0.04), 0.80 for Categorization plus Likert Goodness Rating (SD=0.12), and 0.83 for Categorization plus Continuous Goodness Rating (SD = 0.08). A univariate between-subjects ANOVA with task as the between-subjects factor was statistically significant, F[2,58] = 3.7, p = 0.03, $\eta^2_{\text{partial}} = 0.11$. Bonferroni-corrected post-hoc comparisons showed this difference to be due to the fact that Categorization plus Magnitude Estimation had higher agreement than did Categorization plus Likert Goodness Rating. The values for Categorization plus Continuous Goodness Rating did not differ from those for the other two tasks.

The second measure of reliability was the correlation between the first and second rating for the trials that were judged to be the same sound. The mean correlations were 0.22 for Categorization plus Magnitude Estimation (SD = 0.23), 0.38 for Categorization plus Likert Goodness Rating (SD=0.22), and 0.39 for Categorization plus Continuous Goodness Rating (SD = 0.18). A univariate between-subjects ANOVA with task as the between-subjects factor was statistically significant, F[2,58] = 4.2, p = 0.02, $\eta^2_{\text{partial}} = 0.13$. Bonferroni-corrected post-hoc tests showed that the Categorization plus Magnitude Estimation task was associated with lower reliability than either of the other two tasks, which did not differ from one another.

Predictive Validity: Comparison with Acoustic Measures.

The last analysis examined the relationship between individual participants' ratings of individual tokens and the five acoustic measures of those tokens: M1, M2, M3, M4, and onset F2. This was accomplished by conducting a series of multiple regression models. These were done separately for each listener, and only included responses to the tokens produced by children. They were also done separately for sounds judged to be /s/ and sounds judged to be /ʃ/, as we predicted that the predictors of /s/ and /ʃ/ would be different, based on the findings of Jongman, Wayland, and Wong (2000). All five acoustic measures were entered into the regression simultaneously, as we had no *a priori* predictions for which of these measures would best predict ratings. For the Categorization plus Continuous Goodness Rating task, 76% of the regressions were significant for the sounds judged to be /s/ and 100% for the sounds judged to be /ʃ/. For the Categorization plus Likert Goodness Rating task, 79% of the regressions for the sounds judged to be /s/ were significant, as were 84% of the regressions for the sounds judged to be /s/. For the Categorization plus Magnitude Estimation task, 100% of the regressions were significant for sounds judged to be /s/ and 90% for the sounds judged to be /ʃ/.

The R^2 values were mused as an index of the relationship between the ratings and the sounds’ acoustic characteristics. A two-factor mixed-model ANOVA was conducted with R^2 values as the dependent measure, task (3 levels: Categorization plus Continuous Goodness Rating, Categorization plus Likert Goodness Rating, Categorization plus Magnitude Estimation,) as the between-subject factor and category choice (2 levels: /s/, /ʃ/) as the within-subjects factor. The main effect of category choice was significant, $F[1,58] = 7.8$, $p = 0.007$, $\eta^2_{\text{partial}} = 0.12$. Greater variance was accounted for in the regressions predicting /ʃ/ judgments than in those predicting /s/ judgments. There was no main effect of task, nor did task interact with category choice, though there was a tendency for the most variance to be accounted for in the regressions for the Categorization plus Continuous Goodness Rating data and for the least to be accounted for in the regressions for Categorization plus Magnitude Estimate data (For the regressions predicting /ʃ/ judgments: R^2= 21.3%, SD = 8.2% for Categorization plus Continuous Goodness Rating, R^2=17.5%, SD = 7.4% for Categorization plus Likert Scale Rating, and R^2= 18.6%, SD = 4.7% for Categorization plus Magnitude Estimate; for the regressions predicting /s/ judgments: R^2= 19.1%, SD = 8% for Categorization plus Continuous Goodness Rating, R^2=15.5% , SD = 7.7% for Categorization plus Likert Scale Rating, and R^2= 15.6%, SD = 7.8% for Categorization plus Magnitude Estimate).

Discussion

The purpose of this report was to examine the utility of three tasks for rating children's sibilant fricative productions. A variety of different criteria were used to assess the utility of these measures. The first was whether the measure elicited a continuous response from listeners when presented with a continuous signal. In the most general sense, all three measures fared similarly well in that all elicited a range of responses from the listeners in this study. When the degree of continuity of response was measured, Categorization plus Magnitude Estimation was shown to elicit a narrower range of responses than did Categorization plus Continuous Goodness Ratings or Categorization plus Likert Goodness Ratings.

The second criterion was that the measure should have good intra-rater reliability. When we assessed the proportion of times that the same category judgment was made for repeated tokens, Categorization plus Magnitude Estimation fared better than the other two tasks. The second index of intra-reliability was a measure of the similarity between the first and second category goodness ratings. By this criterion the Categorization plus Magnitude Estimation task fared more poorly than the other two measures. The final criterion was of the correlation between the continuous ratings and the acoustic characteristics of the stimuli being rated. By this criterion the three measures fared similarly, though there was a statistically non-significant numeric advantage for the Categorization plus Continuous Goodness Rating measure.

In sum, the results of the analyses we conducted paint a mixed picture. Categorization plus Magnitude Estimation has the poorest utility overall. It had the lowest intra-rater reliability for within-category ratings, and it elicited the narrowest range of responses. Moreover, the Categorization plus Magnitude Estimation task is arguably the most cognitively taxing of the methods, as it requires listeners to compare two stimuli rather than simply compare a single stimulus to internal prototypes. Moreover, its scoring is complex and requires multiple steps. Indeed, participants in the Categorization plus Magnitude Estimation task commented on how unintuitive the instructions were. This was true even for participants whose responses were well correlated with the sounds' acoustic characteristics. This increases the likelihood that Categorization plus Magnitude Estimation would be hard to implement in real-world clinical settings, where the raters would have competing demands on their attention.

The two remaining scales, Categorization plus Continuous Goodness Rating and Categorization plus Likert Goodness Rating were of similar utility. The only advantage of Categorization plus Continuous Goodness Rating is that the responses it elicits are intrinsically more continuous than those from a Likert scale. This feature could be important if the goal of the measure were to document small changes in children's productions over a course of speech therapy. If the goal were simply to document within-category variation, Categorization plus Likert Goodness Rating is as good a tool as Categorization plus Likert Goodness Rating.

One limitation in this study is that it was impossible to assess inter-rater reliability for within-category ratings. Widely-used statistics for assessing the agreement of multiple raters on continuous scales do exist, such as the interclass correlation coefficient (ICC). However, the nature of the tasks in this study are such that there were many items intermediate between /s/ and some as /ʃ/ that were rated by some listeners as poor examples of /s/ and some as poor examples of /ʃ/. Statistics like the ICC are only appropriate for ratings made under identical circumstances. Hence, we are unable to report whether the three tasks we used differ in inter-rater agreement in category goodness ratings. A second weakness of this study is that it did not compare these methods with the current standard of care in assessment, phonetic transcription. It could be argued that a highly skilled clinician who is fluent with the full set of diacritics from the International Phonetic Alphabet could indeed capture subtle differences among productions using diacritics for productions between fronted variants of /ʃ/, retracted variants of /s/, etc. While that may be so, we argue nonetheless that the rating scales in this study are preferable because of their relative ease of use.

Though Categorization plus Continuous Goodness Rating and Categorization plus Likert Goodness Rating are clearly promising measures, the overall relatively modest variance accounted for by the ratings suggests that these methods need to be refined to before they are of true utility in studies of speech-sound development. There are a number of ways to accomplish this. The first of these is training. Individuals in this study were given no training on the use of these rating scales. A short segment of training with feedback might be sufficient to improve the reliability and the correlation between ratings and acoustics. The second of these is simply experience listening to children's speech. A recent study by Munson, Johnson, and Edwards (2012) measured ratings of a series of contrasts not examined in this study (/s/-/θ/, /t/-/k/, /d/-/ɡ/)

using a slightly different task from those used in this study. Specifically, Munson et al. (2012) asked people to rate sounds on a unidimensional VAS anchored with sound contrasts being rated (i.e., “the ‘t’ sound” and “the ‘k’ sound”). They found that ratings made by individuals with clinical training in speech-language pathology were more strongly correlated with the acoustic characteristics of the stimuli being rated than were those of a group of clinically untrained listeners, like those examined in this study. The degree of exposure needed to improve the correlation between acoustics and ratings is unknown, and is an important topic for ongoing research. A finding that relatively little exposure is needed to improve ratings would further strengthen the argument that rating scales like those examined in this paper should be used in clinical and observational studies.

Acknowledgments

The perception experiments were supported by NSF grant BCS-0729277 to Benjamin Munson. The stimuli were collected with support from NIH Grant DC02932 to Jan Edwards and Mary E. Beckman. Manuscript preparation was supported by a fellowship from the Center for Advanced Studies at Ludwig-Maximillians University, Munich. We generously thank Mary E. Beckman, Jan Edwards, Fangfang Li and Kathleen Hall for help with stimulus selection, acoustic analysis, and experiment design, and Eden Kaiser and Marie K. Meyer for help with participant testing.

References

Baylis, A.L., Munson, B., & Moller, K. (2011). Perception of audible nasal emission in speakers with cleft palate: a comparative study of listener judgments. Cleft Palate-Craniofacial Journal, 48, 399-411.

Berg, T. (1995). Sound change in child language: a study of inter-word variation. *Language and Speech, 38*, 331-363

Carney, A.E., Widin, G., & Viemeister, N. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America, 62*, 961-970.

Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic cues. *Cognition, 108*, 804-809.

Edwards, J., & Beckman, M.E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics, 22*, 937-956.

Engen, T. (1971). Psychophysics II: scaling methods. In J.W. Kling & L. Riggs (Eds.), *Woodworth and Schlossberg's experimental psychology* (pp. 47-86). New York: Holt, Rinehart, & Winston.

Gibbon, F. (1999). Undifferentiated lingual gestures in the speech of children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research, 42*, 382-397.

Holliday, J., Reidy P., Beckman, M.E., & Edwards J. (In Press). Quantifying the robustness of the English sibilant fricative contrast in children. *Journal of Speech, Language, and Hearing Research.*

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252-1263.

Kempster, G., Gerratt, B., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*, 124-132.

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.

Ladd, D.R. (2011). 2011. Phonetics in phonology. In J. Goldsmith, J. Riggle, and A. Yu (Eds.), *Handbook of Phonological Theory, 2nd Edition* (p. 348-373). New York: Blackwell.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Li, F. (2012). Language-specific developmental differences in speech production: a cross linguistic acoustic study. *Child Development, 83*, 1303-1315 .

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*, 999–1011.

Macken, M., & Barton, D. (1980). A longitudinal study of the acquisition of the voicing contrast in American English word-initial stops, as measured by voice onset time. *Journal of Child Language, 7*, 41-74.

Massaro, D., & Cohen, M. (1983). Categorical or continuous speech perception: a new test. *Speech Communication, 2*, 15-35.

Miller, J. (1994). On the internal structure of phonetic categories: a progress report. *Cognition, 50*, 271-285.

Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., & Meyer, M.K. (2010). Deconstructing Phonetic Transcription: Covert Contrast, Perceptual Bias, and an Extraterrestrial View of Vox Humana. *Clinical Linguistics and Phonetics, 24*, 245-260.

Munson, B., Johnson, J., & Edwards, J. (2012). The role of experience in the perception of phonetic detail in children's speech: a comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology, 21*, 124-139.

Munson, B., Schellinger, S., & Urberg Carlson, K. (2012). Measuring speech sound learning using visual analog scaling. *Perspectives in Language Learning and Education, 19*, 19-30.

Nelson, P.B., & Soli, S. (2000). Acoustical barriers to learning: children at risk in every classroom. *Language, Speech, and Hearing Services in Schools, 31*, 356-361.

Nicholson, H., Munson, B., Reidy, P., Beckman, M.E., & Edwards, J. (2015). Effects of age and vocabulary size on production accuracy and acoustic differentiation of young children's sibilant fricatives. *Proceedings of the International Congress on Phonetic Sciences*. Glasgow, Scotland: University of Glasgow.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *Journal of the Acoustical Society of America, 134*, 3781-3792

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*, 779-798.

Smit, A. (1993). Phonologic error distributions in the Iowa-Nebraska articulation norms project: consonant singletons. *Journal of Speech and Hearing Research, 36,* 533-547.

Toscano, J., McMurray, B., Dennhardt, J., & Luck, S. (2010). Continuous perception and graded categorization: Elecrophysiological evidence for a graded relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science, 21*, 1532-1540.

Williams, A.L. (2002). Perspectives in the assessment of speech. *American Journal of Speech-Language Pathology, 11*, 211-212.

List of Figures

Figure 1

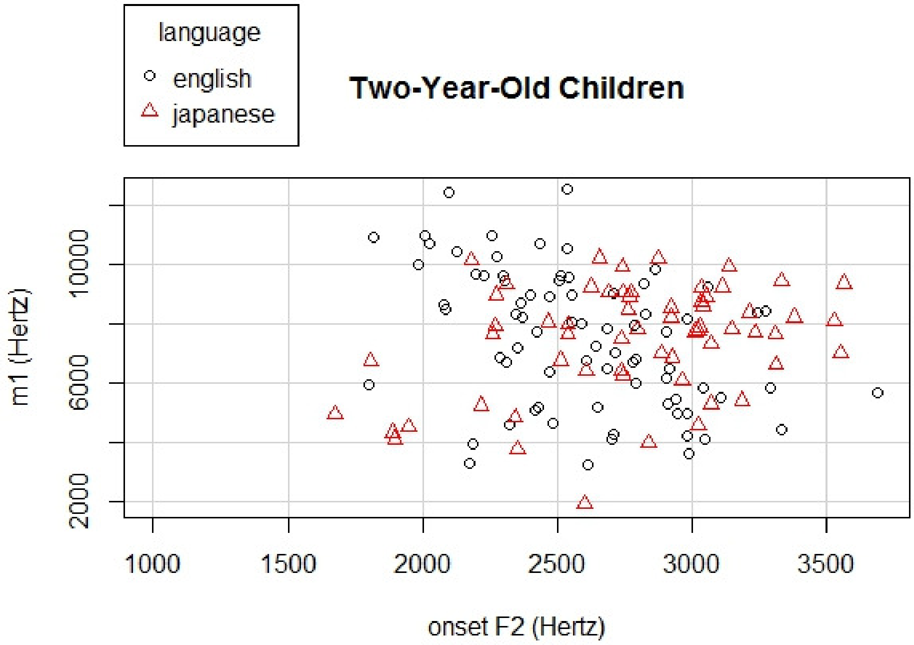

language
○ english
△ japanese
Two-Year-Old Children
m1 (Hertz)
2000
6000
10000
1000
1500
2000
2500
3000
3500
onset F2 (Hertz)

Figure 2

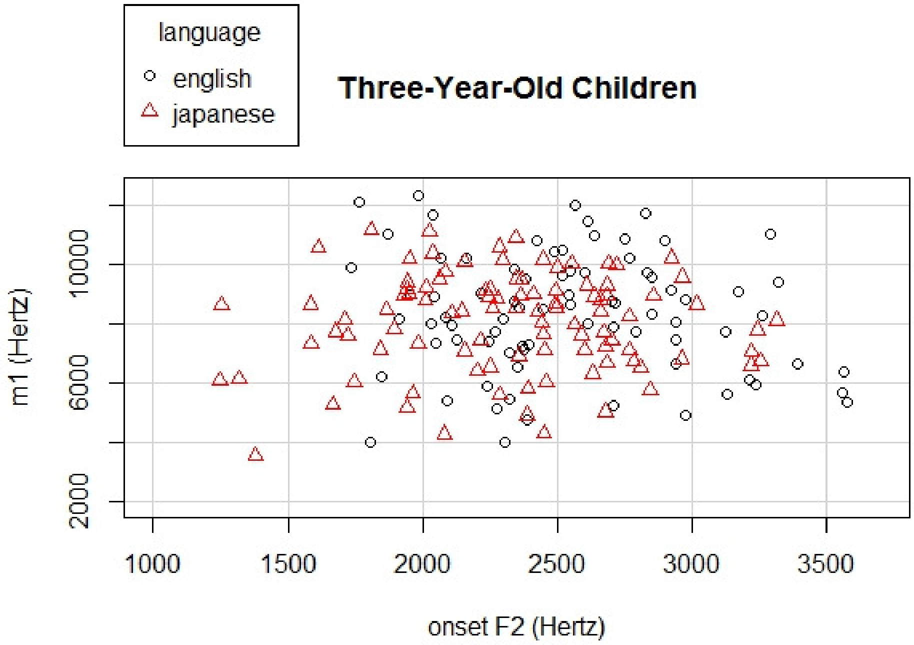

language
○ english
△ japanese
Three-Year-Old Children
m1 (Hertz)
2000
6000
10000
onset F2 (Hertz)
1000
1500
2000
2500
3000
3500

Figure 3

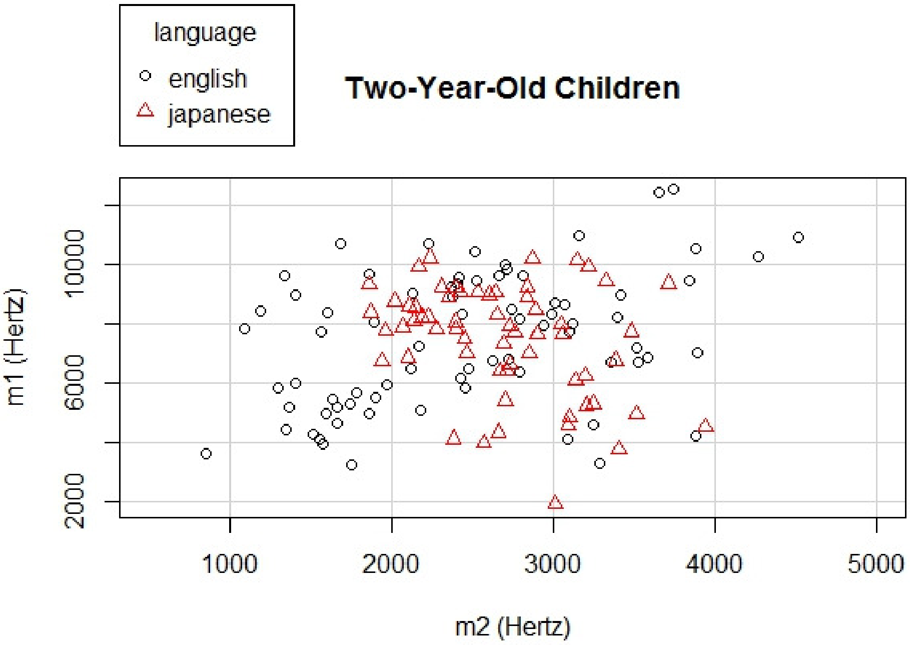

language
○ english
△ japanese
Two-Year-Old Children
m1 (Hertz)
2000
6000
10000
1000
2000
3000
4000
5000
m2 (Hertz)

Figure 4

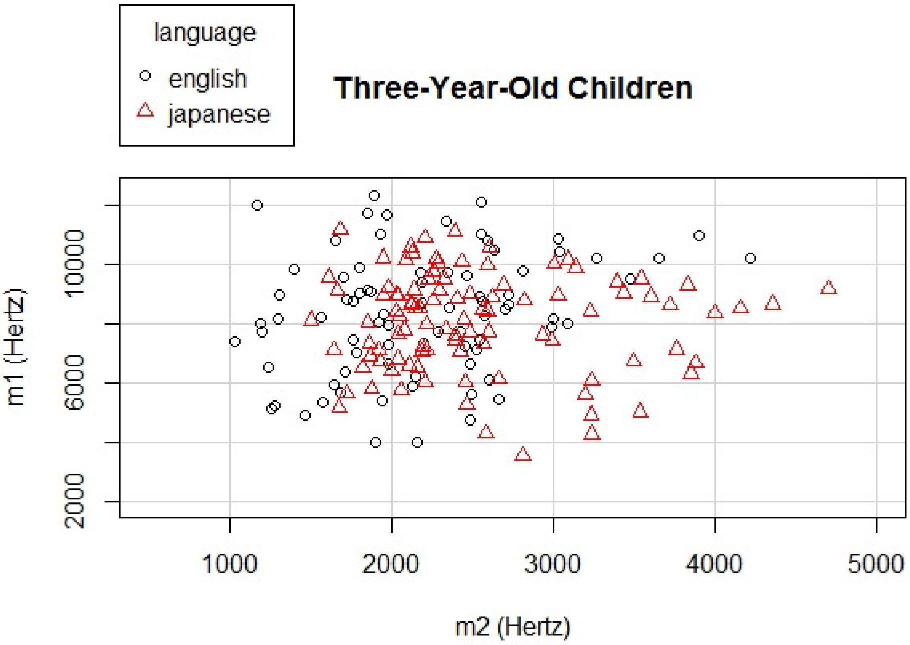

language
○ english
△ japanese
Three-Year-Old Children
m1 (Hertz)
2000
6000
10000
m2 (Hertz)
1000
2000
3000
4000
5000

Figure 5

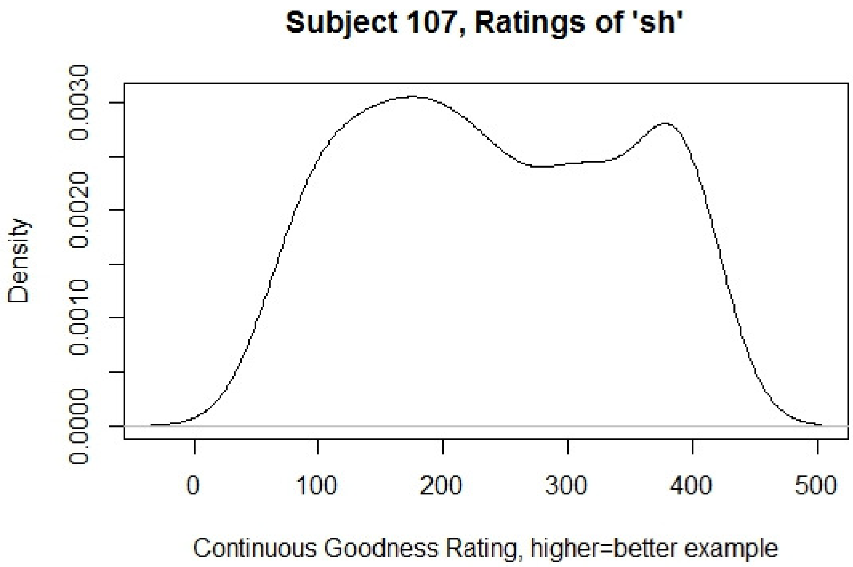

Subject 107, Ratings of 'sh'
Density
0.0000
0.0010
0.0020
0.0030
0
100
200
300
400
500
Continuous Goodness Rating, higher=better example

Figure 6

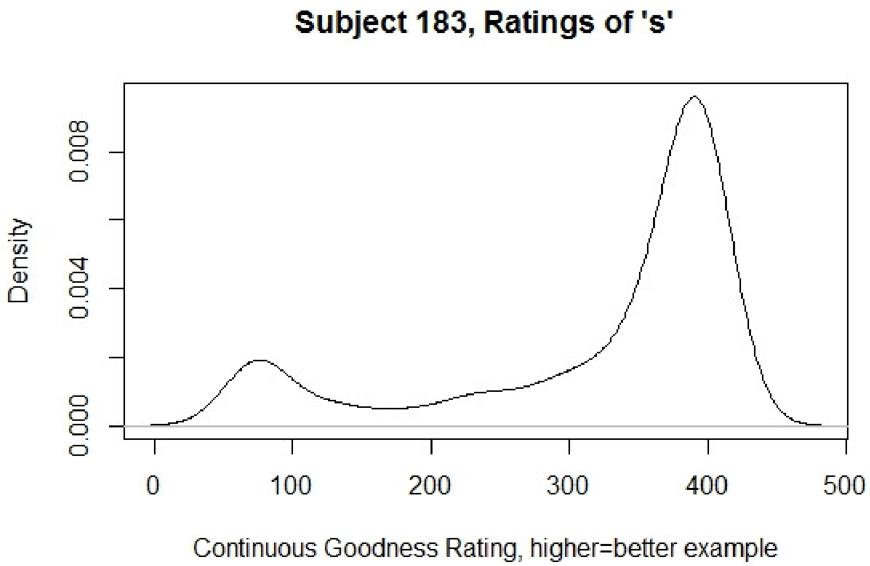
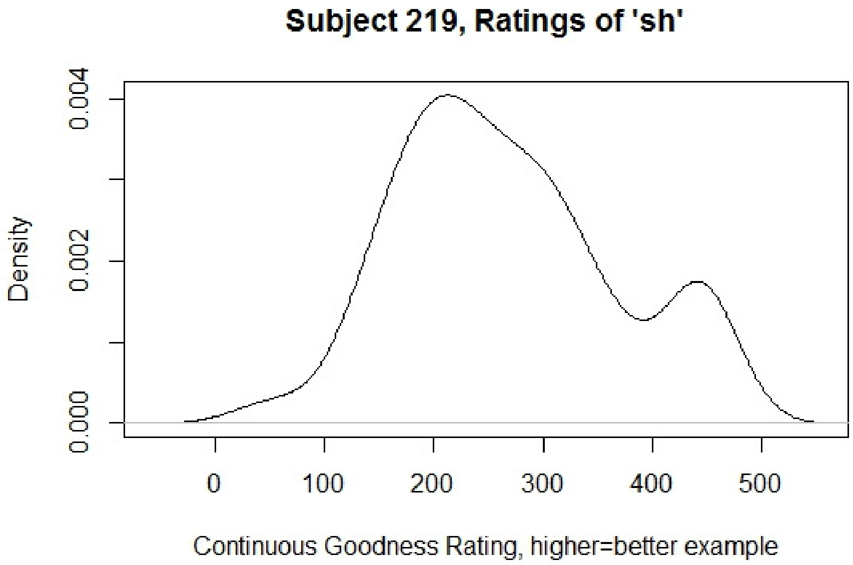

Subject 183, Ratings of 's'
Density
0.000
0.004
0.008
0
100
200
300
400
500
Continuous Goodness Rating, higher=better example

Figure 7


Subject 219, Ratings of 'sh'
Density
0.000
0.002
0.004
0
100
200
300
400
500
Continuous Goodness Rating, higher=better example

Appendix A: Instructions for the Categorization plus Magnitude Estimation Experiment.

For the FC-DME experiment, participants were given the following instructions at the beginning of the experiment:

> When you hear the first sample, give it a number that represents how good the sample is. You can give it any number that you think is appropriate. Give it any number you think is appropriate—we're not going to tell you to choose 1 through 100 or 1 to 7. Just choose a number that reflects your 'gut reaction'—the higher the number, the better the sample. Don't use negative numbers or decimals. You will then be presented the next sample to rate. If the second sample sounds better than the first, give it a higher number. If it sounds worse, give it a lower number. When you hear the next sample, try to make the ratio between the two ratings correspond to the ratio of "goodness" between the two samples. If the second sample is twice as "good" as the first, give it a number that is twice the number you gave before. If it is half as good, give it a number that is half the number you gave before.

After each stimulus was presented for the second time, the participants were prompted with a text box to give a number that represented the goodness of that sample. After the data collection, the responses were separated into two groups based on whether they were identified by the listener as /s/ or /ʃ/. The responses in each group were then modulus-equalized using the following transformation, following Engen (1971), where x_{ij} is the raw score of subject *i* for stimulus *j, n* is the number of stimuli and N is the number of listeners.

$$x'_{ij} = \ln(x_{ij}) + (\frac{1}{nN}\sum_{i=1}^{N}\sum_{j=1}^{n}\ln(x_{ij}) - \frac{1}{n}\sum_{j=1}^{n}\ln(x_{j}))$$

In steps, this process is as follows:

1. Each data point is converted to its natural log.
2. The arithmetic mean of the log-transformed data is calculated for each subject.
3. The arithmetic grand mean of the log-transformed data is calculated.
4. The subject mean is subtracted from the grand mean for each subject.
5. The result of step 4 is added to the log-transformed value for each data point.

The output of step 5 is the value assigned to each token for analysis. This effectively created a dimensionless scale that showed the relative goodness of each of the items.