Published in final edited form as:

J Commun Disord. 2016 ; 64: 91–102. doi:10.1016/j.jcomdis.2016.07.001.

Deriving gradient measures of child speech from crowdsourced ratings

Tara McAllister Byun, Daphna Harel, Peter F. Halpin, and Daniel Szeredi

New York University, New York, NY

Abstract

Recent research has demonstrated that perceptual ratings aggregated across multiple non-expert listeners can reveal gradient degrees of covert contrast between target and error sounds that listeners might transcribe identically. Aggregated ratings have been found to correlate strongly with acoustic gold standard measures both when individual raters use a continuous rating scale such as visual analog scaling (Munson, Johnson, & Edwards, 2012) and when individual raters provide binary ratings (McAllister Byun, Halpin, & Szeredi, 2015). In light of evidence that inexperienced listeners use continuous scales less consistently than experienced listeners, this study investigated the relative merits of binary versus continuous rating scales when aggregating responses over large numbers of naive listeners recruited through online crowdsourcing. Stimuli were words produced by children in treatment for misarticulation of North American English /r/. Each listener rated the same 40 tokens two times: once using Visual Analog Scaling (VAS) and once using a binary rating scale. The gradient rhoticity of each item was then estimated using (a) VAS click location, averaged across raters; (b) the proportion of raters who assigned the “correct /r/” label to each item in the binary rating task ($\hat{p}$). First, we validate these two measures of rhoticity against each other and against an acoustic gold standard. Second, we explore the range of variability in individual response patterns that underlie these group-level data. Third, we integrate statistical, theoretical, and practical considerations to offer guidelines for determining which measure to use in a given situation.

Keywords

crowdsourcing; speech rating; research methods; speech sound disorders; speech perception; visual analog scaling; covert contrast

1. Introduction

1.1 Gradience in the acquisition of phonemic contrasts

Both typically-developing children and children with phonological delay or disorder produce speech patterns that differ systematically from adult inputs. These patterns are often described in terms of substitution of one phoneme for another. For instance, in a child with the phonological pattern termed *stopping*, the words “sea” and “tea” may be transcribed

identically ([ti]). However, adults may be biased by their own phonological knowledge in transcription tasks, favoring categories familiar from their own fully developed phonology (Amorosa, Benda, Wagner, & Keck, 1985; Buckingham & Yule, 1987). Given the anatomical, motoric, and cognitive-linguistic differences between children and adults, these transcriptions may actually stand at some distance from phonetic reality. The term *covert contrast* is used to describe cases where a typical listener would transcribe a categorical error, such as substitution of one phoneme for another, but instrumental measurements reveal a reliable phonetic difference. In one well-known illustration, children who are perceived to neutralize the voiced-voiceless contrast in initial position can be seen to maintain a statistically reliable phonetic distinction in voice onset time (VOT) between voiced and voiceless targets (Hitchcock & Koenig, 2013; Macken & Barton, 1980; Maxwell & Weismer, 1982). The phenomenon of covert contrast in child speech has now been amply documented over more than three decades of research (e.g. Edwards, Gibbon, & Fourakis, 1997; Gibbon, 1999; Munson, Johnson, & Edwards, 2012a; Munson, Schellinger, & Urberg Carlson, 2012b; Richtsmeier, 2010; Scobbie, 1998; Tyler, Edwards, & Saxman, 1990; Tyler, Figurski, & Langsdale, 1993; Weismer, 1984; Young & Gilbert, 1988). Several accounts have suggested that covert contrast could represent a typical or even universal stage in the normal course of acquiring a phonological contrast (Munson et al., 2012a, 2012b; Richtsmeier, 2010; Scobbie, 1998).

Studies of covert contrast in child speech have resulted in a major shift in how speech development is conceptualized. The traditional view, in which acquisition of a new phonemic contrast was framed as an abrupt, categorical change, has given way to a new consensus that speech sound development is phonetically gradient (Hewlett & Waters, 2004; Li, Edwards, & Beckman, 2009). Research on covert contrast also has significance for clinical applications. First, the presence or absence of covert contrast is often interpreted as evidence about the level of processing at which a child's speech errors apply: categorical substitutions imply a grammatical process, whereas covert contrast is often interpreted as evidence that errors are occurring at a more peripheral or articulatory level (Gibbon, 1999; but see McAllister Byun, Richtsmeier, & Maas, 2013). Second, the presence of covert contrast can be interpreted as a positive prognostic indicator for subsequent stages of a child's speech development. Children who maintain a covert contrast between two sound categories are more likely to progress to an overt contrast without receiving formal intervention (Forrest, Weismer, Hodge, Dinnsen, & Elbert, 1990). Once enrolled in treatment, children who show covert contrast are reported to reach the criterion for dismissal in a smaller number of treatment sessions (Tyler et al., 1990).

1.2 Measuring continuous speech development with listener ratings

Despite the well-documented importance of covert contrast in understanding a child's phonological system, it remains rare in clinical practice and some areas of clinical research to collect the instrumental measures that could reveal gradient differences between sound categories. This largely reflects the nontrivial difficulty associated with obtaining and verifying acoustic measurements of child speech. A further deterrent is posed by the inconclusive nature of negative results in a study seeking acoustic evidence of covert contrast in child speech. A phonemic contrast can be realized in any of a large number of

acoustic dimensions, and young children may mark contrast in phonetically unusual ways (Scobbie, Gibbon, Hardcastle, & Fletcher, 2000). If a given study reports no evidence of contrast in the limited set of acoustic parameters that the investigator chose to quantify, it remains possible that a difference was present in another domain that was not measured. Unlike acoustic measures, listeners' perceptual ratings can be collected in a relatively efficient fashion, and the human ear is simultaneously sensitive to numerous acoustic dimensions. These factors have led to a literature investigating whether gradient measures of child speech can be obtained from listener ratings.

Traditional descriptions of human speech perception have emphasized its discontinuous, categorical nature (Liberman, Harris, Hoffman, & Griffith, 1957). From this point of view, listeners tend to overlook covert contrast in child speech because they are perceptually predisposed to ignore phonetic differences within phonemic categories. However, subsequent research has amply demonstrated that within-category discrimination is possible (Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Pisoni & Tash, 1974; Werker & Logan, 1985). These studies emphasize that listeners' responses can be manipulated to be more continuous or more categorical depending on the nature of the task used to elicit perceptual judgments (Pisoni & Tash, 1974; Werker & Logan, 1985). A task that has been successfully used to elicit continuous perceptual ratings is visual analog scaling (VAS). In VAS, the endpoints of a line are defined to represent two different speech categories, and listeners mark any location to indicate the gradient degree to which they perceive a speech token as belonging to one category or the other (Massaro & Cohen, 1983; Munson et al., 2012b). A recent body of work by Munson and colleagues (Julien & Munson, 2012; Munson et al., 2012a, 2012b) has documented the relationship between VAS ratings and continuous acoustic measures of child speech. In a task of rating children's sibilants on a continuum from /s/ to /ʃ/, Julien & Munson (2012) found that mean VAS click location was significantly correlated with a relevant acoustic measure, centroid frequency, for 22 out of 22 listeners. In a multiple regression examining mean VAS click location as a function of both centroid frequency and fricative duration, the acoustic measures accounted for 53% of variance in mean VAS click location. Munson, Johnson, & Edwards (2012a) examined mean VAS click location as a function of acoustic measures for three contrasts in child speech: /s/-/θ/, /d/-/g/, and /t/-/k/. The correlations were significant for the first two contrasts, but not for the /t/-/k/ contrast.

The use of a continuous rating scale, as in VAS, is not the only option to derive continuous measures of child speech from perceptual ratings. Binary responses aggregated over multiple listeners can also be used to derive continuous or near-continuous measures. Perhaps the simplest metric is $\hat{p}$, which estimates the proportion of listeners who identify a token as belonging to a particular category (Ipeirotis, Provost, Sheng, & Wang, 2014). McAllister Byun, Halpin, & Szeredi (2015) reported that the proportion of listeners classifying children's productions as a "correct /r/ sound" was well-correlated with an acoustic measure of rhoticity, F3-F2 distance; this held true in both a sample of 25 experienced listeners ($r = -.77, p < .001$) and a sample of over 150 naive listeners ($r = .79, p < .001$). It is also possible to use Item Response Theory (IRT) (e.g., Baylor et al., 2011) to derive a more sophisticated model accounting for the fact that not all listeners are equally reliable in their perception, and some tokens may be unclear even to highly trained listeners, but McAllister Byun,

Halpin, & Harel (2015) found that this produced limited gains in explanatory power relative to the simpler $\hat{p}$ measure.

In sum, the literature to date suggests that either binary or VAS ratings collected from multiple raters can be used to derive gradient measures of child speech that would allow the identification of covert contrasts in child speech. Which approach, then, should clinicians and researchers favor? Gradient measures derived from binary ratings yield comparatively coarse-grained data, with the degree of granularity depending on the number of listeners rating each token. (For example, a sample of 9 listeners providing binary ratings yields 10 possible values of $\hat{p}$, i.e. between zero and nine raters classifying the token as “correct.”) In VAS, the continuous nature of the response scale opens the possibility that finer-grained differences could be detected. However, if raters are not able to perceive contrasts at a fine-grained level, they may use the VAS scale less consistently than the simpler binary category judgment. Some previous research has suggested that trained or experienced listeners may show greater sensitivity to within-category distinctions than naive listeners (Klein, McAllister Byun, Davidson, & Grigos, 2013; Wolfe, Martin, Borton, & Youngblood, 2003), although other studies have found no effect (Schellinger, Edwards, Munson, & Beckman, 2008) or more categorical response patterns in experienced listeners (Mayo, Gibbon, & Clark, 2013). Munson et al. (2012a) found that experienced listeners (speech-language pathologists) demonstrated greater intrarater reliability than naive listeners (college students) in a VAS task featuring children’s speech sounds. Thus, consistency in using the VAS scale may be of particular concern when raters lack previous training or experience in speech rating tasks. This is an important consideration for present purposes, since this study is part of a larger body of research investigating the utility of crowdsourcing, which supplies mainly untrained listeners, as a means to facilitate the collection of speech rating data.

1.3 Crowdsourcing ratings of speech

Recent technological innovations in the area of online crowdsourcing have provided an efficient means of collecting response data from large pools of individuals. Crowdsourcing platforms create an economy where anyone can post an electronic task or sign up to complete tasks for pay. This paper focuses on the Amazon Mechanical Turk (AMT) crowdsourcing platform, which has been the most widely used and studied by academic researchers to date. While AMT was developed for commercial use, it has been exploited with considerable success as a subject pool for behavioral research in linguistics (e.g. Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011) and psychology (e.g. Goodman, Cryder, & Cheema, 2013; Paolacci, Chandler, & Ipeirotis, 2010). The ease and speed of crowdsourced data collection has been characterized as “revolutionary” (Crump, McDonnell, & Gureckis, 2013); for example, Sprouse (2011) reported that a study that required 88 experimenter hours in a lab setting was replicated in 2 hours on AMT, with equivalent results.

Of course, increasing the speed of data collection is advisable only if it does not compromise data quality, and questions have been raised about the validity of results obtained through AMT and other crowdsourcing platforms (Chandler, Mueller, & Paolacci, 2014). When using online data collection, the researcher has limited control over the properties of the

experimental setup, such as the level of background noise and the quality of audio equipment. It is also more difficult to verify that participants are accurately reporting language/dialect background and hearing status. Crowdsourcing researchers do not dispute that online data collection is typically “noisier” than lab-based data collection. Instead, they argue that it is possible to overcome this inherent noise by recruiting a much larger sample than would be available through conventional channels. This assertion is supported by both computational modeling studies (Ipeirotis et al., 2014) and experimental studies validating results obtained through AMT against results collected in a typical laboratory setting (Crump et al., 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci et al., 2010). McAllister Byun et al. (2015) investigated the validity of crowdsourced listeners’ ratings in a task of assigning binary (correct/incorrect) ratings to /r/ sounds produced by child speakers of English. Ratings were aggregated across 153 naive listeners recruited through AMT and 25 experienced listeners, primarily speech-language pathologists. The modal rating across naive listeners was in agreement with the modal rating across experienced listeners for all but 7 out of 100 tokens, suggesting that crowdsourced listener judgments can represent a valid method for binary categorization of child speech productions. However, McAllister Byun et al. (2015) focused on the validity of binary ratings, with limited consideration of crowdsourcing as a means to obtain gradient characterizations of child speech data.

In the present study, crowdsourced listeners rated 40 child productions of North American English /r/ using either VAS or a binary rating scale. Each listener then re-rated the same 40 tokens, randomly ordered, using the other method. The gradient prototypicality or rhoticity of each item was then estimated in two ways: (a) VAS click location, averaged across raters; (b) the proportion of raters who assigned the “correct /r/” label to each item in the binary rating task ($\hat{p}$). This paper offers a three-part exploration of the relative validity and efficiency of these two methods of obtaining gradient measures of speech through crowdsourcing. First, we validate the two speech measures, mean VAS click location and $\hat{p}$, against each other and against an acoustic gold standard. Second, we explore the range of variability in individual response patterns that underlie these group-level data. Third, we integrate statistical, theoretical, and practical considerations to offer guidelines for selecting which measure to use in a given situation.

2. Method

2.1 Stimuli

This study collected ratings representing different levels of accuracy in the production of a single phoneme, the North American English rhotic /r/. This target was selected because it is late-emerging in the course of typical development (Smit, Hand, Freilinger, Bernthal, & Bird, 1990). In a small subset of speakers, derhotacized production persists through late childhood or adolescence (Ruscello, 1995), continuing into adulthood for 1–2% of speakers (Culton, 1986). Listeners often transcribe children’s derhotactized /r/ productions with the phoneme /w/, but both acoustic measures (P. J. Flipsen, Shriberg, Weismer, Karlsson, & McSweeney, 2001; Lee, Potamianos, & Narayanan, 1999; McGowan, Nittrouer, & Manning, 2004; Shriberg, Flipsen, Karlsson, & McSweeney, 2001) and trained listener judgments (Klein, Grigos, McAllister Byun, & Davidson, 2012) reveal a more nuanced continuum

between fully derhotacized and fully rhotic production. English /r/ is distinguished from other sonorant phonemes by the low height of the third formant, F3 (Delattre & Freeman, 1968; Hagiwara, 1995). The second formant, F2, is relatively high, bringing F2 and F3 particularly close together (Boyce & Espy-Wilson, 1997; Dalston, 1975). Research by Shriberg, Flipsen, Karlsson, & McSweeney (2001) and Flipsen, Shriberg, Weismer, Karlsson, & McSweeney (2001) reported that the distance between F2 and F3 provided a more sensitive characterization of the degree of rhoticity in adolescents' productions of /r/ than raw F3 height. Accordingly, F3-F2 distance will be adopted as the "gold standard" acoustic measure of degree of rhoticity for the purposes of the present study.

In the present study, listeners rated 40 words containing a target /r/ sound. All items were elicited from 12 children at varying stages in the process of remediation of /r/ misarticulation. These children ranged in age from 82 months to 188 months, with a mean of 118.2 months (SD = 28.1 months). Three of the 12 children were female, consistent with previous research indicating that speech errors affecting /r/ are more prevalent in male than female speakers (Shriberg, Paul, & Flipsen, 2009). The number of word tokens per child ranged from 1 to 6, with a mean of 3.3 and a standard deviation of 1.6. The 40-item word list included 10 items representing each of four allophonic variants: syllabic [ɹ] as in *sir*, [ɹ] as a singleton onset as in *red*, [ɹ] as part of an onset cluster as in *grow*, and [ə] in postvocalic position in words like *care* and *fear*, where it has been characterized as the vocalic offglide of a rhotic diphthong (e.g., McGowan et al., 2004). Items were hand-selected so that each category featured varying levels of acoustically determined accuracy (F3-F2 distance). Across the full set of 40 tokens, F3-F2 values were roughly normally distributed (Shapiro-Wilk's $W = 0.98$, $p = 0.79$).

Stimuli were a subset of 100 items that were rated by both naive and expert listeners in a previous study of crowdsourcing for speech ratings (McAllister Byun et al., 2015). The acoustic properties of these items were also measured in connection with that study. Measurements were carried out using Praat acoustic software (Boersma & Weenink, 2015). For each token, a trained graduate student visually inspected the automated LPC formant tracking function in Praat, adjusted filter order as needed, and identified the minimum value of F3 that did not represent an outlier relative to surrounding points. The first three formants were extracted at this point using Burg's method of calculating LPC coefficients. Ten percent of files were re-measured by a second student rater, yielding an intraclass correlation coefficient of .81. For additional detail on the protocol used to obtain and verify acoustic measurements, see McAllister Byun & Hitchcock (2012).

2.2 Task

The task, which was advertised on AMT under the heading "Rate children's /r/ sounds," was implemented in the Javascript-based experiment presentation platform Experigen (Becker & Levine, 2010). Listeners rated the 40 stimulus words described above. Each item was rated two separate times by each rater, once using VAS and once using binary rating. Items were blocked by rating method and then randomized within each block. The order of binary and VAS methods was counterbalanced across participants. Listeners were informed that the use of headphones was mandatory throughout the task, and they were required to report the

brand of headphones they were using; however, no additional tests of the listening equipment and/or environment were conducted. Instead, raters were required to score above a predetermined threshold on a set of attentional “catch trials,” described in more detail below.

Listeners were informed that they would hear words collected from children of different ages, and their task was to rate the /r/ sound in each word. “Correct /r/” and “Not a correct /r/” were presented as the choices in the binary rating task; these terms also served as the endpoints in the VAS task. These terms were used instead of a phoneme pair such as /r/ and /w/ because children’s misarticulated /r/ sounds seldom have phonetic properties consistent with true /w/, which could skew listener responses on the “incorrect” end of the continuum. Instructions for the VAS task were adapted from Julien & Munson (2012), and instructions for binary rating were developed to parallel the VAS instructions as closely as possible. In both tasks, listeners were told “We don’t have any specific instructions for what to listen for when making these ratings. We want you to go with your gut feeling about what you hear.” Five practice trials were presented before each block, to familiarize raters with the interface, but no feedback on accuracy was provided. Complete instructions are reported in the appendix to this paper, along with figures representing the visual analog scale and the binary response buttons. VAS responses were logged in a two-step process: participants clicked on the scale to position the indicator and then clicked a second button to register their response and move on to the next item. If participants clicked on the “record response” button without placing the indicator, they received a warning message.

2.3 Data collection

A total of $n = 726$ workers were recruited via Amazon Mechanical Turk to complete the listening task described above. Raters were required to be self-reported native speakers of American English who had no history of significant speech or hearing impairment. In addition, only responses originating from US IP addresses were considered for inclusion. To identify listeners who were inattentive or unreliable, 18 attentional “catch trials” were randomly interspersed with experimental trials, with half in each rating method block. These items had a predetermined correct response, having been judged by 3 experienced listeners to contain unambiguously correct or incorrect /r/ sounds. Listeners were credited with a correct response if they gave the same answer as the experienced raters’ consensus response in the binary rating task, or if their click location fell in the upper or lower half of the VAS scale in accordance with the correct/incorrect judgment assigned by experienced listeners. Participants were required to demonstrate above-chance accuracy (minimum 14/18 correct) across catch trials.

Data collection was completed in 21.4 hours and cost \$722, including Amazon fees. Results were discarded from 287 participants who did not meet all demographic criteria, 136 participants who did not exceed chance-level performance on attentional catch trials, and 12 participants who had missing or otherwise unusable data, thus resulting in a final group of 291 included participants. The included group of participants had a mean age of 32.4 years, with a standard deviation of 9.8 years.

2.4 Statistical Analyses

To address our first research question, the validity of estimates obtained from the VAS and binary data collection tasks, we compared estimates of /r/ prototypicality for each token obtained using VAS (mean click location across all responses from included raters) versus binary ratings ($\hat{p}$, the proportion of individuals who rated a given item as a “correct /r/”). Concurrent validity was evaluated by calculating the correlation between mean VAS click location and $\hat{p}$. Convergent validity was evaluated by calculating the correlations between each of the two rater-derived measures and the acoustic measure of rhoticity discussed above, F3-F2 distance. Spearman’s ρ was used due to non-normality of the data.

Our second question pertained to the variability in individual response patterns that underlies group-level patterns of rating behavior. Several dimensions of variability were taken into account. The first set of measurements examined differences in how individuals use the rating scales provided. In a binary task, listeners differ in how strict or lenient they are in assigning speech tokens to categories such as “correct /r/” and “not a correct /r/”; this can be referred to as the rater’s *bias*. We quantified the bias of each rater using the proportion of tokens (out of 40) that the rater classified as representing “correct /r/” in the binary rating task. A histogram was used to examine the range and distribution of bias values in our sample of listeners. Previous research has also shown that listeners differ in how they use a VAS scale, particularly in whether they make use of the entire scale or focus on a subset of regions (Munson et al., 2012a). Densities of the click location distributions were used to examine how raters made use of the VAS scale in the present study. A second set of measurements examined differences across individuals with respect to how well their responses agreed with our acoustic measure of rhoticity, F3-F2 distance. For binary response data, we evaluated the degree of acoustic separation between the set of tokens rated “correct /r/” versus “not a correct /r/” for a given listener by calculating the difference in mean F3-F2 distance between the two categories. For VAS data, we calculated the R^2 from a simple linear regression of an individual’s VAS click locations as a function of F3-F2 distance, which represents the percentage of variability in click location explained by the acoustic measure. Finally, we evaluated whether raters whose VAS click locations were well-correlated with the acoustic properties of tokens also showed strong acoustic separation between the “correct /r/” and “not a correct /r/” categories in their binary response data, or if these two properties could dissociate. Specifically, we calculated the correlation across individuals between the difference in means derived from the binary response data and the R^2 derived from the regression model for the VAS data.

Additional analyses were necessary to meet our final aim, which was to integrate statistical and theoretical evidence with practical considerations in order to recommend which rating method(s) to use. The analyses described above were derived from a larger of raters than the average investigator might reasonably be expected to procure. To make practical recommendations, therefore, it is important to evaluate the performance of both methods with a smaller numbers of raters. A sample size of 9 raters was selected as small enough to be practical but large enough to yield valid ratings. This was based on previous research in which we found that responses aggregated across 9 or more naive raters showed the same level of agreement with an expert rater gold standard as responses aggregated across 3

experienced listeners, a standard commonly used in published speech studies (McAllister Byun et al., 2015). In the present analysis, 1000 bootstrap resamples with $n = 9$ were selected with replacement from the total pool of 291 raters. For each resample, mean VAS click location and $\hat{p}$ were calculated for each token using only the responses from the nine raters sampled. We then assessed concurrent validity (via the correlation across tokens between mean VAS click and $\hat{p}$) and convergent validity (via the correlation across tokens between mean VAS click and F3-F2 distance, and between $\hat{p}$ and F3-F2 distance) for each resample. Finally, we computed the mean and standard deviation of these three values across 1000 resamples. A final analysis pertained to average time for task completion for both the binary and VAS methods of data collection. If there is a significant difference in speed between the rating methods, this could constitute a practical argument for adopting the faster method. It could also provide a suggestion that the faster method is less cognitively demanding, although follow-up research would be necessary to draw any concrete conclusion on this point.

3. Results

3.1 Measures assessing the validity of VAS and binary ratings

Two estimates of /r/ prototypicality were obtained. Mean VAS click locations, expressed as a proportion of the total VAS length, ranged from 0.03 to 0.98 across tokens, with a mean of 0.59 and a standard deviation of 0.27. Similarly, the estimates of $\hat{p}$ ranged from 0 to 1 across tokens, with a mean of 0.61 and a standard deviation of 0.32.

Figure 1 depicts the relationships among binary, VAS, and acoustic measures of /r/ prototypicality. Figure 1(a) plots $\hat{p}$ as a function of mean VAS click location. The correlation was very strong (Spearman's $\rho = 0.99$, $p < 0.0001$), indicating a high level of concurrent validity between the two measures. That is, $\hat{p}$ and mean VAS click location can be expected to provide broadly equivalent estimates of /r/ prototypicality, at least when the number of raters is large. Figures 1(b) and 1(c) plot mean VAS click location and $\hat{p}$, respectively, as functions of the acoustic measure, F3-F2. Both perceptual measures were highly correlated with F3-F2 (Spearman's $\rho = -0.81$, $p < 0.0001$ for mean VAS click location; Spearman's $\rho = -0.81$, $p < 0.0001$ for $\hat{p}$). This indicates a high level of convergent validity with the acoustic gold standard for both mean VAS click location and $\hat{p}$.

3.2 Individual-level analyses of VAS and binary response data

Results in the preceding section demonstrated a high level of agreement between VAS, binary, and acoustic measures of /r/ prototypicality. However, these results reflected aggregation of responses across a large number of raters ($n = 291$). It is possible that considerable variability at the level of individual raters could underlie these uniform group-level results and thus impact estimates obtained from smaller samples of raters. This section reports the results of several measures undertaken to quantify and visualize that variability. Figure 2 is a histogram showing the distribution of values of bias in our rater sample, where bias is measured as the proportion of tokens classified as “correct /r/” by an individual rater in the binary task. Although presented with the same set of tokens, some listeners rated as few as 5 out of 40 tokens correct, while others rated as many as 38 out of 40 correct. The

mean number of tokens rated correct was 15.64 with a standard deviation of 5.44. These results reveal considerable heterogeneity in bias across included raters.¹

Similarly heterogeneous results can be seen in an inspection of individual VAS response patterns. Patterns of click location can be visualized using the density of the click location distributions, i.e. smoothed plots representing the likelihood of a click occurring at a given point along the scale. The three individual densities in Figure 3 were selected to illustrate the diversity of VAS response patterns observed. The rater depicted in Figure 3(a) placed most tokens near the middle of the VAS line, assigning few tokens to the extreme ends representing highly correct or highly incorrect production. By contrast, the raters represented in Figures 3(b) and 3(c) show bimodal patterns of click placement, indicating that they tended to favor the endpoints of the scale over the central regions. However, there are clear differences between these two bimodal response patterns. The two peaks in the distribution's density in Figure 3(b) are relatively broad, indicating that this rater marks a separation between "correct" and "incorrect" categories of tokens while still differentiating among tokens in each of those categories. The rater in Figure 3(c) differs in that he/she concentrates the great majority of clicks in two limited subregions of the scale, making almost no use of the middle portion of the VAS line.

Visual inspection of the densities of the click location distributions for all raters showed that most raters used the VAS line in a bimodal pattern, similar to rater in Figure 3(b), however the degree of bimodality varied across raters. Specifically, 47 of the 291 raters had no VAS clicks in the middle 20 percent of the scale, despite the fact that 10 tokens had mean VAS click locations in this range. Of these 47, only 14 raters displayed a pattern similar to the rater in Figure 3(c), where almost all click locations are concentrated around a single point.

The above-reported measures quantified how each rater used the scale independent of any external standard--that is, we do not know which bias measure in Figure 2 or which distribution in Figure 3 represents the best match for the acoustics of the tokens presented. Therefore, it is also necessary to evaluate how individual response patterns agree with acoustic measures for both binary and VAS data. For binary ratings, each rater's agreement with the acoustic gold standard was quantified using the difference in mean F3-F2 distance between the set of tokens rated to represent "correct /r/" and the set of tokens classified as "not a correct /r/". On average, the categories were well-separated (mean 583.99 Hz, standard deviation 130.49 Hz). However, individual raters showed mean differences ranging from 38.82 Hz to 813.21 Hz. Similarly, agreement between an individual's VAS ratings and the acoustic gold standard was measured using the R^2 from a linear regression of VAS rating on F3-F2 distance, representing the percentage of variation in VAS ratings explained by the acoustic measure. The values of R^2 ranged from 0.06 to 0.73 with a mean of 0.44 and a standard deviation of 0.15.

We evaluated whether raters whose binary ratings showed strong agreement with F3-F2 distance were also likely to show strong agreement between F3-F2 distance and VAS click location. Figure 4 shows the correlation between the two measures of agreement across individuals. The relationship is positive and moderate in magnitude (Spearman's $\rho = 0.56$, p

< 0.0001). However, there were also cases of raters who showed strong agreement with the acoustic gold standard on one rating task but not on the other.

In Figure 5, we plot all three types of data: binary ratings, VAS ratings, and F3-F2 distance. We again hand-select examples to illustrate the diversity of relationships among these three measures that can be found in our sample. In Figure 5(a), the rater's VAS ratings show strong agreement with F3-F2 distance ($R^2 = 0.66$), but the difference in means on the binary task is small (489.46 Hz). In Figure 5(b), the rater shows slightly greater separation in his/her binary responses (difference of 545.79 Hz between categories). For VAS, this rater's R^2 is only slightly lower ($R^2 = 0.58$), but his/her click locations cluster at the endpoints of the scale, with little use of the intermediate regions. In Figure 5(c), the rater exhibits poor agreement with the acoustic gold standard for both VAS and binary tasks, with a low difference between category means (305.1 Hz) and low R^2 value ($R^2 = 0.13$). Thus, while the level of agreement between perceptual ratings and acoustics often goes hand-in-hand across binary and VAS tasks, it is certainly possible for individual participants to show stronger agreement with acoustic measures in one task than the other.

3.3 Measures assessing practical considerations for the choice between VAS and binary ratings

The preceding sections demonstrated strong concurrent and convergent validity between binary and VAS ratings obtained from a large sample of crowdsourced listeners; however, we found significant variability in the individual response patterns underlying those group-level results. To meet our aim of providing practical recommendations for researchers or clinicians deciding between binary and VAS methods to obtain gradient measures of child speech, it is necessary to evaluate whether they remain comparable in validity with a sample size of raters that an average investigator could reasonably be expected to procure. Thus, we assessed agreement of VAS and binary ratings with the acoustic gold standard over 1000 repeated resamples of $n = 9$ raters. The correlations between the mean VAS click location and $\hat{p}$ calculated for samples of $n = 9$ raters ranged from 0.8 to 0.98 (Spearman's ρ) across the 1000 resamples. The mean correlation was 0.93, with a standard deviation of 0.02. Thus, even with a small sample of raters, concurrent validity was high, reflecting strong agreement between VAS mean and $\hat{p}$ as a means of estimating gradient degrees of rhoticity with a sample size of 9 raters.

Across 1000 resamples with $n = 9$ raters, the correlation between mean VAS click location and F3-F2 distance ranged from -0.9 to -0.67 , with a mean of -0.79 and standard deviation of 0.04. The correlation between $\hat{p}$ and F3-F2 distance had a similar mean (-0.76) and standard deviation (0.05) across the 1000 resamples. The range of correlations in this case extended into lower values (-0.88 to -0.49). Overall, even with a realistic sample size, convergent validity was moderate to high for both methods of response aggregation.

As a final comparison across the methods, we evaluated the average time to complete the VAS versus binary portions of the experiment. For the 291 raters in the sample, the average time to complete the binary rating task was 3.37 minutes (standard deviation 1.85 minutes), and the average time to rate the same tokens using VAS was 5.16 minutes (standard

deviation 2.4 minutes). Thus, the VAS rating method took an average of 1.79 minutes longer to complete than the binary data collection method.

4. Conclusions

4.1 Validity of crowdsourced ratings

This study assessed the validity of gradient measures of child speech obtained via crowdsourcing and compared two methods for eliciting ratings. Forty child speech tokens were measured using both VAS and binary ratings obtained from 291 naive listeners recruited through Amazon Mechanical Turk. Binary responses were aggregated using the proportion of listeners rating a token as a “correct /r/” ($\hat{p}$), and VAS responses were aggregated using the mean click location across raters for a given token. At the level of the full sample of listeners, these measures were extremely well-correlated with one another, and both measures were similarly well-correlated with an acoustic gold standard, F3-F2 distance. In summary, for the present stimulus materials, both binary and VAS crowdsourced listener ratings yielded high levels of concurrent and convergent validity when a large sample of raters was used.

However, it was additionally hypothesized that considerable individual variability could underlie the uniform results obtained by aggregating across the full group. Thus, we investigated individual response patterns at several levels. As predicted, there was considerable variation across individual raters in how they used both binary and VAS scales, even after excluding raters who did not pass the attentional catch trials. Ratets also varied in the level of agreement between their responses and the acoustic gold standard; this was true of both VAS and binary response data. Those raters whose VAS responses click locations were strongly predicted by F3-F2 distance were also likely to show strong agreement between binary ratings and the acoustic gold standard, with a moderate correlation across members of the group. However, cases of dissociation were also evident. As we explore in more detail below, this suggests that some raters might be better suited to one rating type or the other.

Finally, because the full sample of $n = 291$ listeners is too large to be considered practical for ordinary purposes, we additionally examined the properties of ratings derived from 1000 bootstrap resamples of $n = 9$ listeners. Concurrent and convergent validity remained high, in these samples of $n = 9$ listeners, for both mean click location and $\hat{p}$. This indicates that, despite the individual variability described above, crowdsourced data can yield valid ratings with a sample size small enough to be considered practical for real-world applications.

4.2 Which method is recommended?

Given that both VAS and binary rating methods were found to yield gradient measures of child speech of similarly high validity, any preference for one method over another is likely to hinge on questions of efficiency and ease of data collection. In terms of time required to complete the task, the same set of listeners took longer on average to provide VAS ratings than binary ratings (mean of 5.16 minutes versus 3.37 minutes). Thus, a researcher planning to collect data from nine listeners recruited online might favor binary over VAS data

collection, with the rationale that binary ratings could be obtained more quickly and thus at a potentially lower cost. On the other hand, these differences may be seen as inconsequential in relation to the usual time and cost required to obtain acoustic measurements or solicit expert ratings of child speech. Follow-up research should also investigate whether participants' greater speed in providing binary ratings can be interpreted as evidence that binary classification of speech sounds is cognitively less taxing than VAS rating. A cognitively easier task might be preferable when collecting large numbers of ratings from each individual, since accuracy may fall off at a higher rate as the rater becomes fatigued from a challenging task.

However, recall that $n = 9$ is the number of *naïve* listeners needed to arrive at a confident estimate of token quality. In McAllister Byun et al. (2015), equivalent performance was observed between bootstrap samples of $n = 9$ naïve listeners and $n = 3$ experienced listeners (primarily speech-language pathologists). Thus, if listeners possess special training or expertise, a smaller sample may suffice. For researchers who wish to pursue the direction of identifying skilled raters and reducing the total number of individual ratings collected, there is an increasingly strong incentive to use VAS rather than binary ratings. This is because the information obtainable from binary ratings becomes overly coarse-grained at very low numbers of raters. For example, $\hat{p}$ derived from combinations of binary ratings from three listeners can take on only four possible values. On the other hand, McAllister Byun et al. additionally reported that collecting over 200 naïve listener ratings via crowdsourcing required less than 24 hours, whereas obtaining 25 ratings from experienced listeners took 3 months. Given this asymmetry in the speed with which results can be returned, it seems important to retain access to the convenience of crowdsourcing.

One clear-cut finding of the present investigation is that there is a great deal of individual variability in the performance of naïve listeners who responded to our task posted on AMT. Figure 5a presented a rater whose click locations are very well-correlated with the acoustic properties of tokens. If there were an effective method to identify other crowdsourced listeners who exhibit expert- or near-expert-level performance, VAS ratings from as few as three listeners might yield valid information about the gradient properties of speech tokens. Recent work by Harel et al. (2016) investigated this question of “finding the experts in the crowd.” The performance of individual AMT raters was assessed by measuring both reliability and validity of VAS ratings obtained when a small set of tokens was presented for repeated rating. Validity relative to an external acoustic gold standard was found to be highly correlated with intra-rater reliability across repeated presentations, suggesting that reliability across repeated ratings could represent a useful method to screen raters. However, even when applying a relatively stringent standard and including only the 60 most reliable out of 120 raters, Harel et al. (2016) continued to find considerable variability in bootstrapped resamples of $n = 9$ listeners. Because AMT raters need to be compensated for their work even in the context of a pre-screening test, it was judged impractical to apply an even stricter standard to select only raters with truly expert-like performance.

Ultimately, questions about which rating method to use and how many listeners to recruit may vary with the needs and resources of the individual seeking to obtain ratings. A funded researcher may prefer the convenience of crowdsourcing platforms like AMT, where large

numbers of naive raters can be recruited in an on-demand fashion. In this context, the binary versus VAS distinction is unlikely to be of significance. An intermediate option would be to collect ratings from a large number of students in speech pathology, who are a conveniently accessible population for most researchers, and whose services can usually be engaged at lower cost than experienced speech-language pathologists. Because students are neither experts nor truly naive, researchers should either recruit large numbers of students or use some form of pre-screening to identify students whose perceptual ratings are accurate and reliable (Harel et al., 2016). Finally, a typical clinical practitioner is unlikely to have the time and resources to obtain and analyze blinded listener ratings using currently available systems. However, as technological advances make the process of obtaining ratings more user-friendly, the potential for clinical uptake will increase. A clinician may find it easier to make an arrangement with a group of colleagues to serve as experienced raters for one other's speech samples. In this context, VAS might represent a preferable option. We note that, whatever type of listeners are used, it is essential to obtain informed consent from the client and their family prior to sharing audio samples.

We also note that either researchers or clinicians may have reasons to prefer naive listeners over experienced listeners that are unrelated to convenience or efficiency. The ultimate goal of intervention for speech disorders is to produce improvements in speech production that will be apparent to ordinary listeners with whom the speaker interacts in his/her daily activities. In practice, because of the difficulty of recruiting a large sample of naive listeners, it is more common to use experienced listener ratings and/or acoustic measures to evaluate progress over the course of intervention. When using such sensitive tools, however, there is a risk that a difference may be classified as significant when the degree of change may in fact be too small to have an impact in the real world. By offering easy access to large samples of naive listeners, crowdsourcing technologies could make it easier to evaluate the real-world impact of speech interventions.

4.3 Future directions

While the results of the present study are fairly clear-cut, a number of questions remain to be addressed in future work. We only investigated a single speech sound, English /r/. Previous research has found that the strength of the correlation between VAS click location and acoustic measures of child speech can differ across speech targets (Munson & Payesteh, 2013; Munson et al., 2012a). Thus, future studies investigating the validity of crowdsourced ratings for other speech targets and clinical populations will represent a vital follow-up to the present research. We do note that these results showed strong validity in a task of rating children's /r/ sounds, which has been associated with relatively low correlations in previous VAS work (Munson & Payesteh, 2013) and is characterized as a challenging task even for skilled clinicians (Klein et al., 2012). There is thus reason to be optimistic that similarly strong correlations will be observed for other speech sounds.

In summary, this paper demonstrated the concurrent and convergent validity of two gradient measures of child speech derived from crowdsourced ratings: the proportion of listeners rating a token as "correct /r/" in a binary response task, and the mean click location in a visual analog scaling task. These methods can be considered a reliable and efficient means to

measure the gradient properties of child speech productions, at least for the specific context of English /r/. Future work aiming to measure and predict the properties of individual raters, such as accuracy or variability, could be used to select high-performing raters for future rating tasks, or to provide feedback in a pedagogical or clinical training context.

Acknowledgments

The authors extend their thanks to Elaine Hitchcock and various lab assistants for their role in collecting and measuring the stimuli used in this study. Aspects of this research were presented at the International Congress of Phonetic Sciences (2015). This project was supported by NIH R03DC 012883.

References

- Amorosa H, von Benda U, Wagner E, Keck A. Transcribing phonetic detail in the speech of unintelligible children: A comparison of procedures. *British Journal of Disorders of Communication*. 1985; 20(3):281–287. [PubMed: 4084438]
- Baylor C, Hula W, Donovan N, Doyle P, Kendall D, Yorkston K. An introduction to item response theory and rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*. 2011; 20(3):243–259. [PubMed: 21622595]
- Becker, M., Levine, J. Experigen—an online experiment platform. 2010. Retrieved from <https://github.com/tlzoet/experigen>
- Boersma, P., Weenink, D. Praat: Doing phonetics by computer [computer program]. Version 5.4.20, retrieved 26. 2015 Sep. from <http://www.praat.org/>
- Boyce S, Espy-Wilson CY. Coarticulatory stability in american english/r. *The Journal of the Acoustical Society of America*. 1997; 101(6):3741–3753. [PubMed: 9193061]
- Buckingham HW, Yule G. Phonemic false evaluation: Theoretical and clinical aspects. *Clinical Linguistics & Phonetics*. 1987; 1(2):113–125.
- Chandler J, Mueller P, Paolacci G. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*. 2014; 46(1):112–130. [PubMed: 23835650]
- Crump MJ, McDonnell JV, Gureckis TM. Evaluating amazon’s mechanical turk as a tool for experimental behavioral research. *PLoS ONE*. 2013; 8(3):e57410. <http://doi.org/10.1371/journal.pone.0057410>. [PubMed: 23516406]
- Culton GL. Speech disorders among college freshmen: A 13-year survey. *Journal of Speech and Hearing Disorders*. 1986; 51(1):3–7. [PubMed: 3945057]
- Dalston RM. Acoustic characteristics of english/w, r, l/spoken correctly by young children and adults. *The Journal of the Acoustical Society of America*. 1975; 57(2):462–469. [PubMed: 1117100]
- Delattre P, Freeman DC. A dialect study of american r’s by x-ray motion picture. *Linguistics*. 1968; 6(44):29–68.
- Edwards J, Gibbon FE, Fourakis M. On discrete changes in the acquisition of the alveolar/Velar stop consonant contrast. *Language and Speech*. 1997; 40(2):203–210. <http://doi.org/10.1177/002383099704000204>.
- Flipsen PJ, Shriberg LD, Weismer G, Karlsson HB, McSweeny JL. Acoustic phenotypes for speech-genetics studies: Reference data for residual /r/ distortions. *Clinical Linguistics & Phonetics*. 2001; 15(8):603–630.
- Forrest K, Weismer G, Hodge M, Dinnsen DA, Elbert M. Statistical analysis of word-initial/k/and/t/ produced by normal and phonologically disordered children. *Clinical Linguistics & Phonetics*. 1990; 4(4):327–340.
- Gerrits E, Schouten M. Categorical perception depends on the discrimination task. *Perception & Psychophysics*. 2004; 66(3):363–376. [PubMed: 15283062]
- Gibbon FE. Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research*. 1999; 42(2):382–397.

- Gibson E, Piantadosi S, Fedorenko K. Using mechanical turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass*. 2011; 5(8):509–524. Retrieved from <http://dblp.uni-trier.de/db/journals/lc/lc5.html#GibsonPF11>.
- Goodman JK, Cryder CE, Cheema A. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*. 2013; 26(3):213–224. <http://doi.org/10.1002/bdm.1753>.
- Hagiwara, R. Acoustic realizations of american /r/ as produced by women and men. Vol. 90. *Phonetics Laboratory*. Dept of Linguistics, UCLA; 1995.
- Harel D, Hitchcock ER, Szeredi D, Ortiz J, McAllister Byun T. Finding the experts in the crowd: Validity and reliability of crowdsourced measures of children's gradient speech contrasts. *Clinical Linguistics & Phonetics*. 2016; Epub ahead of print. doi: 10.3109/02699206.2016.1174306
- Hewlett N, Waters D. Gradient change in the acquisition of phonology. *Clinical Linguistics & Phonetics*. 2004; 18(6–8):523–533. [PubMed: 15573488]
- Hitchcock ER, Koenig LL. The effects of data reduction in determining the schedule of voicing acquisition in young children. *Journal of Speech, Language, and Hearing Research*. 2013; 56(2): 441–457.
- Horton JJ, Rand DG, Zeckhauser RJ. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*. 2011; 14(3):399–425.
- Ipeirotis PG, Provost F, Sheng VS, Wang J. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*. 2014; 28(2):402–441.
- Julien HM, Munson B. Modifying speech to children based on their perceived phonetic accuracy. *Journal of Speech, Language, and Hearing Research*. 2012; 55(6):1836–1849.
- Klein HB, Grigos MI, McAllister Byun T, Davidson L. The relationship between inexperienced listeners' perceptions and acoustic correlates of children's /r/ productions. *Clinical Linguistics & Phonetics*. 2012; 26(7):628–645. [PubMed: 22690718]
- Klein HB, McAllister Byun T, Davidson L, Grigos MI. A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology*. 2013; 22(3):540–553. [PubMed: 23813195]
- Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*. 1999; 105(3): 1455–1468. [PubMed: 10089598]
- Li F, Edwards J, Beckman ME. Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in english and japanese toddlers. *Journal of Phonetics*. 2009; 37(1):111–124. [PubMed: 19672472]
- Liberman AM, Harris KS, Hoffman HS, Griffith BC. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*. 1957; 54(5):358. [PubMed: 13481283]
- Macken MA, Barton D. The acquisition of the voicing contrast in english: A study of voice onset time in word-initial stop consonants. *Journal of Child Language*. 1980; 7(01):41–74. [PubMed: 7372738]
- Massaro DW, Cohen MM. Categorical or continuous speech perception: A new test. *Speech Communication*. 1983; 2(1):15–35.
- Maxwell EM, Weismer G. The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics*. 1982; 3(01):29–43.
- Mayo C, Gibbon FE, Clark RA. Phonetically trained and untrained adults' transcription of place of articulation for intervocalic lingual stops with intermediate acoustic cues. *Journal of Speech, Language, and Hearing Research*. 2013; 56(3):779–791.
- McAllister Byun T, Hitchcock ER. Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Speech-Language Pathology*. 2012; 21(3):207–221. [PubMed: 22442281]
- McAllister Byun, T., Halpin, P., Harel, D. Crowdsourcing for gradient ratings of child speech: Comparing three methods of response aggregation. Paper presented at the 18th International Conference of Phonetic Sciences; Glasgow, UK. 2015.

- McAllister Byun T, Halpin P, Szeredi D. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*. 2015; 53:70–83. [PubMed: 25578293]
- McAllister, Byun T., Richtsmeier, PT., Maas, E. Covert contrast in child phonology is not necessarily extragrammatical. Paper presented at the Annual Meeting of the Linguistic Society of America; 2013.
- McGowan RS, Nitttrouer S, Manning CJ. Development of [r] in young, midwestern, american children. *The Journal of the Acoustical Society of America*. 2004; 115(2):871. [PubMed: 15000198]
- Munson, B., Payesteh, B. Clinical feasibility of visual analog scaling. Poster presented at the 2013 Convention of the American Speech-Language Hearing Association; Chicago. 2013.
- Munson B, Johnson JM, Edwards J. The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology*. 2012a; 21(2):124–139. [PubMed: 22230182]
- Munson B, Schellinger S, Urberg Carlson K. Measuring speech-sound learning using visual analog scaling. *Perspectives in Language Learning and Education*. 2012b; 19:19–30.
- Paolacci G, Chandler J, Ipeirotis PG. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*. 2010; 5(5):411–419.
- Pisoni DB, Tash J. Reaction times to comparisons within and across phonetic categories. *Attention, Perception, & Psychophysics*. 1974; 15(2):285–290.
- Richtsmeier, PT. Toronto working papers in linguistics. Vol. 33. Toronto: 2010. Child phoneme errors are not substitutions; p. 343–358.
- Ruscello DM. Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders*. 1995; 28(4):279–302. [PubMed: 8576411]
- Schellinger, S., Edwards, J., Munson, B., Beckman, ME. The role of listener expectations on judgments of children's /s/ productions. Poster presented at the Symposium on Research in Child Language Disorders; Madison, WI. 2008.
- Scobbie, J. Interactions between the acquisition of phonetics and phonology. In: Gruber, M., editor. *Proceedings of the Chicago linguistic society 34, part 2: Papers from the panels*. Chicago: 1998. p. 343–358.
- Scobbie, J., Gibbon, FE., Hardcastle, W., Fletcher, P. Covert contrast as a stage in the acquisition of phonetics and phonology. In: Broe, M., Pierrehumbert, J., editors. *Papers in Laboratory Phonology V*. Cambridge, UK: Cambridge University Press; 2000. p. 194–207.
- Shriberg LD, Flipsen PJ, Karlsson HB, McSweeney JL. Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual /3/ distortions. *Clinical Linguistics & Phonetics*. 2001; 15(8):631–650.
- Shriberg LD, Paul R, Flipsen P. Childhood speech sound disorders: From postbehaviorism to the postgenomic era. *Speech Sound Disorders in Children*. 2009:1–33.
- Smit AB, Hand L, Freilinger JJ, Bernthal JE, Bird A. The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*. 1990; 55(4):779–798. [PubMed: 2232757]
- Sprouse J. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*. 2011; 43(1):155–167. [PubMed: 21287108]
- Tyler AA, Edwards ML, Saxman JH. Acoustic validation of phonological knowledge and its relationship to treatment. *Journal of Speech and Hearing Disorders*. 1990; 55(2):251–261. [PubMed: 2329787]
- Tyler AA, Figurski GR, Langsdale T. Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech, Language, and Hearing Research*. 1993; 36(4):746–759.
- Weismer G. Acoustic analysis strategies for the refinement of phonological analysis. *Phonological Theory and the Misarticulating Child*. 1984:30–52.
- Werker JF, Logan JS. Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*. 1985; 37(1):35–44. [PubMed: 3991316]

Wolfe V, Martin D, Borton T, Youngblood HC. The effect of clinical experience on cue trading for the/rw/contrast. *American Journal of Speech-Language Pathology*. 2003; 12(2):221–228. [PubMed: 12828535]

Young E, Gilbert H. An analysis of stops produced by normal children and children who exhibit velar fronting. *Journal of Phonetics*. 1988; 16(2):243–246.

Appendix

1. Instructions for binary rating

This study investigates adults' perceptions of children's production of the English "r" sound. You will see a written word containing "r," and you will hear a child saying that word. The children in this experiment are of different ages and are at different stages of speech-sound development. Sometimes you will hear "r" productions that are very accurate, and sometimes they will sound inaccurate. Please listen carefully. You can click to listen to each file up to three times.

Your task is to rate the child's production of the "r" sound in each word. When you hear what you think is a correct "r" sound, click the button that says "correct." When you hear what you think is an incorrect "r" sound, or if the "r" sound is missing, click the button that says "incorrect." Sometimes, you won't be sure whether the "r" sound the child produced was correct or not. In those cases, you should take your best guess. We don't have any specific instructions for what to listen for when making these ratings. We want you to go with your gut feeling about what you hear. You do NOT need to assign equal numbers of "correct" and "incorrect" ratings.

2. Instructions for VAS rating

This study investigates adults' perceptions of children's production of the English "r" sound. You will see a written word containing "r," and you will hear a child saying that word. The children in this experiment are of different ages and are at different stages of speech-sound development. Sometimes you will hear "r" productions that are very accurate, and sometimes they will sound inaccurate. Please listen carefully. You can click to listen to each file up to three times.

Your task is to rate the child's production of the "r" sound in each word. This rating will be made on a line. One end of the line says "correct." The other end says "incorrect." When you hear what you think is a FULLY CORRECT "r" sound, click on the line where it says "correct." When you hear what you think is a FULLY INCORRECT "r" sound, or if the "r" sound is missing, click on the line where it says "incorrect." Sometimes, you won't be sure whether the "r" sound the child produced was correct or not. In those cases, you should click somewhere in between the two ends of the line. If the sound was in between but sounded more correct than incorrect, click somewhere on the line closer to the text that says "correct." If it sounded more incorrect than correct, click somewhere on the line closer to the text that says "incorrect." We hope that you will use the whole line when rating these sounds. We don't have any specific instructions for what to listen for when making these ratings. We want you to go with your gut feeling about what you hear.

Figure A1.
Screenshot of VAS rating interface.

FIGURE 1.

Correlation between (a) $\hat{p}$ and mean VAS click location, (b) mean VAS click location and F3-F2 Distance, (c) $\hat{p}$ and F3-F2 Distance.

FIGURE 2. Histogram of rater bias measure (proportion of tokens rated “correct /r/”) for the binary response task.

FIGURE 3.
Density plots of VAS click locations for three selected raters.

FIGURE 4.

Correlation between binary-acoustic agreement (difference in mean F3-F2 distance between sets of tokens rated “correct /r/” vs “not a correct /r/”) and VAS-acoustic agreement (proportion of variation in VAS ratings explained by F3-F2 distance), across individuals.

FIGURE 5. VAS click locations compared to F3-F2 values for three individual raters representing (a) higher performance on VAS and (b) higher categoricity and (c) overall lower performance.