Published in final edited form as:

Clin Linguist Phon. 2017 ; 31(1): 104–117. doi:10.3109/02699206.2016.1174306.

Finding the experts in the crowd: Validity and reliability of crowdsourced measures of children's gradient speech contrasts

Daphna Harel^{*}, Elaine Russo Hitchcock[†], Daniel Szeredi^{*}, Jose Ortiz^{*}, and Tara McAllister Byun^{*}

^{*}New York University, NY, USA

[†]Montclair State University, Bloomfield, NJ, USA

Abstract

Perceptual ratings aggregated across multiple non-expert listeners can be used to measure covert contrast in child speech. Online crowdsourcing provides access to a large pool of raters, but for practical purposes, researchers may wish to use smaller samples. The ratings obtained from these smaller samples may not maintain the high levels of validity seen in larger samples.

This study aims to measure the validity and reliability of crowdsourced continuous ratings of child speech, obtained through Visual Analog Scaling, and to identify ways to improve these measurements. We first assess overall validity and interrater reliability for measurements obtained from a large set of raters. Second, we investigate two rater-level measures of quality, individual validity and intrarater reliability, and examine the relationship between them. Third, we show these estimates may be used to establish guidelines for the inclusion of raters, thus impacting the quality of results obtained when smaller samples are used.

Keywords

child speech ratings; reliability; validity; covert contrasts

When children produce speech that deviates from adult target forms, these patterns are often described in terms of neutralization of contrast between two or more phonemes. For example, in a child who exhibits the phonological pattern of velar fronting, target words like *key* and *tea* may be transcribed identically (i.e., [ti]). However, a substantial literature has demonstrated that listeners' transcriptions cannot be relied upon to paint a complete picture, particularly in studies of developmental phonology. In particular, transcribers inevitably exhibit some degree of bias towards segments and contrasts familiar from their own phonological system. In light of the numerous differences in anatomy, motor control, and linguistic experience between children and adults, transcription is inevitably an imperfect reflection of phonetic reality (Amorosa, Benda, Wagner, & Keck, 1985; Weismer, 1984).

Contact: Dr Tara McAllister Byun, New York University - Communicative Sciences and Disorders, 665 Broadway, New York, New York 10012, United States, tara.byun@nyu.edu.

Statement of interest

The authors report no conflicts of interest.

Numerous acoustic and articulatory studies have gathered the precise phonetic data that are omitted when child speech is studied from broad (phonemic) transcription. With striking frequency, these studies have found evidence of *covert contrasts*, i.e., statistically reliable phonetic differences between categories that are transcribed identically (e.g., Edwards, Gibbon, & Fourakis, 1997; Gibbon, 1999; Munson, Johnson, & Edwards, 2012; Munson, Schellinger, & Urberg Carlson, 2012; Scobbie, 1998; Tyler, Edwards, & Saxman, 1990; Tyler, Figurski, & Langsdale, 1993; Weismer, 1984; Young & Gilbert, 1988). In one well-known phenomenon, children have been found to make a consistent distinction in voice onset time (VOT) between voiced and voiceless targets, although their initial consonants were uniformly transcribed as voiced (e.g., Hitchcock & Koenig, 2013; Macken & Barton, 1980; Maxwell & Weismer, 1982).

The accumulation of evidence of covert contrast has given rise to a shift in the basic conceptualization of speech development. While traditional accounts painted acquisition of a new phonemic contrast as an abrupt, categorical change, the new consensus holds that speech sound development is phonetically gradient (Hewlett & Waters, 2004; Li, Edwards, & Beckman, 2009). This suggests that covert contrast may represent a typical or even universal stage in the acquisition of phonological contrasts (Munson, Johnson, & Edwards, 2012; Munson, Schellinger, & Urberg Carlson, 2012; Richtsmeier, 2010; Scobbie, 1998). However, it is difficult to gauge the actual prevalence of covert contrast. A single phonemic contrast is generally signaled by a number of phonetic cues, of which researchers will typically measure only a limited subset. Children acquiring speech often mark contrast in phonetically unusual ways (Scobbie, Gibbon, Hardcastle, & Fletcher, 2000), and if an instrumental study finds no evidence of contrast in the selected set of parameters, it is still possible that contrast was marked in another acoustic dimension that was not measured. Unlike acoustic measures, the human ear is simultaneously sensitive to contrast in many acoustic dimensions. With appropriate methods, listener judgments might be more effective than traditional acoustic analyses as a means to derive gradient measures of speech (Munson, Schellinger, & Urberg Carlson, 2012).

Visual analog scaling

The task used to elicit perceptual judgments is known to have an important impact on listeners' patterns of response (Pisoni & Tash, 1974; Werker & Logan, 1985). One well-validated method to obtain listener ratings is visual analog scaling (VAS). In VAS, the endpoints of a line are defined to represent two different categories of speech sounds, and listeners click any location along the line to indicate the gradient degree to which they perceive the speech as belonging to one category or the other (e.g., Massaro & Cohen, 1983; Munson, Schellinger, & Urberg Carlson, 2012). For listener ratings to be considered an appropriate means to quantify gradient contrasts in child speech, it must be demonstrated that they are both valid and reliable. Validity refers to extent to which a measurement is representative of the trait it purports to measure. It is often assessed using the correlation between the measure of interest, such as VAS click location, and a gold standard already believed to be valid, such as an acoustic measurement. Reliability, on the other hand, refers to the level of similarity among a set of measurements obtained under consistent conditions.

In this case, it is of interest to assess reliability over repeated administrations of given stimuli, either within a single rater (intrarater) or across multiple raters (interrater).

Several studies (Julien & Munson, 2012; Munson, Johnson, & Edwards, 2012; Munson, Schellinger, & Urberg Carlson, 2012) have investigated the validity of VAS ratings relative to continuous acoustic measures of child speech. In general, they have found reliable correlations between the two. In a task of rating children's sibilants on a continuum from /s/ to /ʃ/, mean VAS click location was significantly correlated with centroid frequency in 22/22 adult listeners (Julien & Munson, 2012). In another study, mean VAS click location was significantly correlated with acoustic measures of children's /s/-/ʃ/ and /d/-/g/ productions, although not for /t/-/k/ productions (Munson, Johnson, & Edwards, 2012). This was true both for a sample of 21 speech-language pathologists and a comparison group of 21 listeners who lacked relevant training or experience. However, the two groups differed with respect to intrarater reliability: speech-language pathologists were significantly more consistent than naive listeners in their VAS ratings across repeated presentations of the same set of tokens. This difference between experienced and naive listeners is of particular interest for the body of research investigating the utility of crowdsourcing as a means to collect speech rating data, since crowdsourcing recruits mainly naive listeners (McAllister Byun, Halpin, & Harel, 2015; McAllister Byun, Halpin, & Szeredi, 2015).

Crowdsourcing ratings of speech

Crowdsourcing makes use of the large pool of individuals who can be recruited through online channels to solve problems of varying levels of complexity. While an individual recruited online is unlikely to display expert-level performance in a given task, responses aggregated over numerous non-experts generally converge with responses obtained from experts (Ipeirotis, Provost, Sheng, & Wang, 2014). In fields such as behavioral psychology (e.g. Goodman, Cryder, & Cheema, 2013; Paolacci, Chandler, & Ipeirotis, 2010) and linguistics (e.g. Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011), online crowdsourcing platforms have come to represent a valued resource where researchers can recruit large numbers of individuals to complete experimental tasks. Tasks that require many hours of the experimenter's time in the lab can be completed in a fraction of the time online (Sprouse, 2011); the speed of crowdsourced data collection has been described as "revolutionary" (Crump, McDonnell, & Gureckis, 2013). This paper will focus on Amazon Mechanical Turk (AMT), the crowdsourcing platform that has been most widely used by academic researchers to date.

There is justifiable concern that the increase in speed offered by crowdsourcing might come at some cost to the quality of results obtained. For example, in an effort to maximize economic gain, paid online users may click through tasks at a high speed with little regard for accuracy. Similarly, crowdsourced data collection does not allow the experimenter to standardize equipment parameters such as processor speed, screen size, or playback volume. Crowdsourcing researchers freely acknowledge that online data collection tends to be "noisier," but they argue that this noise can be overcome by recruiting larger samples than those used in the lab setting. A number of studies have empirically validated results obtained through AMT against results collected in a typical laboratory setting, either by replicating

classic findings (Crump et al., 2013; Horton, Rand, & Zeckhauser, 2011) or using direct comparison of AMT and lab samples (Paolacci et al., 2010).

Crowdsourced ratings have shown promise for measuring gradient degrees of contrast in child speech. McAllister Byun, Halpin, & Szeredi (2015) examined crowdsourced listeners' ratings of the degree of rhoticity in words produced by children at various stages in the acquisition of North American English rhotic sounds. That study found that in both large and small samples of crowdsourced listeners ($n = 250$; $n = 9$), aggregated ratings showed excellent agreement with the responses of experienced listeners, mostly speech-language pathologists. The crowdsourced ratings were also significantly correlated with an acoustic measure of rhoticity, F3-F2 distance ($r = -.79$).¹ McAllister Byun, Halpin, & Szeredi (2015) collected binary (correct/incorrect) rather than VAS ratings from individual listeners; these binary judgments were aggregated into the continuous measure $\hat{p}$, i.e. the proportion of listeners who classified a given item as a "correct" /r/ sound. A follow-up study (McAllister Byun, Halpin, & Harel, 2015) indicated that mean VAS click location was highly correlated with $\hat{p}$ ($r = .99$), and both measures were equally well-correlated with F3-F2 distance ($r = -.99$) when aggregated across a large number of raters ($n = 291$). These results suggest that VAS ratings obtained from crowdsourced listeners can provide a valid measure of gradient degrees of rhoticity in child speech.

However, when smaller numbers of raters are used (which is most likely to be the case in practical settings), outcomes are more variable. For example, when McAllister Byun, Halpin, & Harel (2015) examined subsamples containing a smaller number of raters ($n = 9$), the correlation between F3-F2 distance and mean VAS click location ranged from -0.9 to -0.67 (mean = -0.79 , standard deviation = 0.04). We posit that the range of results observed can be attributed to the fact that in a single sample of nine raters, the results are strongly influenced by the quality of raters who happened to be selected for that sample. Therefore, it should be possible to improve the validity of results from small crowdsourced samples by implementing procedures to exclude raters who exhibit below-average performance.

This study expands on previous work (McAllister Byun, Halpin, & Szeredi, 2015; McAllister Byun, Halpin, & Harel, 2015) by measuring the validity and reliability of crowdsourced measures of gradient degrees of contrast in child speech. Specifically, we first assess overall validity and interrater reliability for measurements obtained from a large set of raters. Second, we investigate two rater-level measures of quality, individual validity and intrarater reliability, and examine the relationship between them. Third, we show how estimates of validity and reliability may be used to establish guidelines for the inclusion of raters, and how these measures can impact the quality of results obtained from smaller samples.

¹Throughout this paper, correlations of click location relative to an acoustic gold standard are negative, because F3-F2 is lower in sounds that are more rhotic and thus judged to be more accurate.

Method

Task and stimuli

VAS ratings were collected from crowdsourced listeners for child speech samples representing different levels of accuracy in a late-developing phoneme, North American English /ɜ/. Both acoustic measures (Lee, Potamianos, & Narayanan, 1999; McGowan, Nitttrouer, & Manning, 2004) and trained listener ratings (Klein, Grigos, McAllister Byun, & Davidson, 2012) indicate that children's attempts to produce /ɜ/ can span a continuum from completely derhotacized to fully rhotic production. Acoustically, American English /ɜ/ is distinguished from other vowels by a particularly low height of the third formant, F3 (Delattre & Freeman, 1968; Hagiwara, 1995). At the same time, the second formant F2 is relatively high; F3 and F2 are close together in an accurate /ɜ/ (Boyce & Espy-Wilson, 1997; Dalston, 1975). Following previous research (e.g., Shriberg, Flipsen, Karlsson, & McSweeney, 2001), this study will treat F3-F2 distance as the gold standard acoustic measure of degree of rhoticity.

The task, which was advertised on AMT under the heading "Rate 'r' sounds produced by children," was implemented in the Javascript-based experiment presentation platform Experigen (Becker & Levine, 2010). Participants were informed that they would be listening to words containing the 'r' sound that had been produced by children of different ages, and that their task was to rate the 'r' sound in each word by clicking anywhere on a line with endpoints labeled "correct" and "incorrect".² Each word could be played up to three times. The orthographic form of the target word was available during rating. (See Figure 1.) Raters received instructions for using the VAS that were adapted from Julien & Munson (2012). These instructions encouraged participants to use the entire line but otherwise left the task open-ended, including the statement "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 so that raters could become familiar with the rating interface, but no feedback on response accuracy was provided at any point during the study. Participants were informed that the use of headphones was mandatory during the task; while this could not be directly enforced, they were required to report which brand of headphones they were wearing.

A single set of 30 monosyllabic words containing /ɜ/ was presented five times, with items randomly ordered within each presentation cycle.³ The selected words were isolated from probe measures administered during baseline, within-treatment, or maintenance intervals in previous studies investigating treatment for rhotic misarticulation (McAllister Byun & Hitchcock, 2012; McAllister Byun, Hitchcock, & Swartz, 2014). Three word productions were selected from each of ten child speakers, ranging in age from 80 months to 187 months (mean = 111 months, SD = 32.1 months). Six out of ten children were female, which is slightly less typical given that most studies find residual speech errors to be more prevalent

²We used "correct" and "incorrect" as the endpoints, rather than the more canonical method of labeling each endpoint with a specific phoneme, because children varied with respect to the specific phonetic character of the sound that they produced for target /ɜ/.

³Because of the randomized order of presentation, there was no concern of a *verbal transformation effect*, i.e., a perceptual change associated with repeated presentation of the same item (Munson & Brinkman, 2004; Pitt & Shoaf, 2002)

in male than female speakers (Shriberg, Paul, & Flipsen, 2009). Items were hand-selected to represent a range of accuracy levels both within and across children, with the additional goal of achieving a roughly normal distribution with respect to the acoustic gold standard measure, described below. Due to the relatively small number of child-friendly monosyllabic words containing /ʒ/, only a few word types (*fur*, *her*, *purr*, *sir*, *stir*, *turn*) were represented in the set of tokens used.

F3-F2 distance was measured in each token to serve as the gold standard acoustic measure of rhoticity. All acoustic measurements were carried out by trained student assistants using Praat software (Boersma & Weenink, 2015). Assistants used the audio and spectrographic record to locate the /ʒ/ interval within each word. They then placed a cursor at the point that they judged to represent the lowest height of F3 in the rhotic portion of each token, avoiding points that represented outliers relative to adjacent points in the formant tracks generated in Praat. The LPC filter order used in formant tracking was selected on an individual basis for each participant via inspection by the authors. Finally, the student assistants ran a script to measure the first three formants at the selected point in each word (Lennes, 2003); in cases where the automatically extracted measurements were judged not to be well-matched to the visible energy contour in the spectrogram, hand measurement was used instead. For additional detail on the protocol used to obtain acoustic measures, see McAllister Byun & Hitchcock (2012).

A total of $n = 170$ 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 ($n = 41$ did not meet this requirement). and to provide responses on all 150 trials ($n = 9$ did not meet this requirement). Per self-report, participants had no history of hearing or speech-language impairment. Thus, a total of $n = 120$ were used for all subsequent analyses. Participants were compensated \$1.75 for their work. Data collection was completed in roughly 10 hours and cost roughly \$350, including Amazon fees. The included group of participants had a mean age of 37.6 years, with a standard deviation of 10.5 years.

Statistical analyses

The first set of analyses investigated the adequacy of the ratings obtained at the group level. Overall validity was calculated as the correlation between the mean VAS click location, pooled across all listeners and presentation cycles, and the acoustic gold standard, F3-F2 distance. Spearman's rank correlation was used due to non-normality in the data (specifically, the heavy-tailed distribution of mean VAS click location). Interrater reliability was assessed by calculating an intraclass correlation coefficient (ICC) over all rating data. The ICC, which ranges from 0 to 1, can be interpreted as the level of agreement across multiple observers' measurements of a single set of items. In the present case, each rater was distilled down to a single score for each token by taking that rater's mean click location across the five presentation cycles. The ICC was then calculated as the consistency in mean VAS click location for the 30 study stimuli across all 120 raters (Shrout & Fleiss, 1979). For the purpose of calculating the ICC, the 120 raters were treated as a random sample and mean

VAS click location across the five presentation cycles was treated as the target for each token.

The next set of analyses aimed to understand how individual raters differed in validity and reliability. In this context, responses from the first four presentation cycles were analyzed, while responses from the fifth presentation cycle were held in reserve as a test set for the final set of statistical analyses, to be discussed in detail below. For each rater, individual validity was calculated as the Spearman's correlation coefficient across tokens between the mean VAS click location and the acoustic gold standard, F3-F2 distance. To assess intrarater reliability, the ICC was calculated to measure degree of consistency within a single rater's click locations for all stimuli across the four presentation cycles, treating the stimuli as a random sample of possible tokens. For both individual validity and intrarater reliability, descriptive statistics were calculated to assess the mean, standard deviation and range across raters in the sample. Furthermore, the relationship between these two measures of rater quality was assessed via Spearman's rank correlation.

In practical settings, it is desirable to obtain only a single set of ratings from a small number of raters, such as $n = 9$ (McAllister Byun, Halpin, & Szeredi, 2015), while taking reasonable steps to ensure that ratings obtained are valid and reliable. This could be accomplished by administering a brief screening measure and using the results to determine which raters will be eligible to complete the full set of study stimuli. To assess this possibility, the third set of analyses examined how well measures of validity and reliability derived from a small subset of the data (the fifth presentation cycle) correlated with a larger subset of the data from the same individual (the first four presentation cycles). First, we examined the Spearman's rank correlation between each individual's validity over the first four presentation cycles, as calculated above, and his/her validity on the fifth presentation cycle. Second, we examined the Spearman's rank correlation between a listener's intrarater reliability on the first four presentation cycles, compared to his or her validity on the fifth presentation cycle.

Lastly, to examine how outcomes can be impacted by setting criteria to control for rater quality, we explored the range of validity estimates obtained when calculating over different subsets of raters. To do this, 10,000 repeated samples of size $n = 9$ were drawn, with replacement, from the full set of raters. Only the data obtained from the fifth presentation cycle were used in this analysis. For each sample, we calculated the Spearman's rank correlation between F3-F2 distance and VAS click location, averaged across the 9 raters selected. This procedure was performed three times: drawing from all possible raters; drawing only from raters whose intrarater reliability across the first four presentation cycles was greater than 0.60; and drawing only from raters whose intrarater reliability across the first four presentation cycles was greater than the median value in the sample, 0.82.

Results

Mean VAS click locations, expressed as a proportion of the total VAS length, ranged from 0 to 1 across tokens, with higher values indicating a greater degree of rhoticity. This distribution of mean VAS click locations had a mean of 0.47, median of 0.44, standard deviation of 0.34, and interquartile range of 0.28. Figure 2 shows that mean VAS click

location was highly correlated with F3-F2 distance when responses were pooled across all listeners and presentation cycles (Spearman's $\rho = -0.74$, $p < 0.001$), indicating a high level of overall validity. As noted above, the relationship between mean click location and F3-F2 distance is negative, since higher VAS ratings and lower F3-F2 distance indicate higher levels of rhoticity. In figure 2, each dot represents a single token, and the size of the dot is determined by the standard deviation across click locations for that token. Tokens at the extreme “correct” end of the scale were rated with the highest precision (smallest standard deviation), and tokens with intermediate acoustic properties showed the lowest precision. Across the full set of ratings obtained, the interrater reliability was 0.60, indicating a moderate level of agreement across raters.

Figure 3 illustrates the range of variation across individuals with respect to both intrarater reliability and individual validity. Correlations reflecting individual-level validity ranged from -0.83 to -0.02 , with a mean of -0.67 and a standard deviation of 0.14 . Thus, although validity was moderate to high when averaging across multiple raters, there were large differences in validity across individual raters. Similarly, intrarater reliability ranged from -0.03 to 0.97 , with a mean of 0.78 and a standard deviation of 0.16 . In this case, the negative values obtained for some raters are sample estimates for population values that are likely equal to zero. The top panel of figure 3 depicts the relationship between intrarater reliability and individual validity. The correlation between these two measures was moderate to high (Spearman's $\rho = -0.62$, $p < 0.001$).

As a concrete illustration of how raters differed in their response patterns, the bottom two panels of figure 3 show one example of a high-quality rater and one low-quality rater. The rater on the left demonstrated both high validity and high reliability. By contrast, the rater on the right showed a virtually random response pattern, yielding very low validity and low reliability.

An individual's validity on the first four presentation cycles was found to be highly correlated with his or her validity on the fifth presentation cycle (Spearman's $\rho = 0.89$, $p < 0.001$). This indicates that estimates of individual-level validity remained reasonably consistent, whether they were calculated from a single set of ratings or from the mean over four sets of repeated ratings. Furthermore, there was also a high correlation between an individual's *reliability* on the first four presentation cycles and his or her *validity* on the fifth presentation (Spearman's $\rho = -0.70$, $p < 0.001$). Thus, if we compute either rater-quality measure (e.g., reliability) for a subset of ratings from one listener, we can form a reasonable estimate of the other measure (e.g., validity) for a different subset of data from the same listener.

Finally, figure 4 illustrates how outcomes are influenced when controls for rater quality are introduced. Each boxplot represents the range of validity values calculated over 10,000 repeated sampling runs, where each subsample had size $n = 9$. As we discuss in more detail below, our focus here is on the use of reliability measures as a means to select high-quality raters. In the leftmost boxplot, all raters were included in the pool for repeated sampling, representing a study design where no quality controls are introduced to screen raters. In this case, the median level of validity across 9 raters was relatively high (Spearman's $\rho = -0.72$).

However, in the worst-case scenario (i.e., the lowest validity observed over 10,000 random resamples with $n = 9$), a VAS-acoustic correlation as low as -0.52 could be observed. The second boxplot depicts the range of outcomes observed over 10,000 resamples when only raters with an intrarater reliability of 0.6 or higher were included ($n = 104$). In this case, the median level of validity across 9 raters is only slightly higher than when using all raters (Spearman's $\rho = -0.73$), and the worst-case level of validity shows a similarly small improvement (Spearman's $\rho = -0.54$). The rightmost boxplot includes only raters with an intrarater reliability above the median in the present sample ($ICC > 0.82$, $n = 60$.) In this case, the median level of validity across 9 raters increases again (Spearman's $\rho = -0.74$), and the worst-case level of validity increases to -0.61 .

Discussion

Review of results

This study aimed to measure the validity and reliability of crowdsourced measures of degrees of gradient contrast in child speech. One hundred and twenty listeners were recruited on Amazon Mechanical Turk to rate a set of 30 word-level productions containing /ʒ/, elicited from child speakers who represented varying stages in the acquisition of the North American English rhotic. Each token was rated by every listener five times, randomly ordered within five presentation cycles. This design made it possible to assess validity and reliability both at the group level and at the individual rater level. At the group level, ratings exhibited a high level of validity (Spearman's $\rho = -0.74$) and a moderate level of interrater reliability ($ICC = 0.6$).⁴ Individual raters were observed to vary widely in quality, both with respect to validity and reliability over repeated ratings. The two measures were significantly correlated with one another, indicating that listeners who were more consistent across repeated ratings of the same token also tended to give more valid ratings.

In a realistic context, we are unlikely to use either the complete set of 120 raters or a single individual. Most researchers would favor a small set of raters, with previous research recommending $n = 9$ as an appropriate sample size (McAllister Byun, Halpin, & Szeredi, 2015). An analysis of repeated resamples with $n = 9$ revealed that validity was generally high (median Spearman's $\rho = -0.72$), but it was also possible to select a sample with a disproportionate representation of low-quality raters, yielding a VAS-acoustic correlation as low as -0.52 . One possible way to maintain the quality of ratings while using a smaller sample is to administer an eligibility test and use these results to select only raters who are likely to perform adequately on the main rating task. A third set of analyses established that measures of validity and reliability derived from a small subset of the data were highly correlated with the same measures derived from a larger subset of the data from the same individual, suggesting that a relatively brief screening measure can be a useful predictor of a rater's overall quality. Finally, we observed that when the eligibility criterion was set to exclude listeners whose intrarater reliability fell below the median for our sample, the range of validity outcomes across repeated resamples with $n = 9$ became more favorable, with a minimum Spearman's rho of -0.61 .

⁴It should be noted that the moderate level of interrater reliability may be due to the high levels of consistency observed on some tokens and could change if different study stimuli were selected.

Implications

This study supports previous research in finding that non-expert listeners' VAS ratings are well-correlated with continuous acoustic measures of child speech (Julien & Munson, 2012; Munson, Johnson, & Edwards, 2012; Munson, Schellinger, & Urberg Carlson, 2012) and that these ratings remain valid when listeners are recruited through online crowdsourcing platforms (e.g., McAllister Byun, Halpin, & Szeredi, 2015; McAllister Byun, Halpin, & Harel, 2015). The most novel contribution of the present study is the finding that individual validity and intrarater reliability are correlated, and that raters can be usefully screened on the basis of reliability in repeated ratings of a subset of tokens. For the reasons discussed above, it is appropriate to use some kind of pre-screening measure when collecting responses from a relatively small number of raters, e.g. $n = 9$. Examining each individual's validity, i.e. the correlation between click location and an acoustic gold standard measure, represents one possible basis for including or excluding raters. However, acoustic measures are time-consuming to obtain, and as discussed above, researchers may be unsure which parameter or parameters should be obtained in a given case. Furthermore, because an individual's level of attention can fluctuate over time and response criteria can shift over the course of accumulated experience, it is generally not sufficient to administer a screening criterion only once. On the other hand, if a single set of stimuli is used repeatedly to make eligibility determinations, listeners may become familiar with the individual tokens, compromising the predictive power of the measure. Thus, a researcher who chooses to screen raters based on validity is likely to need to devote a substantial amount of time to obtaining new gold standard measures when introducing a new target or refreshing the eligibility trial stimuli. By contrast, reliability across repeated ratings can be assessed with any randomly selected subset of the sample of interest. Thus, we maintain that individual intrarater reliability represents a more flexible and repeatable measure that can achieve the same purpose of selecting raters who will produce more valid ratings.

Future directions and conclusions

There is a need for follow-up research to address several limitations of the present study. First and foremost, the present results are based entirely on listeners' ratings of children's productions of rhotic sounds; it is necessary to demonstrate that the relationships described in this paper also hold true in the context of rating other speech targets. Second, previous studies using crowdsourcing (e.g. McAllister Byun, Halpin, & Szeredi, 2015; McAllister Byun, Halpin, & Harel, 2015) have included steps to identify and exclude raters whose responses are suggestive of a lack of attention to the task. These exclusions often add up to a substantial portion of the original sample (e.g., 20.4% in McAllister Byun, Halpin, & Harel, 2015). The present study does not address the relationship between these attentional catch trials and measures of validity or reliability; this represents an important topic for future study.

Despite growing awareness of VAS as a valid means to measure covert contrast in child speech, the method remains relatively underrepresented in the broader literature on speech development. The relatively low rate of adoption of VAS may be partly attributable to a perception that the method is difficult or time-intensive to implement. In actuality, although an initial time investment is needed to set up a system for VAS data collection, in the long

term it can offer a substantial increase in efficiency. It is hoped that the present study, along with our broader efforts to establish online crowdsourcing as a valid means to obtain speech ratings from large numbers of listeners, can lower these perceived barriers and increase the uptake of VAS as an efficient means of measuring gradient degrees of contrast in child speech.

References

- Amorosa H, Benda U von, 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]
- Becker M, Levine J. Experigen—an online experiment platform. 2010 Retrieved from <https://github.com/tlozoot/experigen>.
- Boersma P, Weenink D. Praat: Doing phonetics by computer [computer program]. Version 5.4.20. 2015 [retrieved 26 September 2015] 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]
- Crump MJ, McDonnell JV, Gureckis TM. Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*. 2013; 8(3):e57410. [PubMed: 23516406]
- 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.
- 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.
- 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.
- Hagiwara R. Acoustic realizations of American /r/ as produced by women and men. *UCLA Working Papers in Phonetics*. 1995; 90
- 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 H, Munson B. Modifying speech to children based on their perceived phonetic accuracy. *J. Speech Lang. Hear. Res.* 2012; 55(6):1836–1849. [PubMed: 22744140]
- 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.
- 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]

- Lennes M. Collect formant data from files.praat. 2003 [Retrieved 8 October 2015] <http://www.helsinki.fi/lennes/praat-scripts/public/collectformantdatafromfiles.praat>.
- 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]
- 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.
- 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. *Proceedings of the 18th International Congress 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]
- McGowan RS, Nitttrouer S, Manning CJ. Development of [ɹ] in young, midwestern, American children. *The Journal of the Acoustical Society of America*. 2004; 115(2):871. [PubMed: 15000198]
- Munson B, Brinkman KN. The influence of multiple presentations on judgments of children's phonetic accuracy. *American Journal of Speech-Language Pathology*. 2004; 13(4):341–354. [PubMed: 15719900]
- 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*. 2012; 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*. 2012; 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.
- Pitt MA, Shoaf L. Linking verbal transformations to their causes. *Journal of Experimental Psychology: Human Perception and Performance*. 2002; 28(1):150–162.
- Richtsmeier, PT. *Toronto Working Papers in Linguistics*. Vol. 33. Toronto: 2010. Child phoneme errors are not substitutions; p. 343–358.
- 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 /r/ 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.
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86(2):420. [PubMed: 18839484]

- 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. *J Speech Hear Res*. 1993; 36(4):746–759. [PubMed: 8377487]
- 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]
- 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.

Figure 1.
Screenshot of VAS interface as it appeared to crowdsourced raters.

Figure 2.

Correlation across tokens between F3-F2 distance mean VAS click location, across all raters and presentation cycles. The size of the dot represents the standard deviation across click locations for that token.

Figure 3.

Top panel: Correlation between intrarater reliability and individual validity. Bottom panel: Examples of individual patterns of click location relative to F3-F2 distance. Red triangles indicate the mean click location across the first four presentation cycles for each token; black dots are the individual click locations for each presentation of a token.

Figure 4.

Boxplots representing the distribution of validity values obtained over 10,000 repeated samples with $n = 9$ raters, under three different levels of rater quality control.