CROWDSOURCING FOR GRADIENT RATINGS OF CHILD SPEECH: COMPARING THREE METHODS OF RESPONSE AGGREGATION

Tara McAllister Byun, Peter Halpin, Daphna Harel

New York University
tara.byun@nyu.edu

ABSTRACT

Impressionistic transcription can obscure systematic phonetic differences, but well-constructed perceptual tasks can draw out these covert contrasts. This study assessed the validity of gradient speech ratings obtained via online crowdsourcing. Stimuli were 40 /r/ words from children receiving treatment for /r/ misarticulation. Listeners rated each item twice, once using visual analog scaling (VAS) and once using binary rating. Correlations were examined across mean VAS click location, item parameters from an IRT model estimated over binary ratings, and the proportion of listeners scoring a given item "correct" ($\hat{p}_{correct}$). With 381 listeners, all three measures were highly intercorrelated (Spearman's $\rho > .98$). We also examined mean rankings over 1000 bootstrap resamples of 9 listeners, a sample size considered practical for real-world applications. There was a robust correlation between mean rankings derived from VAS versus $\hat{p}_{correct}$. We conclude that crowdsourcing can be a valid and practical source for gradient measures of child speech.

Keywords: acquisition and disorders; covert contrast; perceptual rating; crowdsourcing.

1. INTRODUCTION

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, or neutralization of a contrast between two phonemes. For instance, a child with the phonological pattern of stopping may be perceived to produce the words "sea" and "tea" as homophones ([ti]). However, instrumental measures may reveal *covert contrast*, i.e. reliable phonetic differences between sounds that adult listeners perceive as neutralized (e.g. [21, 19, 20, 4, 16]).

Covert contrast is described as a widespread phenomenon in child phonology, and existing estimates are thought to understate its actual prevalence ([17, 15]). Acoustic or articulatory studies can only detect covert contrast in the limited set of parameters that the investigator chooses to measure, and children acquiring speech often mark contrast in phonetically unusual ways, creating a high potential for false negative results. Because the human ear can detect contrast in a wide range of acoustic dimensions, a sufficiently fine-grained perceptual rating task may provide a more sensitive indicator of covert contrast than acoustic or articulatory measurements. A recent body of research by Munson and colleagues (e.g. [12]) has advocated for the use of visual analog scaling (VAS), in which the target sound and a perceived substitution are represented as endpoints of a line, and listeners click any location to indicate the prototypicality of each production. Munson and colleagues have shown that mean VAS click location correlates significantly with continuous acoustic measures for various overt and covert contrasts in child speech.

However, VAS has limitations as a method to evaluate covert contrast in child speech. First, it may not be equally reliable in differentiating all contrasts; for example, [11] reported poor interrater reliability when VAS was used to rate children's /r/ sounds. Second, although listeners do have the ability to perceive and respond to within-category distinctions, they detect them less consistently than boundary-crossing phonemic contrasts (see discussion in, e.g., [13]). This could reduce the reliability of VAS measurements for items that fall near a boundary between categories. A preferable approach might capture the gradience of child speech without requiring listeners to override their natural tendency to perceive discrete phonemic categories. Such a solution may be available through Item Response Theory (IRT), which models the probability of a given response as a function of both item and person/rater characteristics (e.g., [1]). When many listeners categorize items in a binary fashion, individual biases yield different boundary locations, and IRT can be used to derive a continuous measure of category prototypicality for each item.

While both VAS and IRT are supported by theoretical and empirical arguments, they are not widely used in clinical research on speech sound disorders.

This is in part a reflection of practical limitations of both methods. Computing a 2-parameter IRT model requires a large number of raters–as a rule of thumb, roughly 10 times as many raters as the *n* of items. VAS is less demanding in terms of manpower, but studies generally report mean click location over roughly 20 listeners. Furthermore, assigning prototypicality judgments by selecting points along a continuum may represent a cognitively more demanding task than simple categorization. Making speech rating tasks available online is one way to facilitate the process of reaching appropriate listeners and collecting their judgments (e.g. [10]). In recent years, further streamlining has been made possible through the rise of online crowdsourcing platforms, where large numbers of individuals are available to complete simple, repetitive tasks for small sums of money.

Numerous crowdsourcing platforms exist, but discussion here will focus on the platform that has been most widely used in academic research, Amazon Mechanical Turk (AMT). Virtually anyone can sign up to post electronic tasks on AMT or to complete tasks for payment as an AMT worker. Although it was designed for commercial rather than research purposes, recent years have seen pronounced interest in AMT as a means of accessing a vast, inexpensive participant pool for behavioral studies in psychology (e.g. [7, 14]) and linguistics (e.g. [18, 6]). These studies have reported that AMT data are broadly comparable to results collected from typical laboratory samples. The ease of AMT data collection has been described as "revolutionary" ([3]); for example, Sprouse [18] reported that it took two hours using AMT to conduct a full replication of a task that required 88 experimenter hours in the lab setting.

There is relatively little literature on the use of crowdsourcing to classify speech produced by children or individuals with a speech disorder. McAllister Byun et al [9] investigated the validity of AMT listeners' ratings in a study that collected binary ratings of individual /r/ words from 250 AMT listeners and 25 experienced listeners. Across items, there was a strong correlation between the proportion of AMT listeners and the proportion of experienced listeners who classified a given item as a "correct /r/ sound" (Pearson's r = .92). A bootstrap analysis found that the "industry standard" level of agreement with an expert listener gold standard was matched when responses were aggregated across samples of at least 9 AMT listeners. Thus, preliminary evidence suggests that crowdsourced listener judgments can represent a valid method for binary categorization of child speech productions. However, McAllister Byun et al [9] did not investigate the validity of crowdsourcing as a means to obtain gradient characterizations of child speech data. The present study took up this question. 40 child productions of North American English /r/ were placed on a continuum of prototypicality (or "/r/-ness") using both continuous and dichotomous rating scales. Three approaches to response aggregation were used to estimate the prototypicality of each item: (a) VAS click location, averaged across multiple raters; (b) IRT item parameters, estimated from binary ratings collected from multiple raters; (c) the proportion of raters who assigned the /r/ label to each item in the binary rating task. Comparisons across these methods offer a preliminary look at the relative validity and efficiency of different methods of obtaining gradient measures of speech through crowdsourcing.

2. METHOD

2.1. Task and stimuli

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 ([2]). Listeners rated 40 words containing a target /r/ sound;[1] words were collected from English-speaking children over the course of intervention for misarticulation of /r/. Each item was rated two separate times by each rater, once using VAS and once using binary rating. The 40-item word list included 10 items representing each of the following categories: syllabic /r/, postvocalic /r/ (offglide of rhotic diphthong), singleton onset /r/, and onset /r/ in a cluster. Items were hand-selected so that each category featured varying levels of acoustically determined accuracy (distance between formants F2 and F3, where a lower F3-F2 distance indicates a more accurate /r/ sound.)[2] Items were presented in random order, blocked by rating method. The order of binary and VAS methods was counterbalanced across participants.

Instructions for the VAS task were adapted from Julien & Munson [8], and instructions for binary rating were developed to parallel the VAS instructions as closely as possible. 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. In both tasks, listeners received the following instruction, adapted from Julien & Munson: "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.[3]

2.2. Data collection

A total of 526 workers were recruited via Amazon Mechanical Turk to complete the listening task described above. Raters were required to be self-reported native speakers of English. To identify listeners who were inattentive or unreliable, 20 attentional "catch trials" were randomly interspersed with experimental trials, with 10 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 16/20 correct) across catch trials.

Data collection was completed in 21.4 hours and cost $722, including Amazon fees. Results were discarded from 28 participants due to incomplete data, and from 131 participants who did not exceed chance-level performance on attentional catch trials. A total of 9 items were discarded because lack of variability in binary responses (too many 1 or 0 responses) would cause failure of model convergence. The exclusion of these items from the analysis was not of concern because their classification was unambiguous.

2.3. Analyses

VAS estimates of /r/ prototypicality were obtained for each item by taking the mean and standard deviation of click location across all responses from included raters. IRT estimates of /r/ prototypicality were obtained by fitting a two-parameter logistic item response model. The parameters estimated were (a) prototypicality or /r/-ness of each item ("IRT /r/-ness"), and (b) the discrimination parameter, representing the slope of the item characteristic curve for each item at its point of inflection. Finally, binary rating data were also aggregated across raters by calculating the proportion of individuals who rated a given item as correct, $\hat{p}_{correct}$. This method was included as a computationally less demanding means of deriving gradient estimates of /r/ prototypicality from binary rating judgments. To evaluate the relative validity of these methods, we calculated the correlation between mean VAS click location and IRT /r/-ness, between IRT /r/-ness and $\hat{p}_{correct}$, and between mean VAS click location and $\hat{p}_{correct}$. Spearman's ρ was used in all cases due to non-normally distributed data.

While the above-described analyses allowed us to compare three different methods to obtain gradient measures of /r/-ness, they were collected under somewhat idealized conditions, i.e. a larger n of raters than the average investigator could reasonably be expected to procure. To estimate the validity of these methods under more realistic conditions, it is necessary to examine their performance with smaller numbers of raters. However, it is a well-known limitation of IRT that 2-parameter models will fail to converge when sample sizes are small. In the present cases, over 300 raters would need to be used in order for the IRT model to converge. Therefore, the second part of our analysis focused on a comparison between mean VAS click location and $\hat{p}_{correct}$ at a number of raters that was judged to be both practical and sufficient to yield valid ratings. A sample size of 9 raters was selected based on McAllister Byun et al [9], where it was 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). In the present study, 1000 bootstrap resamples of the pool of 381 raters were selected with replacement. For each resample, we used only the raters sampled to calculate (a) the ranking of items based on mean VAS click location, and (b) the ranking of items based on $\hat{p}_{correct}$. We then calculated the mean and standard deviation of the ranking of each item over the 1000 resamples. The mean rankings of the items under each of the two methods were compared via Spearman's ρ. To assess the loss of information due to using 9 raters instead of the total 381 raters, bootstrap estimates of Spearman's correlation between the rankings based on 9 raters and the rankings based on all raters were calculated for both mean VAS click location and $\hat{p}_{correct}$.

3. RESULTS AND DISCUSSION

Fig. 1 plots mean VAS click location as a function of the IRT-derived item parameter. These measures were highly correlated (Spearman's ρ = .988,p < .001). Thus, at the level of the full listener sample, strong agreement was observed between VAS and IRT methods of estimating /r/ prototypicality.

Fig. 2 represents $\hat{p}_{correct}$, plotted as a function of

IRT-derived item parameter. Again, a high correlation was observed (Spearman's ρ = .992, p < .001). Thus, at the level of the full listener sample, IRT-estimated item parameters and the proportion of listeners assigning the "correct" rating can be regarded as broadly equivalent.

Figure 1: Mean VAS vs IRT /r/-ness

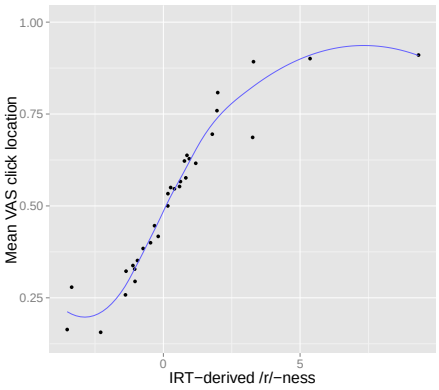
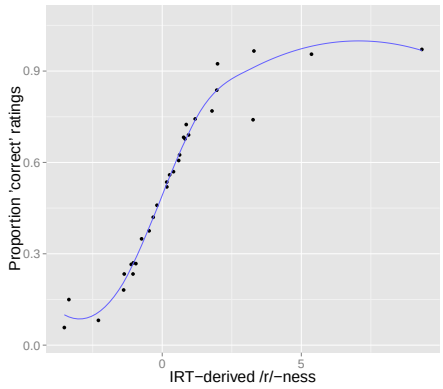


Figure 2: $\hat{p}_{correct}$ vs IRT /r/-ness

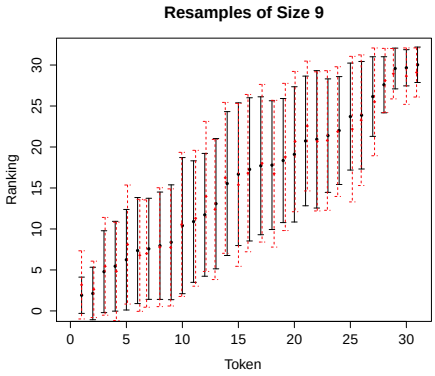


A high correlation was observed between $\hat{p}_{correct}$ and the mean VAS click location (Spearman's rho = .989, p < .001). We conclude that these two methods also return equivalent results at the level of the full listener sample (Figure not shown).

Fig. 3 represents the mean ranking of each item over 1000 resamples with sample size = 9 raters. Mean rankings based on VAS mean and $\hat{p}_{correct}$ are superimposed, with VAS mean in black and $\hat{p}_{correct}$ in red. Mean rankings were highly correlated across the two measures (Spearman's ρ = .992,p < .001). Confidence intervals for the rankings are overlapped for the two methods, and there is no systematic pattern of the confidence interval for one method exceeding that of the other method. These results reflect broad equivalence between VAS mean and $\hat{p}_{correct}$ as a means of estimating gradient degrees of /r/-ness with a sample size of 9 raters.

High correlations were observed between rankings based on 9 raters and rankings based on all raters for both VAS mean (Spearman's ρ = 0.919, p < .001) and $\hat{p}_{correct}$ (Spearman's ρ = 0.902, p < .001), indicating minimal loss of information when using 9 raters instead of the total number.

Figure 3: Mean ranking.



4. CONCLUSIONS

This study assessed the validity of gradient measures of child speech obtained through crowdsourced listener ratings. Consistent with previous research, we found that Visual Analog Scaling can be used to place items on a continuum of prototypicality. The present study further demonstrated that gradient measures of /r/-ness can be derived from binary ratings assigned by a large number of naive listeners. With a sample of 381 naive raters, there was an extremely high correlation between mean VAS click location and item parameters derived by estimating an IRT model from binary ratings. Both methods were also highly correlated with a second measure derived from binary ratings, the proportion of listeners rating a token as "correct" ($\hat{p}_{correct}$).

This study also posited that online crowdsourcing could make gradient measures of speech less challenging to obtain, increasing practical uptake of these experimentally validated methods. To assess the feasibility of our methods, we examined the mean ranking derived from 1000 resamples of n = 9 listeners. The mean ranking derived from VAS click location continued to be very highly correlated with a measure derived from binary ratings (in this case, proportion of listeners rating a token as "correct"), and there were no systematic differences in ranking variability. We conclude that both binary and continuous ratings obtained through crowdsourcing can be used to derive gradient measures of child speech. VAS and $\hat{p}_{correct}$ have comparable validity with n = 9 listeners, a sample

size considered practical for real-world applications.

5. REFERENCES

[1] Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., Yorkston, K. Aug 2011. An introduction to item response theory and Rasch models for speech-language pathologists. *Am J Speech Lang Pathol* 20(3), 243–259.

[2] Becker, M., Levine, J. 2010. *Experigen–an online experiment platform*.

[3] Crump, M. J. C., McDonnell, J. V., Gureckis, T. M. 03 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE* 8(3), e57410.

[4] Edwards, J., Gibbon, F., Fourakis, M. 1997. On discrete changes in the acquisition of the alveolar/velar stop consonant contrast. *Language and Speech* 40(2), 203–210.

[5] Flipsen, P. J., Shriberg, L. D., Weismer, G., Karlsson, H. B., McSweeny, J. L. 2001. Acoustic phenotypes for speech-genetics studies: reference data for residual /r/ distortions. *Clinical Linguistics & Phonetics* 15(8), 603–630.

[6] Gibson, E., Piantadosi, S., Fedorenko, K. 2011. Using mechanical turk to obtain and analyze english acceptability judgments. *Language and Linguistics Compass* 5(8), 509–524.

[7] Goodman, J. K., Cryder, C. E., Cheema, A. 2013. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making* 26(3), 213–224.

[8] Julien, H. M., Munson, B. Dec 2012. Modifying speech to children based on their perceived phonetic accuracy. *J. Speech Lang. Hear. Res.* 55(6), 1836–1849.

[9] McAllister Byun, T., Halpin, P. F., Szeredi, D. Dec 2014. Online crowdsourcing for efficient rating of speech: A validation study. *J Commun Disord* Early online.

[10] Munson, B. 2013. Assessing the utility of judgments of children's speech production made by untrained listeners in uncontrolled listening environments. Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., (eds), *INTERSPEECH*. ISCA 2147–2151.

[11] Munson, B., Payesteh, B. 2013. Clinical feasibility of visual analog scaling. *Poster presented at the 2013 Convention of the American Speech-Language Hearing Association* Chicago.

[12] Munson, B., Schellinger, U. C. K., Sarah K. 2012. Measuring speech-sound learning using visual analog scaling.

[13] Nielsen, K. Dec 2014. Phonetic imitation by young children and its developmental changes. *J. Speech Lang. Hear. Res.* 57(6), 2065–2075.

[14] Paolacci, G., Chandler, J., Ipeirotis, P. G. June 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5), 411–419.

[15] Richtsmeier, P. T. 2010. Child phoneme errors are not substitutions. *Toronto Working Papers in Linguistics, 33* Toronto. 343–358.

[16] Scobbie, J. 1998. Interactions between the acquisition of phonetics and phonology. *Proceedings of the Chicago Linguistic Society 34, Part 2: Papers from the Panels* Chicago. 343–358.

[17] Scobbie, J., Gibbon, F., Hardcastle, W., Fletcher, P. 2000. Covert contrast as a stage in the acquisition of phonetics and phonology. In: Broe, M., Pierrehumbert, J., (eds), *Papers in Laboratory Phonology V*. Cambridge, UK: Cambridge University Press 194–207.

[18] Sprouse, J. Mar 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behav Res Methods* 43(1), 155–167.

[19] Tyler, A. A., Edwards, M. L., Saxman, J. H. May 1990. Acoustic validation of phonological knowledge and its relationship to treatment. *J Speech Hear Disord* 55(2), 251–261.

[20] Tyler, A. A., Figurski, G. R., Langsdale, T. Aug 1993. Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *J Speech Hear Res* 36(4), 746–759.

[21] Young, E. C., Gilbert, H. R. Apr 1988. An analysis of stops produced by normal children and children who exhibit velar fronting. *Journal of Phonetics* 16(2), 243–246.

[1] We use /r/ rather than /ɹ/ to denote the North American English rhotic in an allophone-independent way.

[2] Acoustic measures were not used as a gold standard indicating the correct/incorrect status of a given item, since the formant heights associated with perceptually correct /r/ productions take on different values for children of different ages, and the present stimulus set included children aged 6-15. Age-normalized measures could not be computed because normative data are not available for the full range of allophonic variants of /r/ represented, which have distinct acoustic characteristics (e.g. [5]).

[3] "Correct /r/"/"Not a correct /r/" were used in place of a phoneme pair such as /r/-/w/ because children's misarticulated /r/ sounds seldom have phonetic properties consistent with true /w/, and this could distort listener responses on the "incorrect" end of the continuum.