Effects of Auditory Anchors on Perceptual Judgment of Hypernasality

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Lauren M. Derksen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Associate Professor Benjamin Munson, Chair
Dr. Leslie E. Glaze
Dr. Peggy Nelson, Department of Audiology

July, 2010

Acknowledgements

I would like to first gratefully acknowledge Benjamin Munson for his continued assistance and guidance, and motivation. I would also like to acknowledge Leslie Glaze and Peggy Nelson for their support as committee members. Acknowledgment is given to student research assistants who have contributed to completion of this thesis. Financial support was provided by the HSD Grant from the Department of Speech-Language-Hearing Sciences at the University of Minnesota.

Abstract

Perceptual judgment of hypernasality is a common and widely accepted practice among speech-language pathologists. However, because these judgments are somewhat subjective, reliability is an issue. This study examined the effect of auditory anchors on the validity of judgments of hypernasality in both natural and acoustically manipulated speech samples. In addition, this study investigated the effectiveness of auditory anchors developed using acoustic manipulation of first-formant bandwidth to simulate speech nasality. Anchors consisted of sentences of unprocessed speech and speech that had been acoustically altered by first formant bandwidth to 150 Hz, 300 Hz, and 500 Hz. The wider the bandwidth, the greater the expected nasality. Thirty subjects were assigned one of two groups. The “Anchor” group followed a training procedure to practice rating hypernasality by listening to speech samples and using visual feedback indicating most-correct judgments. The “No Anchor” groups were exposed only to speech with varied apparent nasality. All subjects then rated both acoustically manipulated as well as unprocessed speech and rated on perceived nasality of samples. Results indicated that unprocessed samples were perceived to be most natural, and those in the 150 Hz bandwidth condition were perceived to be the least natural. Surprisingly, samples with bandwidths of 300 and 500 Hz elicited ratings that were intermediate between unprocessed speech and 150 Hz. In this study, auditory anchors did not improve rater accuracy. To conclude, the training regimen presented in this study as a means for improving speech-language pathologists' rating of hypernasal speech can be confidently ruled out for practice of future professionals.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Background

A perceptual judgment of resonance is a clinical skill used by speech-language pathologists to determine the severity of a speech disorder. Perceptual listener judgment is widely accepted as a standard assessment for nasality and is used to validate instrumental measures (Kuehn & Moller, 2000). The use of perceptual judgments as an assessment technique can be validated given that the goal outcome of speech-language treatment is natural sounding speech, and not simply an improved speech signal as measured by a nasometer. Hypernasal resonance is a technical term for the excessive nasality that occurs when someone has either a structural or behavioral problem closing the nasal cavity off from the oral cavity during speech production (Moller & Glaze, 2009). A common characteristic of children with cleft palate is hypernasal resonance due to inadequate closure of the velopharyngeal mechanism. These children produce speech with excessive nasal resonance heard on both vowels and resonant consonants like /l/, /r/, and /w/ (Henningsson et al., 2007). These speech characteristics are assessed by judging the magnitude of hypernasality. Various audio-perceptual scales, such as an equal-appearing interval scale (EAI), a visual analog scale (VA), or direct magnitude estimation (DME), or scalar definitions ranging in severity levels have been used (Henningsson et al., 2007). Given that perceptual ratings are subjective to bias, decision making regarding candidacy and type of treatment is based upon the experience and perception of the clinician. Although speech-language pathologists working in the field of cleft lip and palate are trained to understand differences among mild, moderate and severe

hypernasality, subtle aspects of vocal and/or resonance quality may affect different professionals' ratings differently. In addition, other areas of speech likely to be affected by cleft palate (i.e. articulation deficits, like producing errors on /s/, /z/, and /r/) might interfere with speech-language pathologists' judgments of nasality, simply because these errors bias listeners, even skilled ones, to perceive a child as more severely impaired overall (Benoit, 2008)

When considering clinical implications of rating hypernasality it is important to explain the acoustics and aerodynamics of nasality, clarify the effect of nasality on a speech signal and describe tools for detecting and measuring resonance disorders. Nasalization is characterized acoustically by Lee et al., (2002) and is described as a decrease in amplitude of the first formant (F1). When amplitude of F1 is reduced, an increase in formant bandwidths and upward shifts in formant frequencies occur. A high-energy nasal formant is usually present between 600 to 1000 Hz and is accompanied by low energy of upper formants due to presence of antiformants. These characteristics vary with speakers and phonetic contexts. The nasometer is a computer-based tool designed to objectively measure nasality of a given speaker by computing a ratio of oral acoustic energy to the combined measure of oral and nasal acoustic energy (Mandulak & Zajac, 2009). This instrument uses a sound separator plate that rests on the upper lip and two microphones on either side of the plate that detect amounts of oral and nasal acoustic energy output during speech production. However, both the cost of the equipment and time constraints limit the availability of this objective measure of hypernasality.

Moreover, perceptual measures have greater face validity in assessing resonance disorders, which are perceptual phenomena (Moll, 1964).

Perceptual judgments have been used as standard practice in research alongside acoustic and aerodynamic measurements (e.g. de Krom, 1995, Moll, 1964); however, these measures are partially subjective and are commonly challenged by sub-optimal inter-rater reliability (Chan & Yiu, 2002). External factors, such as internal standards and different personal experiences can affect how each listener perceives various speech qualities and therefore may decrease inter-rater reliability (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). According to Gerratt et al. (1993) and Kreiman et al. (1993) these external factors can be affected by the acoustic context under which speech samples are evaluated. One way to improve reliability of perceptual ratings is to train listeners in their implementation. Moller and Starr (1984) developed training procedures using multiple listeners in an effort to decrease variability of voice and resonance ratings of children with cleft palate. More recently, investigators have examined two kinds of anchors on the reliability of perceptual judgments of different voice disorders. Auditory anchors present a speech sample of known severity while textual anchors provide written definition of the parameter being rated. These studies have shown that ratings of voices with varying types of voice disorders improve in inter-rater reliability when anchors are used (e.g. Chen & Yiu, 2002; Awan & Lawson, 2007). An anchor provides a reference point to a degree of vocal/resonance disorder severity and is thought to create greater consensus among raters. In addition, Chen and Yiu found that anchors made up of

synthesized signals were more effective in improving reliability of perceptual voice ratings than using natural voice anchors.

The current study examines the effects of specific auditory anchors on perceptual ratings of hypernasality made by undergraduate and graduate students of speech-language pathology. The following introductory sections review the literature in this topic and outline the goals of this study.

Literature Review

The literature will be reviewed in four parts: 1) Reliability of Perceptual Voice Evaluation, 2) Perceptual Rating of Hypernasality, 3) Acoustic Theory of Nasalization and 4) Parameters for Measuring Speech Characteristics in Individuals with Cleft Palate. Sections 1 and 2 review perceptual ratings of speech parameters. These ratings have been studied in great detail in the assessment of both voice disorders and resonance disorders; hence, reviews of both literatures will be provided.

Reliability of Perceptual Voice Evaluation

Perceptual evaluation of voice is often used to assess vocal quality during clinical evaluation and to compare with acoustic and aerodynamic measurements (Chan & Yiu, 2002). Like perceptual ratings of nasality, improvement of reliability of ratings of voice can be explored through use of specific training procedures. Research supports the use of anchor training to provide greater inter-rater and intra-rater reliability.

Chen and Yiu (2002) investigated the effect of anchors and training on the reliability of perceptual voice evaluation. Twenty-eight native Cantonese speakers all undergoing training to be speech pathologists judged a set of speech samples ranging in

severity of vocal roughness and breathiness. Speech samples were both recordings of natural speech and synthesized prototypes developed by use of HLSyn Speech Synthesis System. Each participant took a pre-training rating test as a baseline measurement; a training session consisting of listening anchors with suggested ratings of roughness and breathiness as well as written definitions of roughness and breathiness; and a post-training session. The authors found that anchors and training helped to improve the reliability of perceptual voice evaluation. Also, synthesized voice signals were found to be judged more reliably than natural speech samples.

Awan and Lawson (2009) examined the effect of anchors and training and modality of anchors provided (textual versus auditory) on inter- and intra-rater reliability of perceptual analysis of voice types and severities. Forty inexperienced judges rated 36 voice samples based on perceived presence and severity of vocal breathiness, hoarseness and roughness. Subjects were randomly assigned to one of four conditions of: No Anchors, Textual Anchors (written definition), Auditory Anchors (auditory voice samples with type and severity provided), and Combined Textual/Auditory Anchors. All conditions were conducted through use of a computer program. Results indicated inter-rater reliability improved more with auditory anchors in comparison to textual anchors or no anchors at all when rating mild and moderate breathiness and hoarseness in voice disorders. These improvements were determined by mean correlations and 95% confidence intervals. Use of textual anchors resulted in improved inter-rater reliability in comparison to no anchors provided, but not as much compared to auditory anchors. However, the Combined Textual/Auditory Anchor group showed the greatest degree of

inter-rater reliability based on correlations. These results indicate that the use of anchors may improve reliability of rating voice quality and severity of that quality. The authors also concluded that the use of auditory and textual anchors they do not substantially add to time required completing a voice-rating task.

Perceptual Rating of Hypernasality

Like assessment of voice disorders, resonance disorders are most commonly assessed with the use of perceptual listener judgment. As previously stated, this clinical method is regarded as having high face validity, but less than optimal inter-rater reliability. The following study outlines characteristics of perceptual ratings of nasality and potential methods of improving the reliability of these judgments. Lee et al. (2008) provided evidence that practice in general is useful for improving reliability of hypernasality ratings. This study was developed by dividing judges into three training groups. All groups took part in a "listener training session." The first group was simply exposed to a series of speech samples containing various degrees of hypernasality, nasal air emission, voice disorders and articulation errors. The second and third groups' listener training session was composed of four parts. For part 1, listeners were given the opportunity to listen and judge the presence and absence of the various disorders. Part 2 was an identification task in which listeners heard a speech sample then selected the presence of 'hypernasality', 'nasal air emission', 'voice disorder', articulation error' or any combination of these. Part 3 was a paired comparison task where listeners heard two speech samples and selected the one that sounded more hypernasal. Part 4 consisted of

rating hypernasality using Direct Magnitude Estimation (DME) in 5 samples of Dysarthric speakers and 5 samples of cleft palate speakers. Listener training for Group Three varied from Group Two in that Group Three was given feedback of the most accurate answer after each rating. For that reason, Group Two was referred to as the "Practice-Only Group" and Group Three as the "Practice-Feedback Group." For the final task, all groups rated 22 natural male and female speech samples for presence and severity of hypernasality using DME. Results indicated that both the Practice-Only and Practice-Feedback Group showed fair-to-good inter-judge reliability and that ratings were more reliable with female speech samples versus male. The authors concluded that practice in general with or without feedback is beneficial in improving inter-judge reliability of perceptual hypernasality of speech.

Acoustic Theory of Nasalization

These studies outline the process of how speech samples can be acoustically manipulated to increase the perception of nasality. The application of one-third octave analysis for measurement acoustic correlates of hypernasality was performed in a study by Lee et. al (2002). This study consisted of 12 Cantonese speakers with hypernasality as a result of dysarthria, maxillectomy or cleft palate as well as 12 normal Cantonese speakers. Speech samples were derived by segmenting the vowel /i/ from two single Cantonese words. Speech samples were analyzed using Praat version 4.0.1 (Boersma & Weenink, 1999-2001). One third octave analysis was applied to the samples by measuring the amplitudes of one-third octave bands progressing from 125 to 6300 kHz of

the 50 ms portion of the vowel. Results indicated that hypernasality was characterized by an increase in amplitude between F1 and F2 and a decrease in amplitude at F2.

The effects of altered fundamental frequency on nasalance were observed in a study by Mandulak and Zajac (2009). This study examined the effects of altered fundamental frequency (F0) on nasalance levels of speech production of the vowels /i/ and /a/. Participants performed two tasks. For the first task, participants produced tokens of each vowel at a sound pressure level (SPL) of 75-85 dB to represent a medium speaking level for adult men and women. For the second task, participants produced vowels at a the target SPL level as well as a targeted fundamental frequency of 165-175 Hz to represent a range of expected fundamental frequency levels of adult men and women. Percentage of nasalance of vowels was measured by use of the Nasometer 6200 and Computerized Speech Lab Model 4400. Results of these measures revealed that on average male speakers increased their fundamental frequency by 60 Hz during the second task. By increasing fundamental frequency, nasalance during production of the /a/ vowel increased in male speakers. This suggested a possible effect of fundamental frequency and measurement of nasalance.

Chen (1995) performed acoustic analysis and systematic acoustic manipulation to determine acoustic parameters of nasalized vowels in hearing impaired and normal hearing speakers. Through acoustic analysis it was theorized that nasalized vowels are characterized by extra pole-zero pairs and widened formant bandwidths. For example, a speaker with a larger velopharyngeal opening would produce an /a/ vowel that shifts the extra zero to a higher frequency which moves it farther away from the extra pole, causing

a more prominent peak in the spectrum. The prominence of the extra peak is distinguished by it amplitude (P1). Loss is reflected in the bandwidth of the first formant. These losses influence the prominence of that formant peak and are referred to as A1. A larger bandwidth causes the amplitude to decrease (A1-P1). In theory, a vowel with a smaller A1-P1 should be perceived as more nasal. Utterances judged by subjects to be highly nasal had an A1-P1 of less than 10dB. The correlation coefficient between the acoustic manipulation and nasality judgments was -0.82.

Parameters for Measuring Speech Characteristics in Individuals with Cleft Palate

To ensure greater consistency in reporting speech outcomes for treatment planning, Henningsson et al. (2008) proposed parameters of measuring speech characteristics to be used universally by speech-language pathologists and other health professionals. Speech samples should consist of 15-20 sentences. Hypernasality in cleft palate speakers should judged by listening to 25-30 single words or 15-20 short sentences. Sentences should each contain voiced and unvoiced pressure consonants, one target consonant type per sentence and at least two to three target words. An example sentence could be: “Buy baby a bib,” with the target voiced consonant being /b/ or “Sissy saw Sally race” with the voiceless consonant being /s/.

Purpose of the Study

There are two purposes of this study. The first is to simply determine whether anchors altered to simulate hypernasality improve validity of judgments of nasality, just as they have been shown to improve ratings for disordered voices. The second purpose is

to examine whether or not a group that is trained with anchors will rate unmodified speech samples differently when compared to a group that has not been trained with anchors.

This study intends to provide further insight on ideal training procedures for students and newly certified professionals. Experienced professionals may be able to easily detect differences in severity; however, speech-language pathology students and new clinicians typically have heard hypernasal speakers, but have the least amount of experience. For this reason they could be more susceptible to bias when rating hypernasality in a speaker with articulation errors. Previous research has demonstrated that ratings of hypernasality could also be improved with the use of auditory anchors. Improved accuracy and reliability will be demonstrated if a closer relationship is found between known nasality. More specifically, speech samples are acoustically altered to give the illusion of increased of nasality; hence, we are interested in whether there is a closer relationship between 'apparent nasality' conditions and ratings for a group trained with anchors than one trained without. We also examined whether there were differences between those ratings of hypernasality and those made previously for a group who received training with anchors when compared to a group who did not receive such training.

METHODS

Subjects

Subjects for this study were recruited from students from the University of Minnesota with a declared undergraduate major in Speech-Language-Hearing Sciences or Master of Arts program in Speech-Language Pathology. Subjects were recruited by use of flyers posted in Shevlin Hall and announcements made in undergraduate and graduate courses. The selection criteria were as follows:

1) Subjects must be native speakers of English. This was required to eliminate confounding effects of second language learning on speech perception.

2) Subjects must be declared in the undergraduate major of Speech-Language-Hearing-Sciences or Master of Arts program in Speech-Language Pathology. This was required to help ensure a higher level of ecological validity. Students declared in these undergraduate and graduate programs are considered more likely to make perceptual ratings of hypernasality in their future professions compared to other students attending the university.

Using the above selection criteria, 28 females and 2 males constituted the subject group. All subjects were between the ages of 18-27 years with a mean age of 23.07. Recruiting a more gender-balanced sample of subjects was limited by gender demographic based on selection criteria. (See Appendix for Subject Demographic Information)

Procedures:

Speech Samples

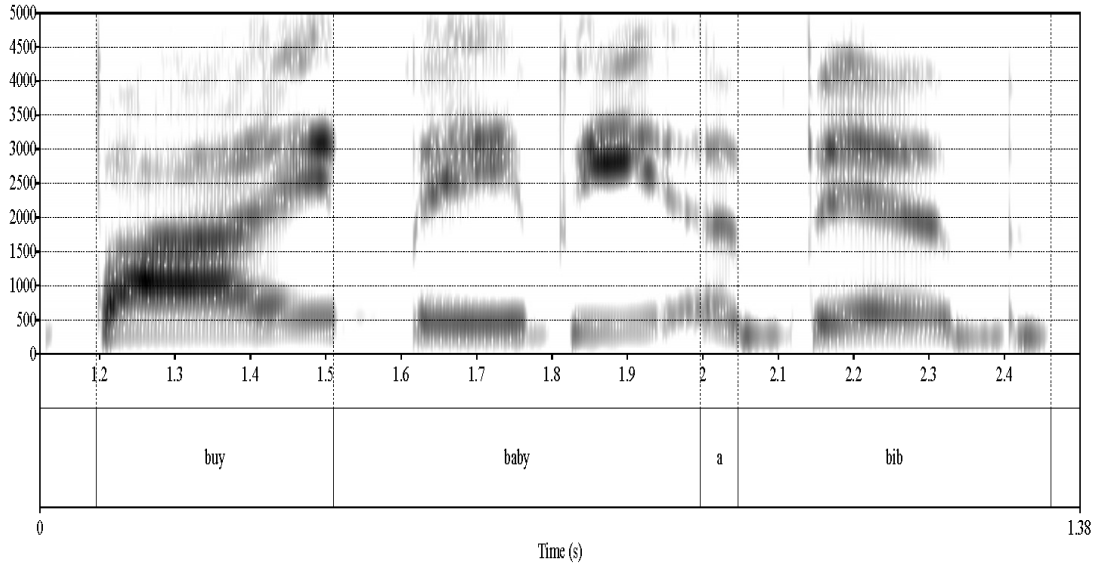
Speech samples presented in the *Exposure, Auditory Anchor,* and *Rating* tasks were derived from a recorded speech of an adult female speaker with normal resonance. An adult female speaker was selected with a clear speaking style and modal voice quality. This speaker produced 15 sentences. Sentences were developed following the framework of Henningsson et al.'s *Universal Parameters for Reporting Speech Outcomes in Individuals with Cleft Palate* (2008). (See Appendix for complete list of sentences). Samples were acoustically altered using the Praat (Boersma, 2001), an acoustic analysis and synthesis computer software program. Sentences were acoustically altered to give the illusion of increased nasality. The acoustic manipulations were based on the acoustic theory of nasalization and nasality in speech described by Chen (1995). Chen found that nasality increased the bandwidth of the first formant frequency. Consequently, to give the illusion of differences in nasality, the first bandwidth of the first formant of each sample was widened in order to increase perceived speech nasality of that sample. Each of the 15 sentences was altered in the first formant bandwidth three times to simulate various levels of nasality. Bandwidths were widened to 150, 300, and 500Hz (See Figure 1). Each sentence with no acoustic alteration was used as well as a "most natural" or "least severe" baseline for subject raters.

Speech samples presented in the *Generalization* task consisted of 10 natural speech samples that varied from normal resonance to severe nasality (as rated by a speech-language pathologist). These samples were unidentifiable and derived from Benoit et al. (2008). All speakers read the standard phonetically balanced passage of "Lazy Jack" (See Appendix for passage). The entire sample was used as the stimulus. As described by Benoit et al., the naturalness of these was rated using a Visual Analog Scale Anchored by the text "Most Severe" and "Least Severe." The ten samples were selected intentionally to represent a range of measured naturalness.

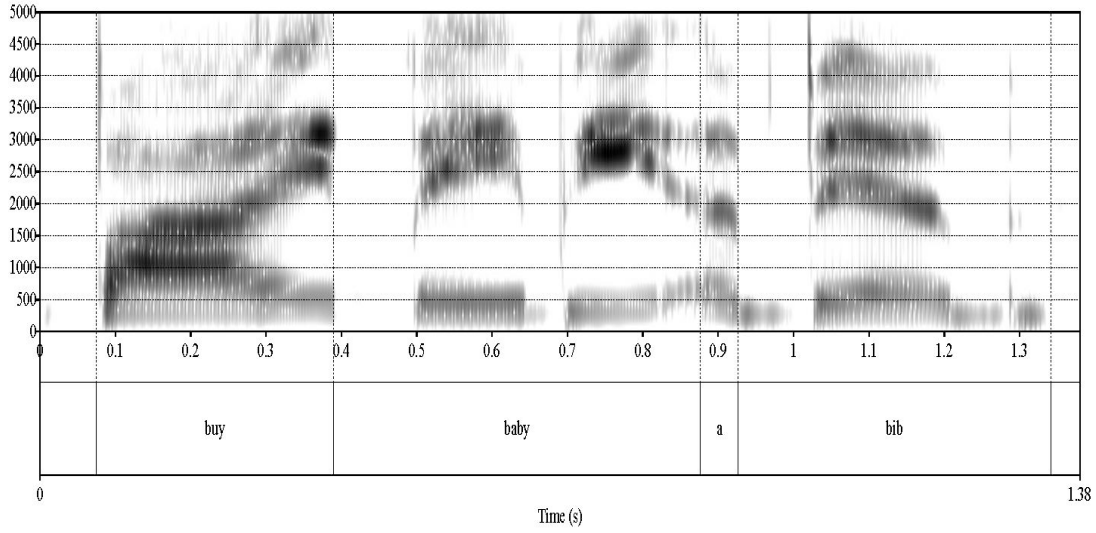
Presentation of Stimuli and Rating Procedure

A total of 70 speech samples (15 sentences of unmodified speech, 45 sentences with first-formant bandwidths altered by Praat, and 10 passages of unmodified speech) were presented as stimuli. The experiment was run using E-Prime software on a computer monitor. Subjects used a computer mouse and keyboard to navigate instructions and rate speech samples. A Visual Analog Scale (VAS) was presented exposure to each auditory stimulus item (See Figure 2). Text presented under the VAS read "Most Severe" at one endpoint and "Least Severe" at the other. Subjects were instructed to click the mouse anywhere along the scale to reflect perceptual ratings of presence and severity of nasality of each of the stimuli.

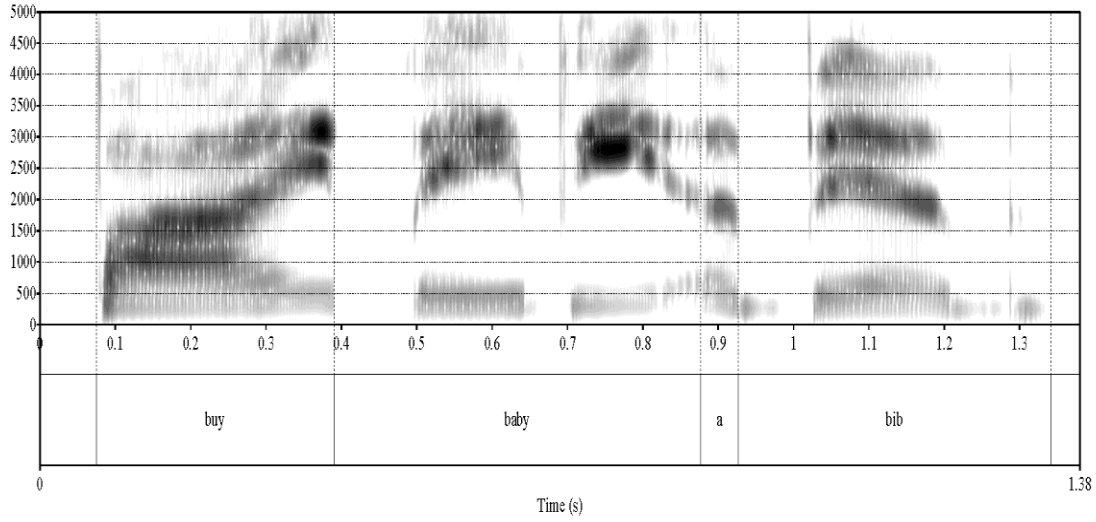
Figure 1. Spectrograms displaying widened first-formant bandwidth at 150, 300, and 500 Hz



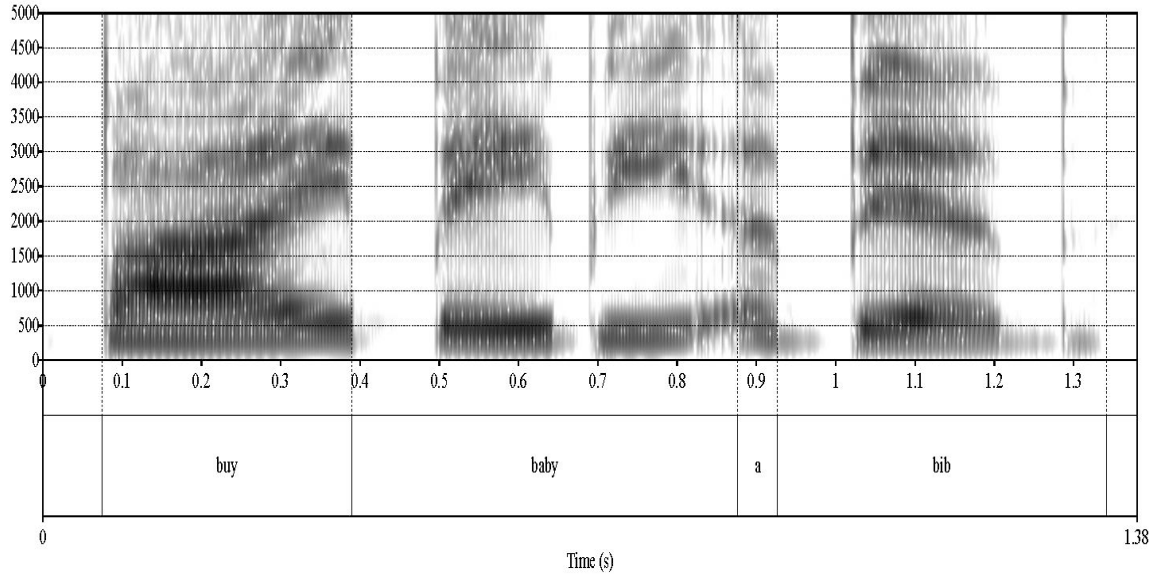
150 Hz


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
buy
baby
a
bib
0
1.38
Time (s)

300 Hz


5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
buy
baby
a
bib
0
1.38
Time (s)

500 Hz



Full Spectrum

Figure 2. Display of the VAS scale used to elicit responses

Subject Testing

All testing was completed during one session, lasting approximately one hour. Each subject was tested individually through headphones in a sound protected booth to control confounding effects of hearing sensitivity on speech perception tasks. Thirty subjects were divided into two groups of 15 subjects each: The "Anchor" Group and the "No-Anchor" Group.

Both groups were presented with a written definition of hypernasality and exposed to 16 auditory samples (8 different samples, each repeated twice in randomized order) of sentences with varied altered bandwidth levels. Following each auditory exposure, a red X was shown on a VAS indicating where on a scale of "Most Severe" to "Least Severe" the stimulus would be ranked in terms of nasality.

"Anchor" Group: Subjects were presented with four tasks. First, an exposure task was presented providing subjects with a written definition of hypernasality, visual

presentation of a Visual Analog Scale and an explanation of how it was used to judge speech in terms of hypernasality. Auditory exposure to 8 speech samples was also presented in this task with visual presentation of a red X on the VAS line indicating where an "expert" might rate the sentence on a continuum of "Most Severe" to "Least Severe." Second, an auditory anchor task was presented in which subjects listened to 16 sentences, independently rated each sentence for perceived nasality by clicking a location on the VAS continuum, followed by visual feedback of the ideal location on the VAS line for each rating. Feedback consisted of a red X presented on the VAS at the "Most Severe" end, "Least Severe" end, or at predetermined points in the center of the continuum line in accordance with ratings of mild or moderate severity. This task was created to serve as a reference point and practice for subjects to increase their rating reliability on subsequent tasks. Third, a rating task was presented in which subjects listened to 60 acoustically altered sentences and independently rated them for perceived nasality. No visual or auditory feedback was given as to correctness of ratings. Fourth, a Generalization Task was presented in which subjects listened to 10 unmodified speech samples consisting of speakers reading the "Lazy Jack" passage. Subjects again rated these stimuli using the same VAS ranging from "Most Severe" to "Least Severe."

"No Anchor" Group: The "No Anchor" Group was presented with the same protocol as the "Anchor Group" with the absence of the "Anchor Task." Therefore, this group was presented with three tasks the Exposure Task, the Rating Task, and the Generalization Task. Each of these tasks was presented identical to that of the "Anchor" Group with the

exception of the Rating Task which was lengthened by 8 additional speech samples to attempt to equate length of total task length between the two groups.

DATA ANALYSIS

For this analysis of data, the dependent measure was "Click location in pixels, ranged from 90 to 535." Each subject's average click locations for the three bandwidths samples were logged. The three manipulated bandwidths and the unmodified samples were used as the dependent measure in an ANOVA examining the effect of bandwidth, the effect of group, and the interaction between them. In addition, the ratings for individual natural-speech samples were used as dependent measures in a linear mixed-effects model, which examined (a) whether these differed as a function of training group, (b) the strength of the association between the ratings from Benoit et al. and the ratings made in the current study, and (c) whether the strength of the association between the previous ratings and the current ratings was statistically equivalent.

RESULTS

Analysis of Variance. The first analysis examined the average ratings for the four bandwidth conditions ($F[3,54] = 217.353$, $p < 0.001$, $\square^2$partial = 0.89). The main effect of training group did not achieve statistical significance at the $\square < 0.05$ level, but did approach this level ($F[1,28] = 3.964$, $p = 0.056$, $\square^2$partial = 0.12). Finally, these two factors interacted significantly ($F[3,84] = 3.379$, $p = 0.022$, $\square^2$partial = 0.11). The interaction between group and bandwidth condition is illustrated by comparing the bar

heights in Figure 2. First, this Figure shows that the marginally significant main effect of group was due to the listeners in the Anchor condition rating the samples as more-natural sounding. Second, this Figure shows that the different bandwidth conditions elicited significantly different ratings. As expected, the unmodified samples elicited ratings indicating that they were perceived to be most natural, and those in the 150 Hz bandwidth condition elicited ratings indicating that they were perceived to be the least natural. Surprisingly, the other two bandwidths elicited ratings that were intermediate between these. This is contrary to predictions, as we expected these ratings to indicate that listeners perceived them as more severe than the 150 Hz condition. Post-hoc Bonferroni-corrected paired comparisons showed that all pairwise differences were significant for both the Anchor and No Anchor groups. The interaction arose because the two groups differed only in their perception of the 150 Hz and 300 Hz samples.

Linear Mixed-Effects Model. A second analysis consisted of a linear mixed-effects model that examined the ratings of the ten natural-speech samples. As described previously, this model assessed (a) whether these ratings differed as a function of training group (b) the strength of the association between the ratings from Benoit et al. and the ratings made in the current study, and (c) whether the strength of the association between previous the previous ratings and the current ratings was statistically equivalent. This was implemented in R using the lmer function. Listeners and items were treated as random effects. The dependent measure was the ratings from the current study. Because the design of this study was nested (i.e., the ten speech samples were nested within the two groups, Anchor and No Anchor), a hierarchical design was used, in which the level 1

factor was ratings from the previous study, and the level 2 factor was group. Two models were run, one in which there was no interaction term between the level 1 and 2 factors, and one in which there was.

The average ratings for the 10 individual samples by the two groups, as compared to the ratings made in Benoit et al, are shown in Figure 3. This figure illustrates two interesting findings. First, consistent with the analysis of the sentences, the No Anchor group rated samples as less natural than did the Anchor group. Second, and quite surprisingly, there was a negative relationship between the measures made in the previous study and those made in the current study, a fact to which we return in the discussion. The results of the statistical model applied to this data showed that the relationship between ratings from the current study and ratings from the previous study was significant ($t = -2.911$, $p = 0.0039$). Moreover, these ratings differed significantly between groups, ($t = -1.976$, $p = 0.0491$). When a second model was run with an interaction term between the level 1 and level 2 factors, this interaction was found not to be significant ($t = -1.2941$, $p > 0.10$). That is, though the regression lines for the two groups in Figure 3 are not strictly parallel, the divergence between them is not so great as to achieve statistical significance in this model.

Figure 3.

Average ratings for the Anchor and No Anchor Groups, separated by bandwidth condition.

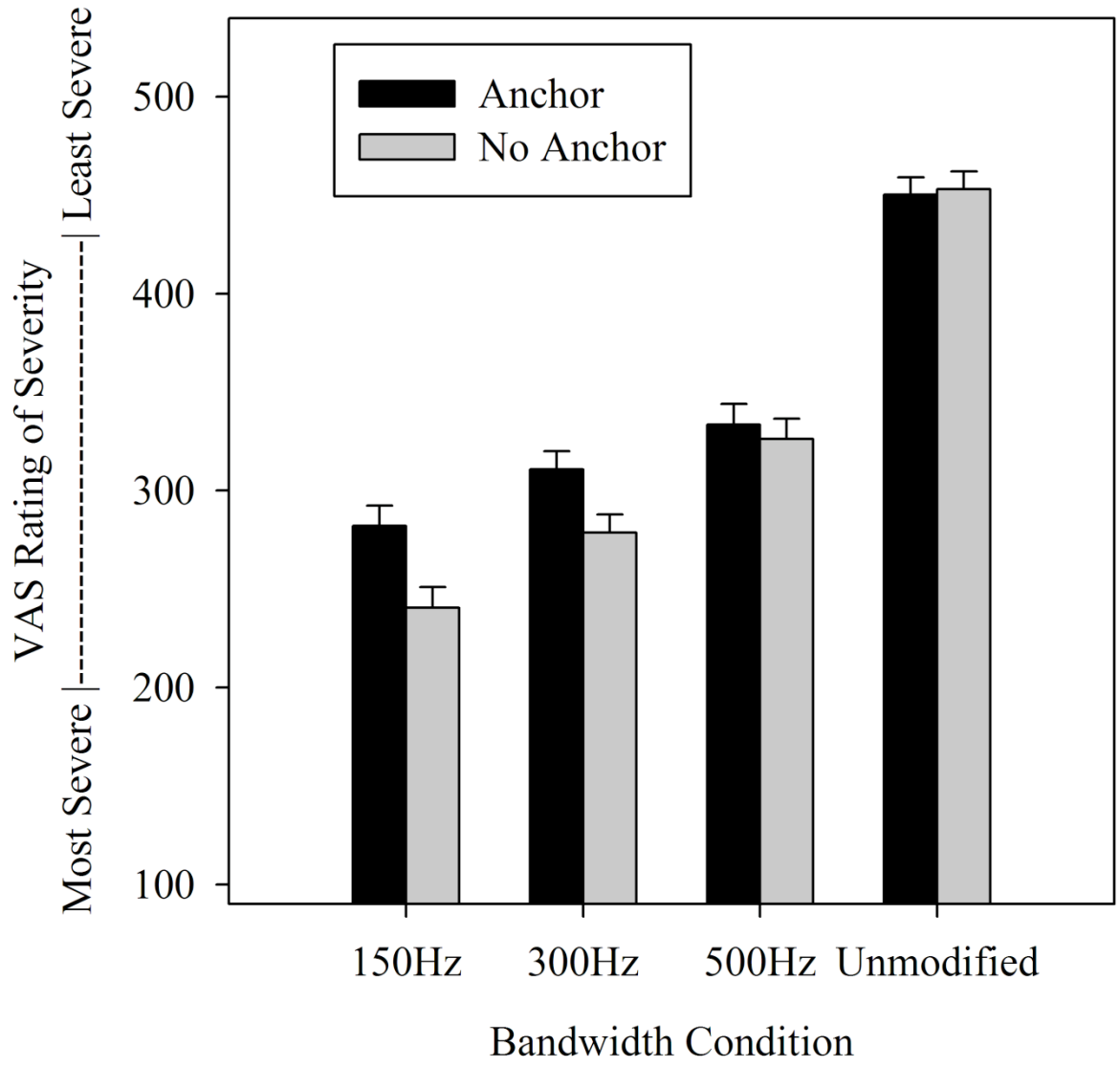
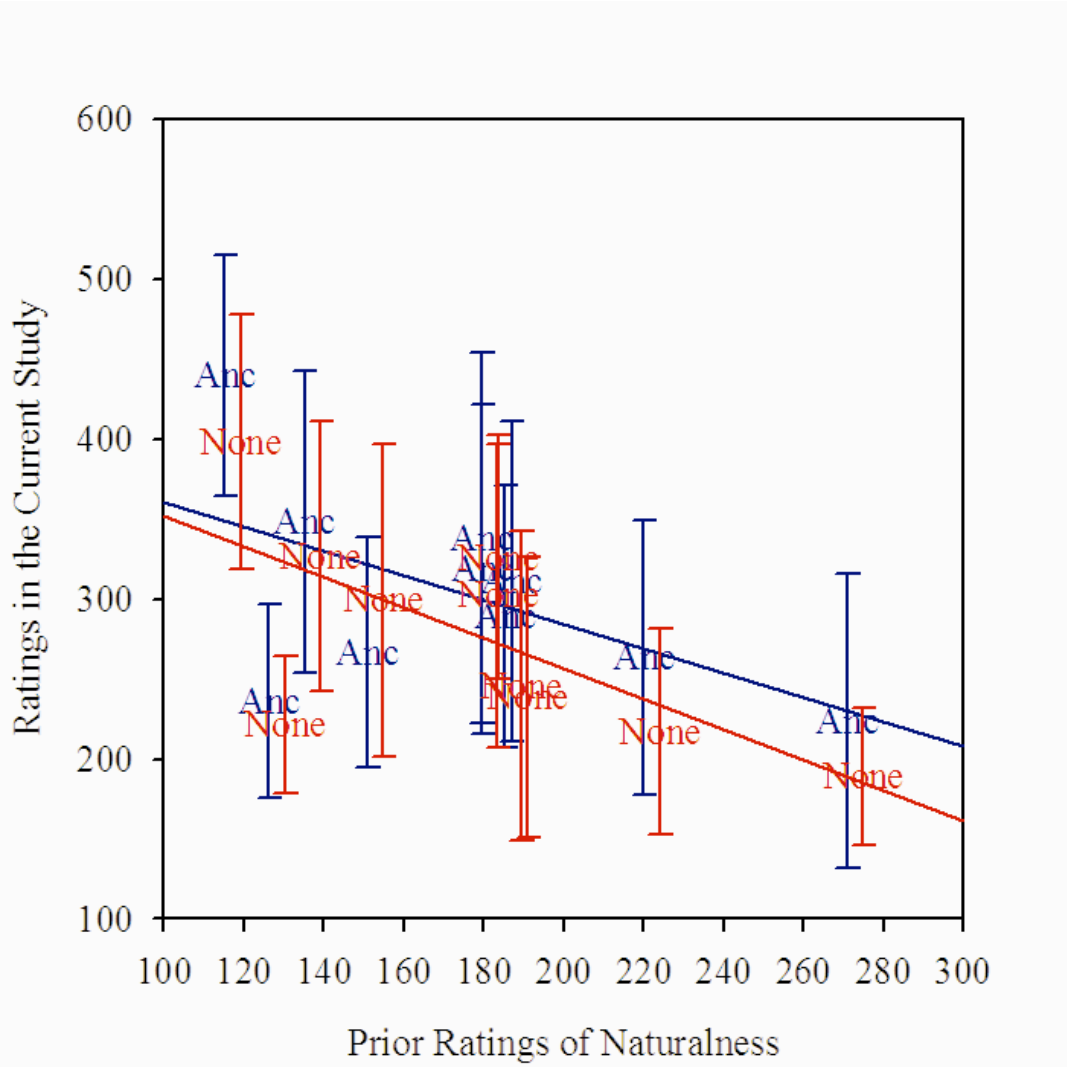


Figure 4.
Correlations between the average ratings of naturalness for the Anchor and No Anchor groups in the current study, as compared to the ratings made in Benoit et al.

Error bars are +/- one Standard Deviation of the mean. Separate linear regressions are fitted by group.



DISCUSSION

This study was meant to further investigate the use of auditory anchors as a training procedure for perceptual speech ratings of nasality. The findings reveal that use of a training regimen in which labels were associated with auditory anchors made by widening of first-formant bandwidth did not directly improve accuracy or reliability of perceptual ratings of speech nasality. Indeed, the bandwidth manipulations themselves did not have the expected effect on the perception of nasality: though more-nasal speech is associated with wider formant bandwidths, listeners did not rate samples with extremely widened bandwidths (i.e. widened to 300 Hz and 500 Hz) as more nasal than samples with more modestly widened bandwidths (i.e., widened to 150 Hz). This suggests that the use of auditory anchors altered acoustically using a computer software tool such as Praat may not be an ideal method of training clinicians to rate presence and severity of hypernasality. The effects of the particular auditory anchors used in the present study were quite different compared to the findings of Lee et al. (2008). Instead of yielding more reliable, accurate perceptual ratings of hypernasality, presentation of these auditory anchors led listeners to perceive the altered samples as more natural, rather than less natural as compared to a control group who was not presented with anchors. Put differently, presentation of these anchors appeared to negatively affect listeners' judgments of both the acoustically manipulated speech samples and the unmodified speech samples. In both the "Anchor" and "No Anchor" groups, listeners did detect a difference in naturalness between the unmodified speech samples and samples with altered bandwidth conditions, but did not perceive samples with a larger bandwidth size

as “More Severe” as was expected. Instead, listeners rated samples with a first-formant bandwidth of 150 Hz as more severely hypernasal than samples with a bandwidth of 300 Hz and 500 Hz. When rating perceived nasality of 10 unmodified speech samples, both listener groups responded with average ratings reversed from the ratings of Benoit et al. (2008). That is, samples rated as “Least Severe” by listeners in the present study were rated as “Most Severe” as indicated in prior ratings of speech naturalness in a previous study by Benoit et al. (2008).

With the exception of the unmodified samples, the “No Anchor” group produced average ratings that were less severe compared the “Anchor” group for samples of altered bandwidth conditions as well as the 10 unmodified speech samples.

These findings suggest that use of speech samples manipulated acoustically to produce an illusion of increased nasality did not enhance training procedures for rating hypernasal speakers. Given that previous research has found success with the use of auditory anchors, it is reasonable to assume that the results of this study may be due the auditory anchors and rating samples themselves. Although these samples were created to coincide with the findings of Chen and Yiu (2002) and provide more consistent ratings by eliminating potential perceptual biases among listeners, they seem to have confused the listeners more than helped them rate more accurately. This confusion extended to the ratings of unmodified samples. Indeed, the confusion was so great as to produce a complete reversal of the ratings that were collected previously for these samples.

Although nasality ratings of natural speech may have sub-optimal inter-rater reliability among listeners, it appears that use of natural speech samples provides a more realistic reference point when developing auditory anchors of resonance compared to altered speech simply because there is more to resonance than first-formant bandwidth. Although manipulation of the samples followed acoustic theory of nasalization, it did not sound like natural speech. This could have contributed to listeners' confusion given that both acoustically altered and unmodified speech samples were included with the same rating scale.

There were limitations of this study that may have affected results. Four of the 15 sentences used as stimuli contained the nasal phoneme /n/. This nasal phoneme is produced with an open velopharyngeal port, versus oral sounds which are produced with a closed nasal cavity. While only present in a few of the sentences, these nasal sounds may have slightly increased listeners' perception of hypernasality than what was actually present.

Further investigation is needed to examine methods of improving perceptual judgment of nasality. Currently there are few published studies reporting the effects of auditory anchors on listener judgment. The fact that Lee et al. found differences in perceptual judgment with exposure to auditory anchors warrants further exploration in this area. All speech-language pathologists would benefit from improved training procedures for this professional practice given that results of these ratings contribute to clinical judgments and to clinical decision-making. While the principle findings of this study are negative, we can confidently rule out the training regimen presented in this

study as a means for improving speech-language pathologists' rating of hypernasal speech.

REFERENCES

Awan, S. N. & Lawson, L.L. (2009). The effect of anchor modality on the reliability of vocal severity ratings. *Journal of Voice, 23(3)* 341-352.

Benoit, K., Munson, B., Thurmes, A., Cordero, K.N., Baylis, A., & Moller, K. (2008). /Factors Affecting Speech Naturalness in Young Adults with a History of Cleft Palate <http://www.tc.umn.edu/%7Emunso005/BenoitEtAl_2008ASHA.pdf>./ Poster presented at the 2008 ASHA Convention, Chicago, IL, November 20-22.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Chan, M.K., & Yiu, E.M-L. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research, 45* 111-126.

Chen, M.Y. (1995). Acoustic correlates of nasalized vowels in hearing-impaired and normal-hearing speakers. *Journal of the Acoustical Society of America, 98(5)* 2443-2453.

De Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38,* 794-811.

Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36,* 14-20.

Henningsson, G., Kuehn, D.P., Sell, D., Sweeny, T., Trost-Cardamone, J.E., & Whitehill, T.L. (2008). Universal parameters for reporting speech outcomes in individuals with cleft palate. *Cleft Palate-Craniofacial Journal, 45(1)* 1-17.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal for Speech and Hearing Research, 36,* 21-40.

Kuehn, D.P., & Moller, K.T. (2000). Speech and language issues in the cleft palate population: the state of the art. *Cleft Palate Journal, 7,* 348.

Lee, A., Ciocca, V. & Whitehill, T.L. (2003). Acoustic correlates of hypernasality. *Clinical Linguistics & Phonetics, 17(4-5)* 259-264.

Lee, A., Whitehill, T.L., & Ciocca, V. (2009). Effect of listener training on perceptual judgment of hypernasality. *Clinical Linguistics & Phonetics, 23(5)* 319-334.

Mandulak, K.C., Zajac, D.J. (2009). Effects of altered fundamental frequency on nasalance during vowel production by adult speakers at targeted sound pressure levels. *Cleft Palate-Craniofacial Journal, 46(1)* 39-46.

Moll, K.L. (1964). 'Objective measure of nasality [letter to the editor]. *Cleft Palate-Craniofacial Journal, 1,* 371-374.

Moller, K.T. & Glaze, L.E. (2009). *Cleft lip and palate: Interdisciplinary issues and treatment.* Austin: PRO-ED, Inc.

Moller, K. T. & Starr, C.D. (1984). The effects of listening conditions on speech ratings obtained in a clinical setting. *Cleft Palate Journal, 21,* 65-69.

Yiu, E.M.L. & Ng, C.Y. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18(3)* 211-229.

APPENDIX A

Subject Demographic Information

Table 1. Demographics and background information for Anchor Group

Subject	Gender	Age	Amount of time spent around children under 5 (scale of 1-10)
800	F	23	3
801	F	20	3
802	F	20	1
803	F	24	1
804	F	23	5
805	M	25	3
806	F	24	4
807	F	24	10
808	F	27	3
809	F	25	5
810	F	26	3
811	F	25	2
812	F	21	2
813	F	23	5
814	F	23	9

Table 2. Demographics and background information for No Anchor Group

Subject	Gender	Age	Amount of time spent around children under 5 (scale of 1-10)
815	M	26	1
816	F	23	3
817	F	21	5
818	F	22	2
819	F	23	7
820	F	25	4
821	F	23	5
822	F	22	1
823	F	23	3
824	F	22	5
825	F	20	2
826	F	27	7
827	F	21	5
828	F	23	5
829	F	18	1

APPENDIX B

Sentence Stimuli

Lazy Jack Passage

Sentences

1. Buy baby a bib
2. Bob is a baby boy
3. Pete got a pipe to keep
4. A pea popped up
5. Sissy saw Sally race
6. See the sassy goose
7. The zebra was at the zoo
8. Keep the can of Coke
9. Ken can keep it cool
10. Gary got the egg a while ago
11. Time to eat at the table
12. Daddy, don't dive too deep.
13. The oddest duck did it.
14. She had to wash her shoes
15. Take a tiny bit of tea.

Lazy Jack Passage

Once upon a time there was a boy named Jack. He lived in a red house with a white roof. His mother worked hard each day feeding the pigs and the chickens or washing clothes in a big tub. But all Jack did was to play with his play with the squirrels, or sit in a chair by the stove and sleep. In the summer, Jack liked very much to drink cold milk or eat ice cream under a shady tree. Some days he would visit the zoo to see the animals and some days he would go to a large river to fish. But when winter came he often stayed home to play with his bicycle, his blocks, his toy soldiers, or his yellow ball. And sometimes he just sat watching the other children skate on the smooth ice. But if anyone mentioned work, nothing could make him leave his place before the warm fire. As usual, he thought only of his own pleasure.