Evaluating listener bias introduced by two speaker groups on VAS ratings: A methodological investigation

by

Rebecca S. Higgins

Undergraduate Honors Thesis
Department of Hearing and Speech Sciences
University of Maryland, College Park

Advisory Committee:
Dr. Jan Edwards, Chair
Dr. Margaritis Fourakis
Dr. Jared Novick

Graduate Co-Advisor:
Allison Johnson

Table of Contents

List of Figures

List of Tables

Acknowledgements

Throughout my undergraduate career, I have been blessed with an abundance of support and guidance. First, I want to thank my advisor, Dr. Jan Edwards for her flexible support, her genuine encouragement, and for her wonderful leadership over the lab. Thank you for making L2T the family it has been to me the past four years. To Allie Johnson, my graduate student co-advisor, you have been so patient, so selfless, and so understanding. Thank you for holding my hand through mixed effects models, but pushing me outside of my comfort zone so that I could succeed. I hope the students you someday teach know how lucky they are to have someone like you. I would like to thank my committee members as well. Dr. Fourakis, thank you for challenging me intellectually, both in the classroom and throughout this process. Dr. Novick, thank you for exploring a topic a little outside of your expertise and bringing a refreshing perspective to this project. I would also like to thank Max Scribner for being my coding hero this semester. Thank you for bringing my design to life. Zachary Maher, another coding hero, has provided me with all kinds of support the past two years (yes- I know box and whisker plots show medians). Thank you to Michele Liquori for helping out with the very mundane task of listening to all the selected tokens. Thank you to my fellow L2T Undergraduate students who helped run my study, and especially to Ileana Thompson for her support throughout testing. And finally, I want to thank my family, my friends, and especially Jakob for their steadfast support and for listening to me talk about my thesis way too often.

Abstract

Objectives: The goal of this experiment was to determine whether listeners rate /t/ and /k/ consonants produced by 3 to 5-year-old children with cochlear implants (CIs) differently when in a blocked condition that did not include productions by children with normal hearing (NH) compared to an unblocked condition which did include such productions. Blocking has been shown to influence results on categorical tasks. Some research suggests that judgments made using continuous a visual analog scale (VAS) are less susceptible to bias than judgments made using a categorical system. However, the research on VAS and bias is sparse, with no research on the interaction of VAS ratings and perceptual bias when bias is introduced to an experiment via a comparison of two groups. This study used word-initial /t/ and /k/ CV sequences produced by children with NH and children with CIs to investigate the effect of blocking on VAS judgments. **Design:** 48 adult participants were recruited at the University of Maryland. Each participant rated 500 CV tokens in a VAS experiment. Half of the participants were assigned to the blocked condition, and half to the unblocked condition. Mixed-effects models were used to analyze the ratings of the tokens produced by children with CIs to see if there was a significant change from the blocked to the unblocked condition. **Results:** For both the t-like and k-like tokens, models showed significant effects of intercept and transcription category. Ratings for production by children with CIs were not significantly different across the two conditions. **Conclusions:** The VAS ratings of tokens from children with CIs did not differ in the blocked and unblocked condition. This result supports the finding that VAS may be less susceptible to bias than categorical judgments. In future studies, researchers may choose blocked or unblocked designs to compare these two groups of speakers, depending on which design is better suited to answer individual research questions.

Introduction

Strong research rests on strong methodologies. When methodologies stumble, research does too. When researchers or speech-language pathologists need to describe speech – in order to describe children's speech sound development, or to determine if a particular child has a speech sound disorder – they usually rely on phonetic transcription. Transcription involves listening to speech and choosing a categorical symbol to represent a child's production of a speech sound. Although transcription is the standard method, this approach has shortcomings.

The process of transcription rests on the assumption that speech production is categorical. However, research has shown that productions can be intermediate between two phonemes (Gibbon, 1999). It has been argued that even though speech is produced in a continuous fashion, speech perception is categorical. Early behavioral research suggested that individuals did not have within-category representations of phonemes (Liberman, et al., 1957). Recent research has challenged this idea. For example, in Tuscano et al. (2010), two peaks in the event related potentials (ERPs) recorded as participants categorized words that varied on a VOT spectrum showed that fine-grained acoustic features were maintained through late phase perceptual processing. This research provides strong evidence that speech is encoded in a continuous fashion before it is categorized.

Nonetheless, a transcriber must pick a categorical symbol even if the production is intermediate. When this occurs, clinicians and researchers are excluding information carried by a rich acoustic signal, which can be problematic. For example, transcription is not well-suited to capture productions of covert contrasts. A covert contrast occurs when a listener would transcribe two sounds using the same phonemic symbol, even when acoustic measurements show a reliable difference between the two sounds (e.g. Gibbon et al., 1990; McAllister Byun et al.,

2016; Munson et al., 2012). Researchers believe that children who produce covert contrasts have underlying representations of the two phonemes, whereas children who produce no contrast do not. Covert contrasts are relevant to the treatment and prognosis of phonological disorders in children, because children who produce covert contrasts progress faster through therapy compared to those who produce no contrast (Tyler et al., 1993).

Another shortcoming of the transcription process is that listeners are biased. Research has shown that people interpret what they hear based on prior expectations. In a speech perception experiment, researchers manipulated a carrier phrase to bias listeners' perception of the speaker's age. When the carrier phrase was "a weawwy yike," listeners were more likely to categorize a speech token as incorrect compared to when the same token was presented in the carrier phrase "I really like" (Munson et al., 2010). In another speech perception experiment, listeners were more likely to categorize a speech token as incorrect if the word "misarticulation" was mentioned in the task instructions (Schellinger, Edwards, Munson, & Beckman, 2008).

There is also evidence to suggest that listeners' impressions can impact speech perception of adult speakers. In New Zealand, there is a current shift in the vowels in the word TRAP and DRESS. According to sociolinguistics, this shift had been led by young, female speakers. Drager (2011) showed pictures of people of different ages and sexes at the same time as words along the DRESS-TRAP continuum were presented. Even though the same words were played to each participant, changes in the speaker's perceived sex and age changed perception of the vowel. These studies give speech-language pathologists and others who rely on the accuracy of their perceptions reason to consider additional, more objective measures of speech production that are less susceptible to bias.

In light of the mismatch between a categorical system (transcription) and a continuous signal, as well as the presence of listener bias, better characterizations of speech production must be created to ensure efficient treatment of speech sound disorders and enhance the validity of published research findings. Two alternative methods of characterizing speech production are acoustic measures and visual analog scale (VAS) ratings. Acoustic waveforms and spectra reflect articulatory movements by displaying disturbances of air pressure and can be used to derive measures that more appropriately reflect the continuous nature of speech. However, acoustic analysis is a complicated and time-consuming process that requires researchers or clinicians to obtain high-quality sound recordings and determine which parameters to use from a long list of possible options. This type of analysis requires expertise and choosing the wrong features or parameter settings may misrepresent important aspects of the acoustic signal (Munson et al., 2012).

VAS may be a more practical measure than acoustic analysis for clinicians and clinical researchers. VAS ratings are straightforward and easy-to-collect, and they have been shown to correlate with acoustic measures. Several studies have shown that VAS ratings reliably depict the difference between a broad range of contrasts, including /s/ and /ʃ/ (Munson, Schelligner, Beckman, & Meyer, 2010; Munson, Johnson, & Edwards, 2012), /d/ and /g/ (Munson, Johnson, & Edwards, 2012), /s/ and /θ/ (Schellinger, Munson, & Edwards, 2017), and the presence of rhotic /ɹ/ (McAllister Byun et al., 2016). Additionally, VAS ratings have high intra-rater reliability, even among inexperienced listeners (Munson et al., 2012).

Less known is the effect that listener bias could have on VAS. Studies have shown that categorical judgments are sensitive to listener bias. For example, listeners were more likely to categorize a production along the /f/-/θ/ continuum as /f/ when the speaker was female compared

to when the speaker was male (Babel & McGuire, 2013). There is some research that suggests that judgments made using continuous VAS are less susceptible to bias than judgments made using a categorical system. Munson et al. (2017) tested the effect of two conditions that introduced bias to VAS judgments of children's /s/ and /θ/ productions. One condition involved simultaneously rating a continuous variable (gender typicality), and the other involved simultaneously rating a categorical variable (vowel identity). When using a continuous VAS, neither biasing condition changed the VAS judgments. However, when a categorical system was used to characterize the /s/ and /θ/ productions, judgments were influenced by biasing conditions.

This presents one important consideration for using VAS ratings: if listeners do not use the scale as truly continuous, acoustic detail is still missed. Unfortunately, there is evidence to suggest that listeners sometimes use a continuous scale in a more categorical fashion. In McAllister Byun et al. (2016), the majority of raters showed a bimodal response pattern when using a VAS, indicating they only used the endpoints of the scale rather. The degree of bimodality was variable between listeners. It is important to point out that in this study a binary task was mixed into the VAS ratings, which could have biased the rater into using the scale in a more categorical fashion (despite the evidence from Munson, 2017, showing that continuous VAS ratings are not influenced by simultaneously performing a categorical task). The discrepancy between these two studies could be attributed to differences in the administration of the experiments. In Munson et al. (2017), participants came into the lab to complete the VAS task. In McAllister Byun et al. (2016), participants were recruited on Amazon Mechanical Turk, a crowdsourcing platform where "turkers" are paid a few cents for each task they complete. Noise from the crowdsourced data may be playing a role in why listeners seem to be using the continuous scale more categorically.

If listeners are somehow influenced to use the endpoints of the scale rather than the entire continuum, the task becomes more categorical, and VAS ratings may become more susceptible to perceptual bias. This bias could be introduced in a number of ways. Participants could be misled by task instructions to believe that the task was more categorical than continuous. Or, one stimulus could influence a listener's perception of the next stimulus, thereby altering his or her use of the scale.

The possibility that listeners' perceptions could be influenced by stimuli within a VAS task is particularly problematic when researchers are looking for potential group differences. Currently, there is no research on whether listener bias is introduced by comparing the speech of different populations within a VAS experiment. It is possible that presenting speech from two groups would cause listeners to use VAS more categorically, with tokens from one group clustering around one endpoint, and tokens from the other group clustering on the other endpoint. On the other hand, if group membership does not introduce listener bias within a VAS experiment, ratings for each group would span the entire continuum. Also, even if listener bias is introduced by presenting two groups, VAS could be somewhat resistant to bias compared to a categorical task.

VAS could be used to compare the perception of speech of two groups of individuals. For example, VAS could be used to compare the productions of children with cochlear implants (CIs) and children with normal hearing (NH) to reveal whether their speech differs systematically. VAS could also be used to determine whether there are group differences in productions of covert contrasts, which could inform better approaches to speech therapy. However, without foundational knowledge about how stimuli from multiple groups in a VAS paradigm impact perceptual bias, results may be misleading.

This study is designed to examine how VAS impacts perceptual bias by examining VAS ratings for two groups of speakers under two presentation conditions. The two groups of speakers are children with cochlear implants and children with normal hearing. Children with CIs have less intelligible speech than children with NH (Peng, Spencer, & Tomblin, 2004). Even productions from children with CIs that are transcribed as "correct" have acoustic differences compared to children with NH of the same chronological age (Reidy et al., 2017; Todd, Edwards, & Litovsky, 2011). This is not surprising given that CIs transmit a degraded signal compared to the output of a healthy cochlea (Dorman et al, 2000). Also, children with CIs experience a period of auditory deprivation before implantation. Even when children are implanted as early as one year of age, their speech and language development is delayed relative to their peers with normal hearing (Ertmer & Goffman, 2011). Due differences in speech production accuracy and intelligibility, listeners may be more likely to rate productions from children with CIs as categorically worse than productions from children with NH, especially if they've recently heard productions from children with NH.

This recency effect may cause listeners to use the scale more categorically, rating each group's productions at the endpoints of the scale. Alternatively, if listeners hear productions from one group at a time, they may be more likely to rate productions within each group over a larger range, ultimately decreasing the effect of bias. Previous research has looked at how blocking impacts bias in categorical designs. In Johnson (1991), listeners identified fricatives produced by male and female speakers. In the speaker sex effect, listeners were more likely to identify a consonant as /s/ than /ʃ/ when it was spoken by a male. When the experiment was blocked by speaker, the speaker sex effect decreased. Results from Babel and McGuire (2013), corroborate this finding: when /f/ and /θ/ tokens were presented in a design blocked by speaker,

raters exhibited more sensitivity to the /f/ and /θ/ contrast. However, if continuous VAS ratings are not susceptible to listener bias, there may be no differences in ratings across groups when a blocked design is used compared to an unblocked design.

The proposed experiment will determine whether listeners rate consonant goodness of /t and /k/ productions from children with CIs differently when tokens are presented in a blocked condition (productions only from children with CI) compared to an unblocked condition (productions from children with CIs intermixed with productions from children with NH). If there is no difference between ratings in the blocked and unblocked conditions, this result would provide evidence that listeners are not inherently biased by grouping or that VAS could be more resistant to bias than other tasks. However, if there is a difference, additional result is needed to determine which set of ratings correspond more closely to objective measures of speech production, such as acoustic features or articulatory gestures. This research will help future studies obtain valid results by determining the best methodological approach for comparing groups using ratings from a VAS task.

Methods

Speakers

The speech tokens used as stimuli in this study were recorded during a larger longitudinal study at the University of Wisconsin-Madison and the University of Minnesota-Twin Cities. Stimuli were produced by 52 children, 26 children with cochlear implants and 26 children with normal hearing. A total of 40 sessions were recorded for each group, as some children returned for multiple visits over the course of 2-3 years. The groups were matched on maternal education, sex, and age across sessions. Table 1 displays the demographic information for the speakers.

Table 1: *Speaker demographics*.

Group	Age (Months)	Sex (Male:Female)	Maternal Education
CI n = 26	50 Range: 31-66	11:15	Some high school, high school diploma, or equivalent: 2 Some college/2 yr degree: 5 College or Graduate Degree: 19
NH n = 26	50 Range: 31-66	11:15	Some high school, high school diploma, or equivalent: 2 Some college/2 yr degree: 5 College or Graduate Degree: 19

Speaking Task

Stimuli were recorded using a picture-prompted, auditory word-repetition task. Words for this task were chosen if at least 80% of children within the selected age range knew the word according to the *MacArthur-Bates Communicative Development Inventories* norms (CDI; Fenson, et al., 2007), and the word could be easily represented by a picture. Three wordlists were used depending on the age of the child at each visit. These wordlists are shown in Appendix Table 9.

During the task, a picture was presented on a computer screen while a female speaker's production of the word (e.g., "cat") was presented over loudspeakers. The child repeated the word into a microphone, and digital recordings were saved. Some words were repeated in order to balance the tokens across high back, high front, low back, and low front vowel contexts. Approximately 32 /t/- and /k/-initial words were elicited from each child.

Stimuli

The stimuli for the current perception experiment were CV segments, consisting of a word-initial /t/ or /k/ consonant plus 150 ms of the following vowel, extracted from a real word.

The /t/-/k/ contrast was chosen because previous studies have shown that listeners can reliably rate productions along the /t/ and /k/ continuum using VAS (Arbisi-Kelm et al., 2010). Also, the /t/-/k/ contrast is differentiated by spectral cues that are heavily impacted by processing limitations of cochlear implants. For this study, tokens that were transcribed as correct productions (t, k), intermediate productions (t:k, k:t), and substitutions transcribed as t or k ($t, $k) were included as stimuli. All initial consonants had been transcribed for a previous study (Johnson, Bentley, Munson, & Edwards, 2019).

Stimuli were screened, edited, and extracted before they were pooled for selection. A script in *Praat* (Boersma & Weenink, 2018) automatically marked token onset at the burst and token offset at 150ms after the voice-onset time (VOT). Then, I listened to each token and looked at the spectrogram and formants to determine if the token was usable or needed to be edited.

Tokens transcribed as something other than a lingual stop and tokens with a VOT of less than 0.20 ms were excluded. Tokens with background noise, whispered phonemes, poor sound quality, devoiced vowels, and clipping were also excluded. In some cases, the offset needed to be adjusted. For example, if the vowel was shorter than 150ms, then the boundary line was moved to the end of the vowel so that no final consonant was present. Image 1 shows the user interface for selecting a token. Ultimately, the tokens ranged in length from 109 ms to 421 ms. Across all tokens, the average duration was 242.7 ms, and the average duration for each group, (NH k, NH t, CI k, and CI t), was within 4 ms.

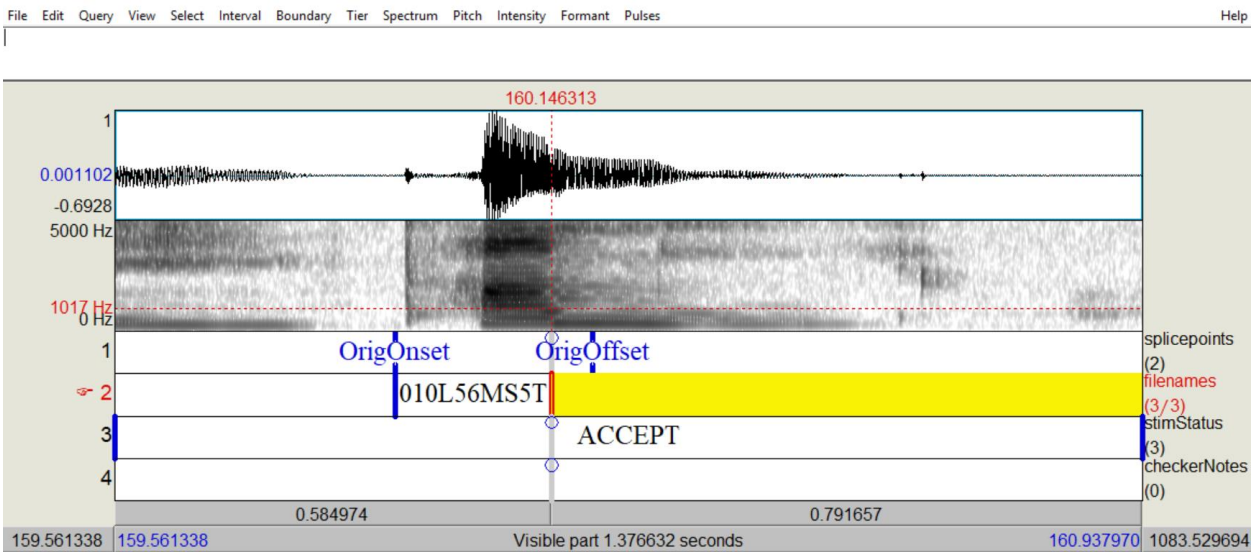

Figure 1. Praat selection script interface. Line 1 represents the boundaries created by the script and Line 2 represents the final boundaries. As shown by the red line, this boundary was moved to the end of the vowel.

After the initial selection, the tokens were extracted from Praat and saved as individual .wav files. Each file was amplitude normalized to an intensity of 70dB. Any token that had clipping after the normalization was excluded. There were a total of 653 usable tokens from speakers with CIs and 845 usable tokens from speakers with NH. A Masters student and I listened to the remaining tokens again, double-checking that all of the tokens were free from background noise. A few further tokens were discarded.

In order to select 125 t-like tokens and 125 k-like tokens from each group, between 3 and 5 tokens were randomly chosen from each speaker in both the NH and CI conditions. The files were then placed in an “Included t” or “Included k” folder depending on their respective population (CI/NH). Finally, these tokens were uploaded to a public folder in terp.connect, an online server to be accessed for the experiment. The final token lists consisted of 250 tokens from children with CIs and 250 tokens from children with NH.

Data Collection

Listeners

48 participants (33 females, 15 males) were recruited at the University of Maryland, College Park. The average age of the participants was 20 years old, with a range of 18 to 26 years old. All of the participants self-reported that their primary language was English and that their speech, language, and hearing was within normal limits or that they had no previous speech, language, or hearing diagnosis. Participants had varying degrees of exposure to children's speech, with some participants interacting with children very rarely while others interacted with children daily. Each listener was assigned to either the blocked or unblocked condition.

Listening Task

Listeners were asked to rate consonant goodness along a VAS, which reflected the perceptual auditory space is reflected on a continuous, visual line. In this study, the scale was designed to optimize precision and decrease the amount of bias present. In Matejka et al. (2016), different VAS designs were compared on a range of criteria. Banded, two labeled scales were associated with high levels of precision and low levels of bias. Therefore, the VAS for this study had a series of grey and black bands with two labels, 'very bad' on the left side and 'very good' on the right side. An example of the VAS for a /t/ token is shown in Figure 2.

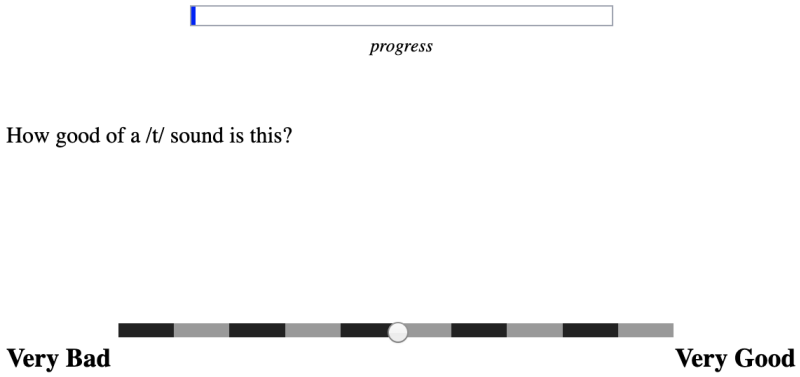

Figure 2. An example of the Visual Analog Scale. The banded, two-labeled scale used to rate consonant goodness.

For each token, participants clicked a location on the scale that represented their perceived goodness (or badness) of the production. Listeners heard productions from children with normal hearing and children with cochlear implants that were transcribed as /t/, /k/, or intermediate between /t/ and /k/. In one block, they heard only /t/-like productions, including correct [t] for target /t/, substitution of [t] for target /k/, and productions intermediate between /t/ and /k/ that were perceived as closer to /t/ ([t:k]). In another block, they heard only /k/-like productions, including correct [k] for target /k/, substitution of [k] for target /t/, and intermediate productions between /k/ and /t/ that were transcribed as closer to /k/ ([k:t]). A sample of blocks that a participant would receive in each condition are shown in Table 2.

A Latin Square design was used to assign participants to conditions. First, a listener was assigned to a condition (blocked or unblocked). Then, the listener was assigned to hear either /t/-like tokens or /k/-like tokens first. Listeners in the blocked condition were also assigned to hear either children with CIs first or children with NH first. All listeners heard alternating blocks of

/t/-like and /k/-like tokens. The 125 tokens within each block were always presented in random order.

Table 2. *An example of blocks of tokens presented in each condition.*

Blocked Condition	Unblocked Condition
• 125 NH /t/-like tokens • 125 NH /k/-like tokens • 125 CI /t/-like tokens • 125 CI /k/-like tokens	• 125 NH+CI /t/-like tokens • 125 NH+CI /k/-like tokens • 125 NH+CI /t/-like tokens • 125 NH+CI /k/-like tokens

The experiment was designed on Ibex Penn Controller. A UMD short link was created to access the experiment quickly. Participants completed the experiment in a quiet room in the Learning to Talk Lab in Taliaferro Hall. After providing consent, participants answered demographic questions including age, place of birth, language history, history of speech, language, or hearing services, and their familiarity with children's speech. They were verbally instructed to listen to the consonant-vowel segments and rate how good the 't' or 'k' sound was produced, ignoring the goodness of the vowel. Following the demographic questionnaire, participants were taken to an instructions page shown in Figure 3.

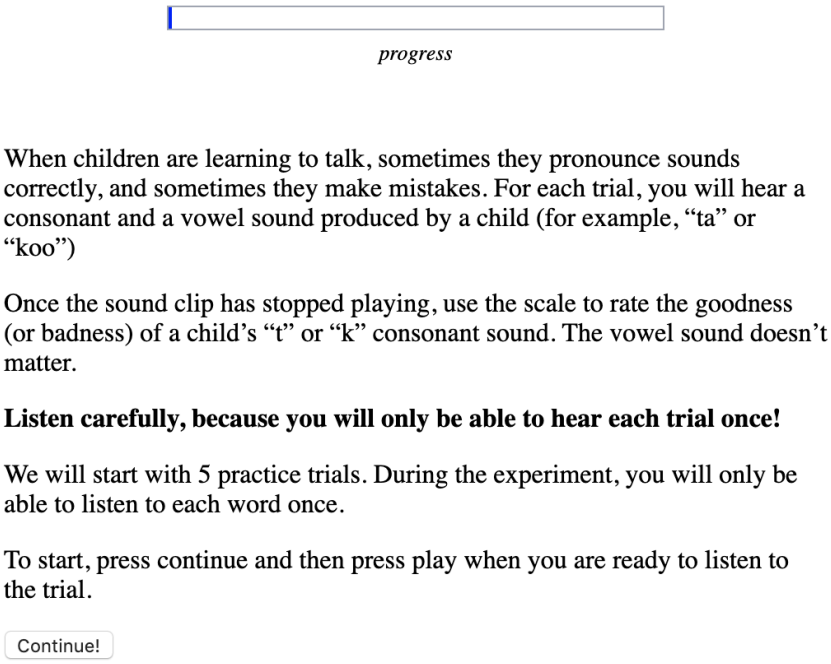

Figure 3. Listener Instructions.

All trials were completed on a MacBook Pro computer with Sennheiser HD 280 Pro headphones. Participants were asked to adjust the volume to a comfortable level during the five practice items and were encouraged to take breaks throughout the experiment. After completing the experiment, listeners were compensated financially for their participation. The entire task, including consent and compensation, took approximately 1 hour.

Analysis

In order to determine whether tokens produced by children with CIs were perceived differently when presented in a blocked condition compared to an unblocked condition, only ratings from the children with CIs were analyzed. Eight linear mixed-effects regression models were fit, four models for the t-like tokens and four models for the k-like tokens. These models were run using the *R* lme4 package (Bates et al. 2014, v. 1.1-21). All models predicted rating

based on the fixed-effects of condition (blocked vs. unblocked), transcription category (clearly produced /t/ or /k/ consonants vs. intermediate t:k or k:t productions), target consonant (/t/ or /k/), as well as all 2-way interactions and the 3-way interaction between the fixed-effects. In order to account for non-independence due to multiple ratings per speaker and multiple ratings per listener, the model included by-speaker and by-listener random intercepts. For both the t-like and k-like datasets, each model had a different reference category, which allowed comparisons within and across conditions, transcription categories, and target consonants. The *R* formula used for the model is shown in Equation 1. To account for running four regression tests on the same set of data, the alpha level to test for significance was adjusted to 0.0125.

Equation 1. Mixed-effects model formula coded in *R*.

$$Rating \sim Condition \times Transcription\ Category \times Target\ Consonant + (1 \mid SpeakerID) + (1 \mid ListenerID)$$

Predictions

Based on the previous research, I predicted that ratings of speech produced by children with cochlear implants would be lower in the unblocked condition compared to the blocked condition. Also, I predicted that there would be a significant effect of transcription category, with tokens transcribed as a clear consonant rated higher than intermediate productions. Finally, I predicted that there would be no difference in ratings of the /t/-like and /k/-like tokens.

Results

All 500 tokens (250 from each group, half /t/-like, and half /k/-like) were rated by each participant. The same 500 tokens were rated in both the blocked and unblocked conditions. Tokens were unevenly distributed across transcription categories. Table 3 shows the number of

tokens in each category within the 500 tokens chosen. The mean rating of tokens in the blocked condition was 59.64 (*SD* = 25.25) and in the unblocked condition was 58.59 (*SD* = 23.00). The mean rating of tokens produced by children with NH was 59.54 (*SD* = 24.00) and the mean rating of tokens produced by children with CIs was 58.69 (*SD* = 24.32).

Table 3. *Number of tokens per transcription category and target consonant for each group*.

	[k] for target /k/	[k] for target /t/	[k:t] for target /k/	[k:t] for target /t/	[t] for target /t/	[t] for target /k/	[t:k] for target /k/	[t:k] for target /t/
NH	103	0	22	0	109	6	6	4
CI	97	9	12	7	94	17	8	6

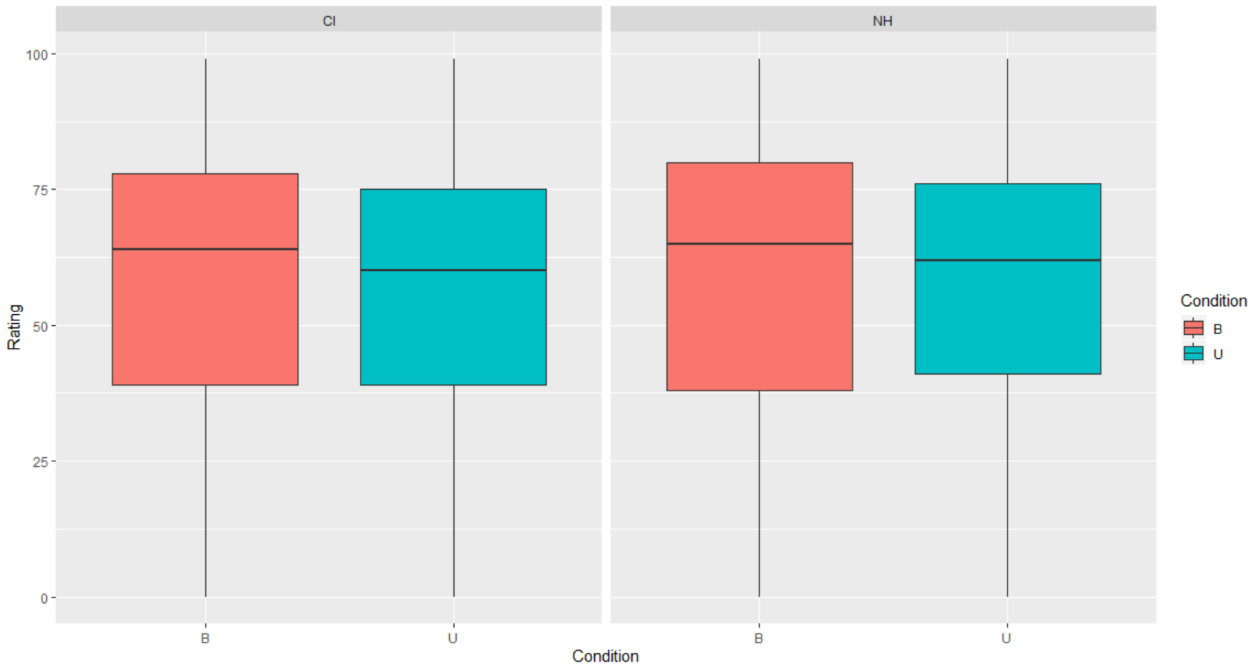

Figure 4. Ratings by condition and group. The black line shows the median for each condition and the box represents the interquartile range. The whiskers visualize the range of ratings in each

condition. Ratings from children with cochlear implants are in the left panel and children with normal hearing in the right panel. Red boxes are ratings in the blocked condition, while blue boxes are from the unblocked condition.

Ratings varied substantially across listeners. Some listeners were more likely to rate productions as bad, and others were more likely to rate productions as good. The mean rating by listener ranged from 42.13 (*SD* = 23.89) to 87.42 (*SD* = 15.57). Also, the mean rating differed by speaker, with some speakers more likely to have high ratings compared to others. The mean rating by speaker ranged from 43.51 (*SD* = 27.38) to 70.98 (*SD* = 20.97). This variability was built into the models though random intercepts by-listener and by-speaker. Ultimately, across listeners, the entire scale was used, with ratings ranging from 0-99.

Listeners tended to use more of the scale when rating tokens produced by children with CIs in the blocked condition (mean range = 93.2) compared to the unblocked condition (mean range = 90.0). Mean ratings of tokens from children with CIs varied across transcription categories. Tokens transcribed as [t] or [k] were rated higher (mean = 60.04, *SD* = 23.38) than tokens transcribed as intermediate [t:k] or [k:t] (mean = 49.81 , *SD* = 26.03). Ratings also varied based on the target consonant. In general, the ratings changed in a logical manner, with correct productions rated the highest, followed by substitutions, then intermediate productions. However, the [k] for target /t/ tokens were rated the highest, contrary to the pattern of previous research (e.g., Munson et al., 2010). Figure 5 shows the distribution of ratings by transcription category in the CI population.

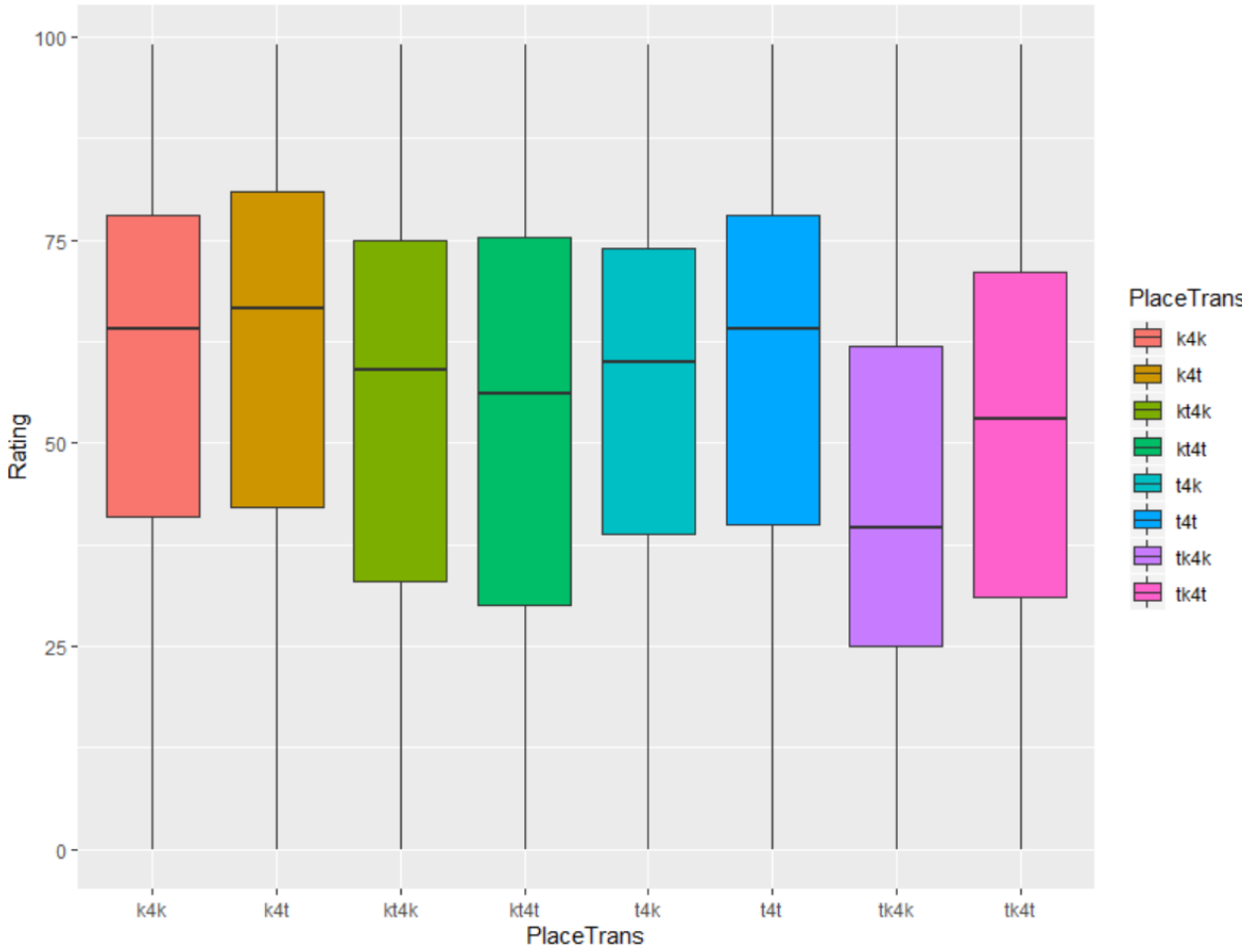

Figure 5. Rating by Transcription Category. Ratings by transcription category are displayed for the CI speaker group, with k-like tokens on the left and t-like tokens on the right.

Results for T-like Tokens

The tables with the results from models T1-T4 are in appendices 1-4, respectively. Model T1, which had a reference category of Blocked, Clear Consonant, and Target /t/, showed significant main effects of intercept ($\hat{\beta}_0$= 60.03, *SE* = 2.18, t =27.59, p< 0.001) and transcription category ($\hat{\beta}_0$ = -6.20, *SE* = 1.92, t =-3.23, *p* = 0.00124).

Model T2, which had a reference category of Blocked, Intermediate, and Target /k/, showed significant main effects of intercept ($\hat{\beta}_0$ = 47.99, *SE* = 2.68, t =17.91, *p*< 0.001) and transcription category ($\hat{\beta}_0$ = 11.59, *SE* = 2.07, t =5.59, *p*< 0.001). The main effect of target

consonant ($\hat{\beta}_0$ = 5.84, *SE* = 2.42, t =2.42, p = 0.0158) did not reach significance after the adjusted alpha-level was applied.

Model T3, which had a reference category of Unblocked, Intermediate, and Target /t/, showed significant main effects of intercept ($\hat{\beta}_0$ = 51.56, *SE* = 2.82, t =18.29, p< 0.001) and transcription category ($\hat{\beta}_0$ = 6.23, *SE* = 1.92, t =3.24, p = 0.00119). The main effect of target consonant ($\hat{\beta}_0$ = -4.96, *SE* = 2.42, t =2.42, p = 0.04) did not reach significance after the adjusted alpha-level was applied.

Model T4, which had a reference category of Unblocked, Clear Consonant, and Target /k/, showed significant main effects of intercept ($\hat{\beta}_0$ = 58.05, *SE* = 2.43, t =23.92, p< 0.001) and transcription category ($\hat{\beta}_0$ = -11.45, *SE* = 2.07, t =-3.23, p <0.001).

Overall, all T-like models show significant main effects of intercept and transcription category. The significant main effects of transcription category indicate that across conditions and target consonants, tokens transcribed as clear consonants were rated significantly higher than tokens transcribed as intermediate productions. The effect of condition was not significant in any model, suggesting that tokens were rated similarly across the blocked and unblocked condition. The non-effect remains consistent across all reference categories and is visualized in Figure 6. Although the effect of target consonant did not reach significance after the alpha level was adjusted, there was a trend in models T2 and T3 for intermediate productions to be rated higher when the target was /t/ compared to /k/.

Figure 6. Mean ratings of /t/-like tokens produced by children with CIs across conditions, transcription categories, and target consonants. Tokens presented in the blocked condition are in green, and tokens produced in the unblocked condition are in purple. Tokens with a target of /t/ are in the left panel and tokens with a target of /k/ are in the right panel. The tokens transcribed as a clear /t/ are on are on the left side within each panel, and tokens transcribed as intermediate closer to /t/ are on the right side. Light circles represent individual data points, dark circles represent the mean, and the lines show +/- 2 standard error from the mean.

Results for K-like Tokens

The tables with the results from models K1-K4 are in appendices 5-8, respectively. Model K1, which had a reference category of Blocked, Clear Consonant, and Target /t/, showed

significant main effects of intercept ($\hat{\beta}_0$= 62.18, *SE* = 2.80, t =22.20,, p< 0.001) and transcription category ($\hat{\beta}_0$ = -10.50, *SE* = 2.29, t =-4.59, p < 0.001).

Model K2, which had a reference category of Blocked, Intermediate, and Target /k/, showed significant main effects of intercept ($\hat{\beta}_0$ = 51.77, *SE* = 2.71, t =19.105, p< 0.001) and transcription category ($\hat{\beta}_0$ = 9.71, *SE* = 1.40, t =6.94, p< 0.001).

Model K3, which had a reference category of Unblocked, Intermediate, and Target /t/, showed significant main effects of intercept ($\hat{\beta}_0$ = 52.88, *SE* = 2.92, t =18.14, p< 0.001) and transcription category ($\hat{\beta}_0$ = 8.87, *SE* = 2.29, t =3.88, p < 0.001).

Model K4, which had a reference category of Unblocked, Clear Consonant, and Target /k/, showed significant main effects of intercept ($\hat{\beta}_0$= 59.63, *SE* = 2.41, t =24.74, p< 0.001) and transcription category ($\hat{\beta}_0$ = -9.15, *SE* = 1.40, t =-6.53, p <0.001).

Congruent to the T-like models, the K-like models show significant main effects of intercept and transcription category. The significant main effects of transcription category indicate that across conditions and target consonants, k-like tokens transcribed as clear consonants were rated significantly higher than tokens transcribed as intermediate productions. Like the T-like model, condition did not produce a significant result for any reference category, indicating that k-like tokens were rated similarly in the blocked versus the unblocked condition. Unlike the T-Models, there was no trend for intermediate productions to be rated higher when the target consonant was /t/ compared to /k/.

Figure 7. Mean ratings of /k/-like tokens produced by children with CIs across conditions, transcription categories, and target consonants. Tokens presented in the blocked condition are in green, and tokens produced in the unblocked condition are in purple. Tokens with a target of /k/ are in the left panel and tokens with a target of /t/ are in the right panel. The tokens transcribed as a clear /k/ are on are on the left side within each panel, and tokens transcribed as intermediate closer to /k/ are on the right side. Light circles represent individual data points, dark circles represent the mean, and the lines show +/- 2 standard error from the mean.

Discussion

The goal of this experiment was to determine whether listeners rated /t/ and /k/ consonants produced by 3 to 5-year-old children with CIs differently when listeners heard tokens from children with CIs and NH intermixed within a block (unblocked) compared to hearing each group separately (blocked). If listener bias impacts the way adult participants use a continuous scale when two groups are present, and group membership within an experiment introduces listener bias, ratings of speech produced by children with cochlear implants would be different in a blocked experiment compared to an unblocked experiment. Alternatively, if continuous scales are less susceptible to bias, as suggested by Munson et. al. (2017), then blocking would have no effect on ratings.

My hypothesis was that ratings of /t/ and /k/ produced by children with CIs would be lower when presented in an unblocked condition compared to a blocked condition was not borne out. Ratings of /t/ and /k/ consonants produced by children with CIs were not statistically different in the blocked and unblocked conditions. The current results conflict with previous studies that show that blocking can induce bias (Babel & McGuire, 2013; Johnson, 1991). It is possible that blocking induces bias when using categorical, forced-choice measures, but does not affect judgements made using a continuous scale (VAS). The difference between the previous studies and the current study is the use of a continuous scale (VAS) instead of a categorical, forced-choice task. Similarly to Munson, et al., 2017, a biasing condition had no effect when a continuous scale is used.

Although this study focused on how the ratings from the children with CIs changed in the presence of blocking, it is also interesting to look at the effect of blocking on the ratings of the children with NH. If blocking did impact VAS ratings, I would expect that the ratings of tokens produced by children with CIs would be lower and the ratings of tokens produced by children

with NH would be higher in the unblocked condition relative to the blocked condition. However, the mean ratings of both groups were slightly lower (see Figure 4) in the unblocked condition compared to the blocked condition.

Across both conditions and target consonants, the ratings of clear consonants were significantly higher than the ratings of intermediate consonants. Also, when considering both transcription category and target consonant together, the mean ratings generally followed the expected pattern, with ratings decreasing from correct productions to substitutions to intermediate productions. This result confirms that even untrained listeners are sensitive to subtle, within-category differences when asked to rate productions on a VAS.

There are some limitations of this study. First, because tokens were randomly selected and there were very few instances of [k] for target /t/ or [k:t] for target /t/ produced by children with NH, there were no instances of [k] for target /t/ or [k:t] for target /t/ selected from the children with NH. One unique result in this study was that [k] for target /t/ substitutions were rated the highest out of all the transcription categories. Previous VAS studies have found that listeners are able to differentiate between correct productions and substitutions- rating substitutions slightly lower than the correct productions (e.g., Munson et al., 2010). This inconsistency could be caused by many factors. The higher rating of [k] for target /t/ productions could be emerging solely from the inadequate number of examples. If more examples of [k] for /t/ productions were chosen, the mean rating might shift down closer to the true population mean. Since the speech of children with cochlear implants is less intelligible than the speech of children with normal hearing, and because the speech sound development of children with CIs is delayed and different than their peers with normal-hearing, there were many more intermediate productions from the children with CIs in total. However, if more intermediate productions from

the NH children were purposefully selected, the token set would not reflect the naturally occurring ratio of correct productions, substitutions, and intermediate productions of this group. As a result, this could over-represent productions that were rarely produced.

As mentioned above, there were no [k] for target /t/ productions from the children with NH included in the study. It could be the case that there are acoustic differences between the way children with NH and children with CIs produce [k] for target /t/ substitutions. Children with cochlear implants may be exhibiting no covert contrast at all. In that scenario, their [k] for target /t/ productions would sound identical to their [k] for target /k/ productions. Therefore, the target consonant could have a different effect on the goodness rating of a token.

One consideration of the study is whether group membership between children with cochlear implants and normal hearing is distinct enough to evoke bias. In other studies, listener bias occurs even when the listeners are not made aware of the group differences. For example, in Johnson (1991), speakers are not told that they will be listening to male and female speakers, yet the speaker sex effect still occurs. Since it is well documented that children with cochlear implants have less intelligible speech than peers with NH, if VAS ratings were susceptible to bias, and if listeners were comparing NH tokens they recently heard to tokens from children with CIs, it is logical to assume they would rate the productions from children with CIs as "worse" or less intelligible. However, the groupings in this study may have characteristics that don't produce bias. It could be the case that a CV sequence is not enough time for a listener to develop expectations about the speaker, so the differences between the speech of children with CIs and NH in CV sequences is not sufficiently large to create bias in the unblocked condition. Because this is the first experiment to investigate listener bias in VAS when comparing two groups, further research is needed to determine whether these findings generalize to other consonants and

groups. It is possible that another consonant contrast, longer speech segments, or different clinical populations could introduce listener bias within a VAS experiment.

Conclusion

The current study investigated whether blocking impacted ratings of /t/ and /k/ consonants from children with CIs. Specifically, the study considered whether the ratings of /t/ and /k/ consonants in CV sequences produced by children with cochlear implants changed when a blocked (only CI productions) or unblocked (mixed CI and NH productions) design was utilized. The results imply that blocking by group does not change the way adults rate /t/ and /k/ productions from children with CIs. The results from the current study provide researchers with data-driven evidence of how VAS might function when comparing two groups. My results also provide continuing support of the conclusion that VAS is less susceptible to bias than previously used categorical methods.

It is also important to note that phonetic transcription, a method often used by researchers and clinicians alike, is a categorical task and therefore, may be more sensitive to bias than VAS. Clinicians and researchers should be inspired to incorporate additional continuous measures to describe speech. VAS ratings provide a quick, straightforward, acoustically valid method to capture fine-grained detail of speech. This method allows clinicians to see the gradiency of speech sound development, including the presence of covert contrast, and can inform clinical decisions.

In the future, this study should be replicated with different consonant contrasts and different clinical populations. As mentioned in Munson et al. (2010), different contrasts may have different levels of susceptibility to bias. When comparing /s/ and /θ/, /θ/ was more susceptible to bias than /s/. Other contrasts, such as the /ɹ/-/w/ contrast- which is often a salient

cue to a speaker's age or articulatory skills- could be more susceptible to bias than /t/ or /k/. Finally, the results of this study should be used to design new experiments to explore the differences between /t/ and /k/ consonants produced by children with CIs compared to children with NH to determine if acoustic differences significantly impact perception.

References

Arbisi-Kelm, T., Edwards, J., Munson, B., & Kong, E.-J. (2010). Cross-linguistic perception of velar and alveolar obstruents: A perceptual and psychoacoustic study. Poster presentation at the Acoustical Society of America. Available in Journal of the Acoustical Society of America, 127, 1957.

Babel, M., & McGuire, G. (2013). Listener expectations and gender bias in nonsibilant fricative perception. *Phonetica, 70*(1-2), 117-51. doi:10.1159/000354644

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *lme4: linear mixed-effects models using Eigen and S4*. R package version 1.1-21.

Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer. [Computer program]. Version 6.0.30. http://www.praat.org/

Drager, K. (2011). Speaker age and vowel perception. *Language and Speech, 54*(1), 99-121.

Dorman, M., Loizou, P., Kemp, L., & Iler Kirk, A. (2000). Word recognition by children listening to speech processed into a small number of channels: Data from normal-hearing children and children with cochlear implants. *Ear and Hearing, 21*(6), 590-596.

Ertmer, D., & Goffman, L. (2011). Speech production accuracy and variability in young cochlear implant recipients: Comparisons with typically developing age-peers. Journal of Speech, Language, and Hearing Research: Jslhr, 54(1), 177-89.

Fenson, L., Marchman, V.A., Thal, D.J., Dale, P.S., Reznick, J.S., & Bates, E. (2007).

MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.). Baltimore, MD: Brookes.

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

Harel, D., Hitchcock, E. R., Szeredi, D., Ortiz, J., & McAllister Byun, T. (2017). Finding the experts in the crowd: Validity and reliability of crowdsourced measures of children's gradient speech contrasts. *Clinical Linguistics and Phonetics, 31*(1), 104–117.

Johnson, A., Munson, B., Bentley, D., and Edwards, J. (2019). What can speech acquisition data tell us about cochlear implant device limitations? Analyzing accuracy, error patterns, and spectral features of children's /t/ and /k/ productions. Poster [accepted to be] presented at the 2019 Conference on Implantable Auditory Prostheses. Lake Tahoe, CA, 14-19 July.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Language and Speech, 35*(3), 265.

Liberman, A., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Diego, CA, 7-12, May.

McAllister Byun, T., Harel, D., Halpin, P. F., & Szeredi, D. (2016). Deriving gradient measures of child speech from crowdsourced ratings. *Journal of Communication Disorders*, *64*, 91–102.

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

Munson, B., Johnson, J. M., & Edwards, J. (2012) The role of clinical experience in speech-language pathologists' perception of subphonemic detail in children's speech. *American Journal of Speech Language Pathology, 21,* 124-139.

Munson, B., Schellinger, S. K., Carlson, K. U., & Urberg-Carlson, K. (2012). Measuring speech-sound learning using visual analog scaling. *Perspectives on Language Learning and Education, 19*(1), 19–30.

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics and Phonetics, 31*(1), 56–79.

Peng, S., Spencer, L., & Tomblin, J. (2004). Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. Journal of Speech, Language, and Hearing Research: Jslhr,47(6), 1227-36.

Reidy, P., Kristensen, K., Winn, M., Litovsky, R., & Edwards, J. (2017). The acoustics of word-initial fricatives and their effect on world-level intelligibility in children with bilateral cochlear implants. *Ear and Hearing, 38*(1), 42-56.

Schellinger, S., Edwards, J., Munson, B., & Beckman, M. E. (2008). Assessment of phonetic skills in children 1: Transcription categories and listener expectations. Poster presented at the 2008 ASHA Convention, Chicago, 20-22, November 2008.

Todd, A. E., Edwards, J. R., & Litovsky, R. Y. (2011). Production of contrast between sibilant fricatives by children with cochlear implants. *Journal of the Acoustical Society of*

America, 130(6), 3969–3979.

Tuscano, J., McMurray, B., Dennhardt, J., & Luck, S. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *12*(10), 1532-1549.

Tyler, A. A., FigurskiI, G. R., & Langsdale, T. (1993). Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech and Hearing Research, 36*(4), 746–759.

Appendix

Appendix Table 1: Results from Model T1 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of blocked, clear consonant, and target t.

	Estimate	Std. Error	t-value	p-value
Intercept	60.03318	2.17572	27.592	<2e-16***
Condition	-2.24823	2.66535	-0.844	0.40320
Transcription Category	-6.20338	1.91989	-3.231	0.00124**
Target	-0.45321	1.28372	-0.353	0.72407
Condition x Transcription Category	-0.02261	2.53711	-0.009	0.99289
Condition x Target	0.71391	1.58801	0.450	0.65304
Transcription	-5.38195	2.77047	-1.943	0.05211

Category x Target				
Condition x Transcription Category x Target	0.16109	3.62084	0.044	0.96452

Appendix Table 2: Results from Model T2 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of blocked, intermediate, and target k.

	Estimate	Std. Error	t-value	p-value
Intercept	47.9946	2.6793	17.913	<2e-16***
Condition	-1.3958	3.3550	-0.416	0.6781
Transcription Category	11.5853	2.0711	5.594	2.33e-08***
Target	5.8352	2.4163	2.415	0.0158*
Condition x Transcription Category	-0.1385	2.5833	-0.054	0.9573
Condition x Target	-0.8750	3.2540	-0.269	0.7880
Transcription Category x Target	-5.3820	2.7705	-1.943	0.0521
Condition x Transcription Category x Target	0.1611	3.6208	0.044	0.9645

Appendix Table 3: Results from Model T3 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of unblocked, intermediate, and target t.

	Estimate	Std. Error	t-value	p-value
Intercept	51.55896	2.81876	18.291	<2e-16***
Condition	2.27083	3.57332	0.635	0.52606
Transcription Category	6.22599	1.91989	3.243	0.00119**
Target	-4.96016	2.41628	-2.053	0.04013*
Condition x Transcription Category	-0.02261	2.53711	-0.009	0.99289
Condition x Target	-0.87500	3.25404	-0.269	0.78802
Transcription Category x Target	5.22086	2.77047	1.884	0.05955
Condition x Transcription Category x Target	0.16109	3.62084	0.044	0.96452

Appendix Table 4: Results from Model T4 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of unblocked, clear consonant, and target k.

	Estimate	Std. Error	t-value	p-value
Intercept	58.0457	2.4268	23.918	<2e-16***
Condition	1.5343	2.9755	0.516	0.6076
Transcription Category	-11.4469	2.0711	-5.527	3.41e-08***
Target	-0.2607	1.2837	-0.203	0.8391
Condition x Transcription Category	-0.1385	2.5833	-0.054	0.9573
Condition x Target	0.7139	1.5880	0.450	0.6530
Transcription Category x Target	5.2209	2.7705	1.884	0.0596
Condition x Transcription Category x Target	0.1611	3.6208	0.044	0.9645

Appendix Table 5: Results from K1 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of blocked, clear consonant, and target t.

	Estimate	Std. Error	t-value	p-value
Intercept	62.1836	2.8008	22.202	< 2e-16***
Condition	-0.4352	3.7436	-0.116	0.908
Transcription Category	-10.5029	2.2865	-4.593	4.45e-06***
Target	-0.7014	1.5606	-0.449	0.653
Condition x Transcription Category	1.6316	3.0683	0.532	0.595
Condition x Target	-1.4175	2.1216	-0.668	0.504
Transcription Category x Target	0.7907	2.6118	0.303	0.762
Condition x Transcription Category x Target	-1.0671	3.5897	-0.297	0.766

Appendix Table 6: Results from K2 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of blocked, intermediate, and target k.

	Estimate	Std. Error	t-value	p-value
Intercept	51.76991	2.70971	19.105	<2e-16***
Condition	-1.28819	3.60345	-0.357	0.722
Transcription Category	9.71228	1.40057	6.935	4.52e-12***
Target	-0.08922	2.10136	-0.042	0.966
Condition x Transcription Category	-0.56447	1.86314	-0.303	0.762
Condition x Target	2.48462	2.89564	0.858	0.391
Transcription Category x Target	0.79067	2.61182	0.303	0.762
Condition x Transcription Category x Target	-1.06714	3.58967	-0.297	0.766

Appendix Table 7: Results from Model K3 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of unblocked, intermediate, and target t.

	Estimate	Std. Error	t-value	p-value
Intercept	52.8771	2.9151	18.139	<2e-16***
Condition	-1.1964	3.8976	-0.307	0.759496
Transcription Category	8.8713	2.2865	3.880	0.000106***
Target	-2.3954	2.1014	-1.140	0.254361
Condition x Transcription Category	1.6316	3.0683	0.532	0.594909
Condition x Target	2.4846	2.8956	0.858	0.390897
Transcription Category x Target	0.2765	2.6118	0.106	0.915699
Condition x Transcription Category x Target	-1.0671	3.5897	-0.297	0.766262

Appendix Table 8: Results from Model K4 predicting Rating based on Condition, Transcription Category, and Target Consonant with a reference category of unblocked, clear consonant, and target k.

	Estimate	Std. Error	t-value	p-value
Intercept	59.6295	2.4098	24.744	<2e-16***
Condition	1.8527	3.2059	0.578	0.566
Transcription Category	-9.1478	1.4006	-6.531	7.07e-11***
Target	2.1189	1.5606	1.358	0.175
Condition x Transcription Category	-0.5645	1.8631	-0.303	0.762
Condition x Target	-1.4175	2.1216	-0.668	0.504
Transcription Category x Target	0.2765	2.6118	0.106	0.916
Condition x Transcription Category x Target	-1.0671	3.5897	-0.297	0.766

Appendix Table 9: Wordlists used in the speaker task.

Word	IPA transcription	Vowel Context
Cake	/keɪ̯k/	Front
Candle	/kændl̩/	Front
Candy	/kændi/	Front
Car	/kɑɹ/	Back
Cat	/kæt/	Front
Catch	/kætʃ̯/	Front
Coat	/koʊ̯t/	Back
Coffee	/kɑfi/	Back
Comb	/koʊ̯m/	Back
Cookie	/kʊki/	Back
Cousin	/kʌzɪn/	Back
Cup	/kʌp/	Back
Cutting	/kʌɾɪŋ/	Back

Keys	/kiz/	Front
Kitchen	/kɪt͡ʃɪn/	Front
Kitten	/kɪtn̩/	Front
Kitty	/kɪɾi/	Front
Table	/te͡ɪbl̩/	Front
Take	/te͡ɪk/	Front
Tape	/te͡ɪp/	Front
Teacher	/tit͡ʃɹ̩/	Front
Teddy bear	/tɛdibe͡ɪɹ/	Front
Tent	/tɛnt/	Front
Tickle	/tɪkl̩/	Front
Tiger	/ta͡ɪgɹ̩/	Back
Toast	/to͡ʊst/	Back
Toaster	/to͡ʊstɹ̩/	Back
Tongue	/tʌŋ/	Back

Tooth	/tuθ/	Back
Toothbrush	/tuθbɹʌʃ/	Back
Towel	/taʊl̩/	Back
Tummy	/tʌmi/	Back