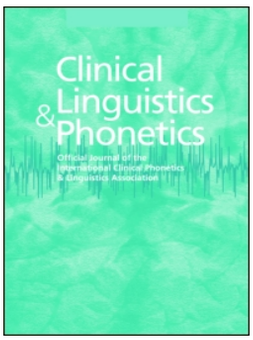

Clinical Linguistics & Phonetics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iclp20

Perceptual measurement of articulatory goodness in young children: Relationships with age, speech sound acquisition, and intelligibility

Ashley Sakash, Tristan J. Mahr & Katherine C. Hustad

To cite this article: Ashley Sakash, Tristan J. Mahr & Katherine C. Hustad (2023): Perceptual measurement of articulatory goodness in young children: Relationships with age, speech sound acquisition, and intelligibility, Clinical Linguistics & Phonetics, DOI: 10.1080/02699206.2022.2150893

To link to this article: https://doi.org/10.1080/02699206.2022.2150893

Published online: 02 Jan 2023.

CLINICAL LINGUISTICS & PHONETICS
https://doi.org/10.1080/02699206.2022.2150893

Taylor & Francis
Taylor & Francis Group

Perceptual measurement of articulatory goodness in young children: Relationships with age, speech sound acquisition, and intelligibility

Ashley Sakash[a], Tristan J. Mahr[a], and Katherine C. Hustad[a,b]

[a]Waisman Center, University of Wisconsin-Madison, Madison, Wisconsin, USA; [b]Department of Communication Sciences & Disorders, University of Wisconsin-Madison, Madison, Wisconsin, USA

ABSTRACT

Speech language pathologists regularly use perceptual methods in clinical practice to assess children's speech. In this study, we examined relationships between measures of speech intelligibility, clinical articulation test results, age, and perceptual ratings of articulatory goodness for children. We also examined the extent to which established measures of intelligibility and clinical articulation test results predicted articulatory goodness ratings, and whether goodness ratings were influenced by intelligibility. A sample of 164 (30–47 months) typically developing children provided speech samples and completed a standardised articulation test. Single word intelligibility scores and ratings of articulatory goodness were gathered from 328 naïve listeners; scores on a standardised articulation test were obtained from each child. Bivariate Pearson correlation, linear regression, and linear mixed effects modelling were used for analysis. Results showed that articulatory goodness ratings had the highest correlation with intelligibility, followed by age, followed by articulation score. Age and clinical articulation scores were both significant predictors of goodness ratings, but articulation scores made only a small contribution to prediction. Articulatory goodness ratings were substantially lower for unintelligible words compared to intelligible words, but articulatory goodness scores increased with age at the same rate for unintelligible and intelligible words. Perceptual ratings of articulatory goodness are sensitive to developmental changes in speech production (regardless of intelligibility) and yield a different kind of information than clinical articulation scores from standardised measures.

ARTICLE HISTORY
Received 21 October 2021
Revised 29 September 2022
Accepted 13 November 2022

KEYWORDS
Speech development; articulation; speech intelligibility; motor speech disorders; assessment of speech disorders

Introduction

Articulation and the acquisition of speech sounds are critical components of overall speech development. In typical development, speech sounds, which are building blocks for words, develop in a relatively predictable developmental sequence (Crowe & McLeod, 2020). The study of speech sound development has longstanding historical roots, dating back to the early 1900s (for example, (Conradi, 1904; Poole, 1934)). Methods for the examination of speech sounds, even today, employ the same approach as early work, namely an expert listener makes perceptual judgements of production features of individual sounds. Productions can be phonetically transcribed using broad or narrow levels of detail from

CONTACT Katherine C. Hustad kchustad@wisc.edu Waisman Center, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53706, USA

the International Phonetic Alphabet, binary scoring (incorrect/correct) can be applied, and/or productions can be judged as omission, substitution or distortion of the target speech sound. When standardised clinical articulation tests are administered, an aggregate score that can be compared against age norms is generated. Studies using perceptual methods suggest that, in the English language, most children can produce all but 4 consonants by five years of age (Crowe & McLeod, 2020) in monosyllabic words. A cross-linguistic review of 27 languages also found that on average, five-year-old children have acquired most consonants (McLeod & Crowe, 2018). It is noteworthy that about 10% of children who are otherwise typically developing have a developmental speech sound disorder (Bishop, 2010). Many other children experience disruptions in speech development due to a range of different neurodevelopmental disorders, genetic syndromes, or hearing impairment.

Speech intelligibility, defined as how well a speaker's acoustic signal can be accurately recovered by a listener, is a multidimensional construct influenced by many variables (Kent et al., 1989; Yorkston & Beukelman, 1980). Speech intelligibility is measured at a lexical level using methods such as orthographic transcription by naïve listeners in which words are scored as correct or incorrect based on their lexical match to the target attempted by the child. Here, productions with speech sound errors can be scored as intelligible if a listener assigns the correct word. Intelligibility can also be measured at the utterance level using methods such as proportion of complete and intelligible utterances via language sample analysis. The former method is considered by some to be a gold standard (Connolly, 1986; Kent, 1993; Kent et al., 1994; Lagerberg et al., 2014). Like speech sound acquisition, speech intelligibility increases developmentally over a protracted period of time. Large scale efforts to characterise the scope of speech intelligibility development have only recently emerged. In 2020, Hustad and colleagues found that typically developing children make steady gains in intelligibility development from 30 to 47 months (Hustad et al., 2020). They found that on average children were 46% intelligible during single word productions at 30 months. By 36 months of age, average single word intelligibility increased to 55%. By 42 months, single word intelligibility averaged 65%, and by 47 months, it was 70%. Subsequent work has shown that intelligibility development continues to advance through 9 years of age (Hustad et al., 2021). A key result from intelligibility studies is that there is considerable variability among children of the same age, especially younger children, but variability reduces with age. This variability can make differentiation between typical and atypical development difficult, particularly for younger children.

It is clear that speech intelligibility development has a different time course than speech sound acquisition. A definitive one-to-one relationship between speech sound accuracy and speech intelligibility has not been established, due in great part to methodological limitations with available measurement approaches. Studies examining the contributions of different speech sounds to intelligibility have been very limited, and have generally shown that clinical articulation scores (reflecting standardised inventories of sounds mastered) have a modest correlation with speech intelligibility in children with different speech disorder conditions (Ertmer, 2010; Natzke et al., 2020). One explanation is that speech sounds do not require perfect production in order for listeners to identify the words they comprise. Hence, a child can be understood without having all speech sounds mastered. Moreover, some developmental features like phonological processes can affect a child's speech in systematic and predictable ways.

The use of ratings scales for quantifying speech sounds in a more refined manner than simple binary or categorical judgements has received increasing attention as a way to characterise speech features. In recent studies, Munson and colleagues compared different scales to rate specific sounds produced by typically developing children (Munson & Urberg Carlson, 2016; Munson et al., 2017; Schellinger et al., 2017). Collectively, these studies showed that although there was variability, continuous ratings of the goodness of children's consonant productions were correlated with acoustic measures. However, it is unknown how this type of rating relates to measures of speech sound development such as clinical articulation scores and how such ratings might be influenced by the intelligibility of individual words being rated. Further, we do not know know how more holistic ratings of articulation goodness at the word level might be related to the aforementioned measures.

In the present study, we examine ratings by naive human listeners of the goodness of children's articulation at the level of single word productions, irrespective of the accuracy of individual speech sounds or the intelligibility of the word. Our ratings of the goodness of word production differs from speech sound accuracy as measured by articulation tests that use binary correct/incorrect production scoring. It also differs from orthographic transcription intelligibility measures that score words as correct/incorrect. Specifically, our goal was to explore how listener perception of children's word-level articulation goodness relates to other well established perceptually-based clinical constructs. We sought first to describe relationships among variables (e.g. word-level perceptual articulatory goodness ratings, scores from a standardised articulation test, speech intelligibility scores, and child age). We were especially interested in determining if articulation scores from standardised clinical testing as obtained from expert judges (certified speech language pathologists) predicted goodness ratings made by naïve listeners for individual words. If clinical articulation scores are highly predictive of articulatory goodness ratings after age effects are accounted for, this would suggest that the two measures are capturing the same construct and that goodness ratings could potentially be indicative of speech sound accuracy as a screening tool for speech disorders or for measuring change in speech production. If, however, clinical articulation scores are not highly predictive of articulatory goodness ratings after age effects are accounted for, it would suggest that goodness ratings by naïve listeners are not sensitive to speech sound accuracy, but may capture a different construct. We also sought to examine the extent to which articulatory goodness ratings change with age for words that are intelligible versus unintelligible and whether the effect of age differed for intelligible versus unintelligible words. If goodness ratings change with child age similarly for intelligible and unintelligible words, it may suggest that listeners are sensitive to developmental speech advancements that are independent of intelligibility. Such a finding would support the validity of perceptual ratings of articulatory goodness as a means of characterising developmental changes in speech, and may have the potential to be useful clinically for characterising change in children with speech disorders, even if that change does not involve a perceptual categorical threshold shift for speech sounds.

In the present study, we asked the following research questions:

(1) What are the relationships among perceptual articulatory goodness ratings, clinical articulation scores, age, and single-word intelligibility scores?

(2) Do articulatory goodness ratings show change with age, specifically for words that listeners were able to understand and for words that listeners were not able to understand?

Materials and methods

Participants

Approval for this study was granted by the University of Wisconsin-Madison Institutional Review Board. Informed consent was obtained for all participants.

A total of 164 typically developing (TD) children (92 females, 72 males) between 30–47 months contributed speech samples for this study. Children were recruited from the local community in Madison, WI. All participants were required to speak American English as their native language. In addition, they were required to have typical speech, language, and hearing development as determined by the following: (1) parent report of normal hearing and either pure-tone hearing screening or distortion product otoacoustic emission screening; (2) speech sound production within normal limits as indicated by articulation scores on the Arizona Articulatory Proficiency Scale – Third Edition (AAPS-3; (Fudala, 2001); and (3) language skills within normal limits as indicated by the Preschool Language Scale – Fifth Edition screening test (PLS-5; (Zimmerman et al., 2012)). Two children (1 female, 1 male) were excluded from analyses for having incomplete speech samples or incomplete responses from listeners. Children were stratified by age such that there were 56 children between 30 and 35 months of age; 49 children between 36 and 41 months; and 57 children between 42 and 47 months.

328 healthy adults participated as listeners. Listeners were primarily undergraduate students recruited from the university through public postings and social media. All listeners met the following inclusion criteria: (1) be between 18 and 45 years of age; (2) be a native speaker of American English; (3) have no identified language, learning or cognitive disabilities per self-report; (4) pass a pure tone hearing screening administered via headphones at 25 dB HL at 250, 500, 1000, 4000, and 6000 Hz in both ears; and (5) have no more than incidental experience listening to or communicating with persons having communication disorders. Listeners were 238 females and 90 males; their mean age was 20.5 (SD = 3.6) years.

Child and listener measures

Children completed a standard research protocol administered by a licenced speech language pathologist during a single in person visit to the laboratory. Details regarding data collection procedures from children and from adult listeners are provided in Hustad et al. (2020). Of interest to the current study were 1) single-word intelligibility scores obtained from naïve listeners based on elicited speech samples produced by children; 2) ratings of articulatory goodness made by naïve listeners, and 3) child clinical articulation scores from the AAPS-3. A brief description of these measures is provided below.

Acquisition of speech samples from children

Each child completed a structured imitative speaking task during their visit to the laboratory. During this task, children were audio-recorded while repeating a list of sentences from

the Test of Children's Speech (TOCS+) (Hodge & Gotzke, 2014, 2014; Hodge et al., 2007), a developmentally appropriate set of speech stimuli that systematically vary in length. Eliciting the same set of stimuli from children ensured that intelligibility scores reflected listeners' perception of target words relative to a known set of items. Stimuli consisted of sentences that ranged from 1 to 7 words. There were 38 single words from the TOCS+, which were repeated in isolation, and 10 sentences each for 2-words, 3-words, 4-words, 5-words, and 6-words, for a total of 60 sentences. In the current study, only data from the child's single word productions were analysed. Characteristics of the stimulus words are described in a previous paper (see Hustad et al., 2019)

The speaking task occurred in a sound-attenuating suite. Speech samples from children were recorded using a digital audio recorder (Marantz PMD 570, D & M Holdings Inc., Tokyo, Japan) at a 44.1-kHz sampling rate (16-bit quantisation). A condenser studio microphone (Audio-Technica AT4040, Audio-Technica U.S., Inc., Stow, OH) was positioned next to each child using a floor stand and was located approximately 18 inches from the child's mouth. Throughout the session, the level of the signal was monitored and adjusted on a mixer (Mackie 1202 VLZ, Mackie Designs Inc., Woodinville, WA) to obtain optimised recordings and to avoid peak clipping.

Productions of the target stimulus items were elicited from children using a 12.9-inch Apple iPad Pro. For each stimulus item, children were presented with an image, an orthographic representation, and a recorded production of the item. Children were instructed to repeat what they heard immediately following the recorded adult production. All child productions were monitored in real time by a student research assistant to ensure that speech samples were free from overlap with the model and free from extraneous noises. Repetitions were requested from the participant if needed.

Single-word intelligibility scores

Digital audio recordings from children were transferred to a desktop computer and edited to remove extraneous noises and the pre-recorded adult model. Individual files were then created for each stimulus item produced by each child. Audio samples were peak amplitude normalised to ensure that maximum loudness levels were the same across children and stimulus items while preserving the amplitude contours of the original productions.

Using in-house software, speech stimuli were presented to listeners via a desktop computer in a sound-attenuating suite. The external speaker on the desktop was calibrated to ensure the peak output level was 75 dB SPL from where listeners were seated. During the task, each listener was presented with all speech stimuli spoken by a single child. The in-house software randomised the presentation order of stimulus items for each listener. Listeners were instructed to provide orthographic transcriptions of each word (i.e. to type what they thought the child had said). Two unique listeners provided transcriptions for all stimuli produced by one child.

Listener orthographic transcriptions were scored as either correct or incorrect based on whether they matched target transcription (of child productions) phonemically. Misspellings and homonyms were accepted as correct if all phonemes in the transcription matched the target. The total number of words transcribed correctly by each of the two listeners per child was added together, then divided by the total number of words possible (across the two listeners) and multiplied by 100 to yield a percent intelligibility score for each child. We calculated the interrater reliability of intelligibility scores with the intraclass

correlation coefficient (ICC). We used an average-score, absolute agreement, one-way random effects model, and we found high agreement among average ratings, ICC(2) = .94, 95% CI = [.92, .96].

Intelligibility data and associated growth curves for intelligibility from the same children have previously been published (Hustad et al., 2020). In the present paper, we consider additional variables that were not previously analysed, including child articulation test scores (standardised articulation test results) and perceptual articulatory goodness ratings by listeners, both detailed below.

Perceptual articulatory goodness ratings by listeners

Listeners made ratings of articulatory goodness for each single word production that they transcribed in the intelligibility listening task. Specifically, after listeners typed their response for a given word-level stimulus item, they were asked to respond to the prompt 'rate this child's articulation' using a continuous 7-point sliding scale (anchors were: 1 = very poor and 7 = very good). Listeners were able to move a slider anywhere on the 7-point scale, and a continuous numeric result was saved for each rating. Listeners were asked to rate articulation of each word they heard. It is important to note that the articulatory goodness rating always occurred after the child's production of the word and after the listener transcribed the word. Listeners did not have knowledge of whether their word transcriptions were correct or incorrect when making goodness ratings.

Goodness ratings for all words were pooled within each individual listener and child and divided by the total number of words rated, to obtain an average articulatory goodness score for each child and listener. Ratings from each of the two listeners per child were then further averaged to yield one goodness score per child. We also examined goodness ratings in a disaggregated form by taking goodness ratings for words that each listener identified correctly in the intelligibility task and goodness ratings for words that each listener identified incorrectly in the intelligibility task. We then examined differences in goodness ratings for those words that both listeners transcribed correctly versus those words that both listeners transcribed incorrectly.

Child clinical articulation scores

The AAPS-3 (Fudala, 2001) was administered as part of the research protocol and administration was performed as outlined by the testing manual. Scoring consisted of making binary (correct/incorrect) judgements of each speech sound, also as outlined by the manual. Total (raw) scores were calculated and used for each child for articulation scores in the current study.

Statistical Analysis

We first examined the relationships among child-level variables using correlations including between-listener correlations. We then modelled overall articulatory goodness using linear regression. The outcome variable for this model was each child's mean articulatory goodness rating, averaging over ratings from two listeners. Predictor variables were the child's age in months and the child's total score on the AAPS-3 (clinical articulation measure).

We next looked at the differential effects of intelligibility on perceptual articulatory goodness ratings at the item level. That is, we examined whether perceptual articulatory goodness differed for intelligible versus unintelligible words. The outcome measure here was average articulatory goodness rating for each child on each item, so we used a linear mixed effects model to handle repeated measurements at the child level and item level. Each child produced items from a word list and had their productions transcribed by two listeners. We grouped items based on the intelligibility of the children across listeners. Items where both listeners correctly transcribed the word were *intelligible* and items where both listeners incorrectly transcribed the word were *unintelligible* (e.g. the listener did not type the word spoken by the child). We omitted items where one listener correctly transcribed the word and the other listener did not (19.0% of items). In the modelled dataset, there were 4,811 observations (words with mean goodness ratings) for 162 children and 21–36 items per child.

The model's fixed effects included age (centred at 36 months), intelligibility (intelligible vs. unintelligible), and the age-by-intelligibility interaction. The two main effects estimate how articulatory goodness changes with age and changes between intelligible versus unintelligible productions. The interaction term then estimates whether the effect of age on goodness ratings differs for intelligible versus unintelligible productions. The model included by-item and by-child random intercepts and by-item and by-child random slopes for intelligibility. Statistically, the random effects adjust the average goodness ratings of individual items and children in intelligible and unintelligible productions, so these effects allow items to vary in average easiness and allow children to vary in ability. We also fitted a fully disaggregated model with by-listener random intercepts added, and the estimates from this alternative model are reported after the main results.

Analyses were carried out the R programming language (vers. 4.2.1; (R Core Team, 2019)). Mixed models were estimated using the lme4 package (vers. 1.1.30; (Bates et al., 2015)). We estimated marginal means and contrasts using the emmeans package (vers. 1.8.1.1; (Lenth, 2019)).

Results

Research question 1

We asked: What are the relationships among perceptual articulatory goodness, clinical articulation scores, age, and single-word intelligibility scores? How well do clinical articulation scores predict goodness ratings? Correlations between each pair of variables were examined. Scatterplots are shown in Figure 1 with Pearson correlation coefficients reported in each panel. Results showed that all variables were significantly correlated ($p < 0.01$), but that the correlations between goodness ratings and intelligibility, intelligibility and clinical articulation scores, intelligibility and age, and goodness ratings and age were the strongest. The correlation between clinical articulation test scores and articulatory goodness ratings was the weakest, $r(160) = .32$.

We modelled average perceptual articulatory goodness using stepwise regression. Table 1 reports coefficients from each linear model. First, we regressed each child's average articulatory goodness rating onto their age to account for expected developmental changes. There was a significant effect of age such that a 1-month increase in

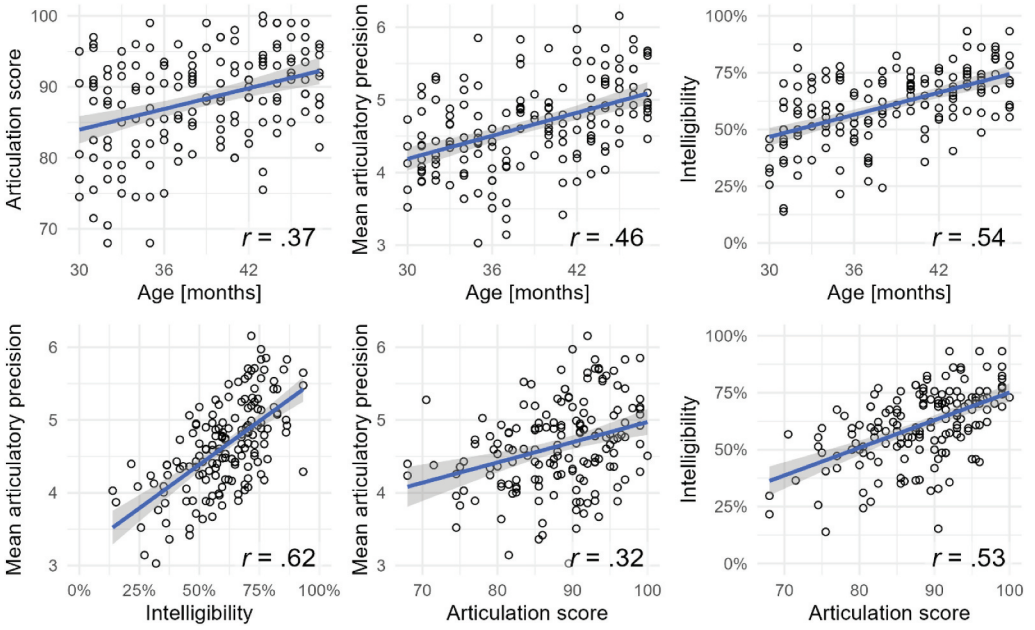

Figure 1. Scatterplots of the key measures for this study. N = 162 children in each panel. Lines and error bands are the regression lines and 95% confidence intervals for the predicted means.

Table 1. Model coefficients from linear models estimating children's average articulatory goodness rating from other child-level predictors. Clinical articulation scores were centred at the sample mean (88 points).

	Age				+ Artic Score			
	Est	SE	*t*	*p*	Est	SE	*t*	*p*
Intercept (36 months, mean artic)	4.505	0.047	95.4	<.001	4.522	0.047	95.9	< .001
Age (+1 month)	0.053	0.008	6.6	<.001	0.046	0.009	5.3	< .001
Artic Score (+1 point)					0.015	0.006	2.3	.022

age predicted an increase in articulatory goodness of 0.05, SE = 0.008, 95% CI [0.04, 0.07], $p < 0.001$, adj. $R^2 = .207$. We then added clinical articulation scores to the model, and there was a small but significant effect of articulation score on articulatory goodness over and above age, $b = 0.02$, SE = 0.01, 95% CI [0.00, 0.03], $p = 0.022$, adj. $R^2 = .228$.

Research question 2

We asked: Do goodness ratings show developmental change with age, specifically for words that listeners were able to understand and for words that listeners were not able to understand? Are the effects of age different for intelligible versus unintelligible words? Model coefficients are reported in Table 2, and the regression lines for the model's fixed effects are visualised in Figure 2. For intelligible words, the expected (i.e. model-estimated average) articulatory goodness rating for an average child and average word at 36 months of age was 5.17, SE = 0.08, 95% CI [5.02, 5.32]. For unintelligible words, the corresponding expected

Table 2. Model coefficients from the mixed effects model estimating children's articulatory goodness rating from age, item intelligibility and the age by intelligibility interaction. The top rows summarise the model's fixed effects which estimate the expected goodness rating for an average child and average item, and the bottom rows summarise the model's varying (random) effects which estimate the variability in goodness ratings among children and items and the variability of intelligibility effects among children.

	Est	SE	95% CI
Intercept (36 months, intelligible)	5.169	0.078	[5.016, 5.322]
Age (+1 month)	0.027	0.008	[0.012, 0.042]
Intelligibility (intelligible → unintelligible)	−1.409	0.097	[−1.600, −1.219]
Age × Intelligibility	0.000	0.008	[−0.016, 0.015]
Child: sd(Intercept)	0.441		
Child: sd(Intelligibility)	0.296		
Child: cor(Intercept, Intelligibility)	−0.187		
Item: sd(Intercept)	0.382		
Item: sd(Intelligibility)	0.511		
Item: cor(Intercept, Intelligibility)	−0.374		
sd(Residual)	1.021		

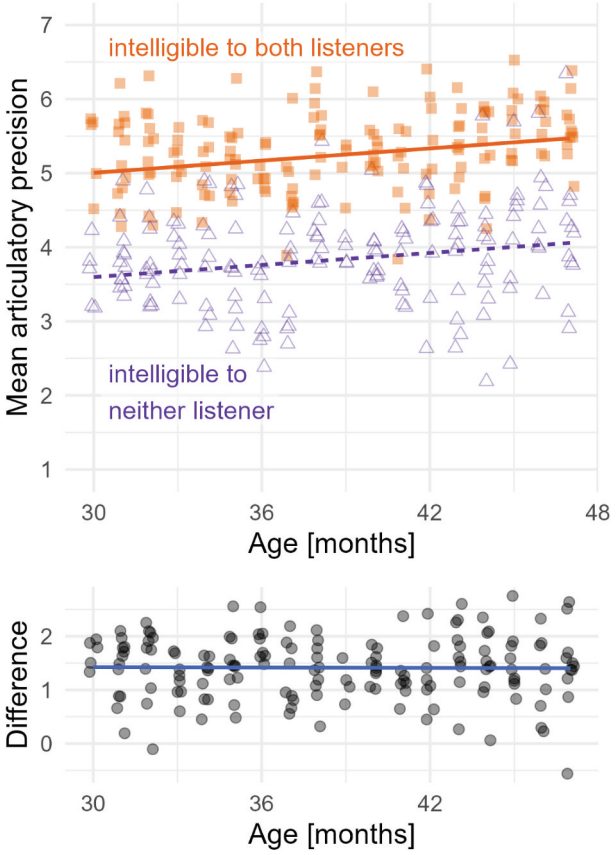

Figure 2. Top panel: Estimated articulatory goodness rating for intelligible (squares) versus unintelligible (triangles) items by age. Each point represents one child's average articulatory goodness rating for either intelligible or unintelligible items. Regression lines reflect model results: a clear intelligibility effect (gap between the lines), a clear age effect (lines increasing with age), the intelligibility effect being larger than the age effect (gap is larger than amount either line rises), and no clear age-by-intelligibility interaction (parallel lines). Bottom panel: Observed differences in articulatory goodness rating between intelligible and unintelligible productions for each child. The flat linear regression line also shows that there was not an age-by-intelligibility interaction; the average difference does not change with age.

articulatory goodness rating was 3.76, SE = 0.10, 95% CI = [3.57, 3.95]. Intelligible productions received significantly higher perceptual articulatory goodness ratings on average than unintelligible ones, $b = 1.41$, SE = 0.10, 95% CI [1.22, 1.60], $p < 0.001$.

Although there were significant main effects of age and intelligibility on perceptual articulatory goodness ratings, there was a statistically significant effect of age such that goodness ratings on productions from older children were higher on average than goodness ratings on productions from younger children. A one-month increase in age predicted an increase in average goodness rating of 0.03 points, SE = 0.01, 95% CI [0.01, 0.04], $p < 0.001$. Thus, a one-year increase in age would predict an increase in expected articulatory goodness of 0.33 points, SE = 0.09, 95% CI [0.15, 0.51]. We note that the magnitude of this age effect (.33 increase for one year of age) was much smaller than the intelligibility effect (1.41 increase for intelligible over unintelligible items).

There was not a statistically significant age-by-intelligibility interaction, $b = 0.000$, SE = 0.008, 95% CI [−0.016, 0.015], $p = 0.97$. Thus, there was not a significant difference in the month-over-month change in average articulatory goodness rating for intelligible versus unintelligible transcriptions.

In the preceding analysis, we used the average perceptual articulatory goodness rating for items in which two listeners either both correctly transcribed the word or both incorrectly transcribed the word. As noted earlier, 19.0% of ratings were excluded because the production was intelligible to one listener but not the other listener. We repeated the preceding analysis but used disaggregated ratings, included the previously excluded ratings, and expanded the model's random effects to include by-listener random intercepts and random slopes for intelligibility. The model's fixed effects estimates matched the previous estimates in direction, magnitude and significance. The estimated perceptual articulatory goodness rating for an average child on an average intelligible word by an average listener at 36 months of age was 4.98, 95% CI [4.82, 5.14]. For an unintelligible word, the expected rating decreased by 1.26, [1.07, 1.45]. For a 1-month increase in age, expected articulatory goodness increased by 0.04, [0.02, 0.05], and there was not a significant age-by-intelligibility interaction, $b = -0.01$, [−0.03, 0.00], $p = 0.08$.

Discussion

In this study, we examined relationships between speech intelligibility, clinical articulation test results, and perceptual ratings of articulatory goodness. Although intelligibility and clinical articulation test results are well-established clinical measures in speech language pathology, the construct of articulatory goodness, as we use it in this study, has not been examined previously. We measured articulatory goodness in a lexical intelligibility task: listeners orthographically transcribed a series of children's single-word productions, and after each transcription, they rated the goodness of a child's articulation. Goodness ratings are meant to capture impressions of articulation that are finer grained than intelligibility; thus, we hypothesised that they might be strongly related to speech sound production as measured by clinical articulation test scores. However, we also considered the possibility that goodness ratings could reflect listeners' certainty of word identity as captured by intelligibility measures.

Our results revealed three key findings. First, we found that perceptual articulatory goodness ratings had moderate to strong correlations with age, intelligibility, and clinical articulation test scores. Second, we found that clinical articulation test scores made a very

small contribution to prediction of goodness ratings after accounting for age effects. Third, when we examined articulatory goodness ratings for intelligible versus unintelligible words, we found a clear developmental effect such that goodness ratings increased by the same amount for intelligible and unintelligible words. Collectively, these findings suggest that perceptual ratings of articulatory goodness are sensitive to developmental changes in speech production, regardless of the intelligibility of words. Findings are discussed in detail below.

Relationships between articulatory goodness ratings and other measures

Results of this study showed significant correlations between all variables of interest; however, the strength of those correlations varied. Given the nature of speech development it might be expected that age would be the strongest correlation observed with each speech variable. However, that was not the case. Although age was significantly correlated with intelligibility, articulatory goodness ratings, and clinical articulation scores, our findings showed that correlations between intelligibility and each of the other variables were generally stronger than correlations between age and the other variables. However, the correlation between intelligibility and age, although moderate in absolute strength, was one of the strongest observed in this study ($r = .53$) - not surprisingly, children became more intelligible as they got older, and intelligibility was reliably correlated with our articulation measures. One notable feature, established in our earlier work (Hustad et al., 2020) and shown in Figure 1, is that intelligibility is highly variable between individual children within age, especially for younger children. In the present study, there was a similar strong positive correlation between age and articulatory goodness ratings ($r = .47$). That is, single words produced by younger children received lower (or poorer) perceptual articulatory goodness ratings by listeners and single words produced by older children received higher (or better) perceptual articulatory goodness ratings by listeners. Listeners in this study did not know children's ages and thus were not biased by age expectations. In addition, listeners heard only one child, so their articulatory goodness ratings were not made relative to other children. Articulatory goodness ratings showed considerable variability by age, similar to our intelligibility findings, and this variability was maintained over the age range of this study. The correlation between clinical articulation test scores (note that these were raw scores and not standardised scores) and age, though significant, was weak ($r = .36$). Articulation scores also showed the greatest amount of within age variability across all ages. This finding is surprising given the well-established effects of age on articulation development. One explanation is that standard articulation tests are designed to identify children who are not meeting age expectations for speech sound acquisition and therefore may benefit from speech therapy. In the present study, we included only typically developing children with articulation test scores within normal limits for age; however, some children acquire speech sounds prior to age expectations and are thus advanced in their articulation development. The variability observed in this study was likely a reflection of performance at and above average within typical development. We expect that stronger correlations between age and clinical articulation test scores would be observed in a larger sample of children that included clinical populations.

The correlation between single-word intelligibility scores and clinical articulation scores was moderate in strength ($r = .54$) in this study. This finding is consistent with other studies showing that transcription intelligibility is not closely related to speech sound accuracy

(Ertmer, 2010; Natzke et al., 2020). Measurement differences between intelligibility and AAPS-3 scoring may be one explanation for this finding. For intelligibility scores, transcription is orthographic, measured at the word level, and made by unfamiliar listeners. For the AAPS-3, transcription is phonetic and is made by speech-language therapists. We would expect these constructs to be closely related, however, words can be intelligible even when speech sounds are not produced correctly. Thus, there is not a clear one to one relationship between speech sound accuracy and intelligibility.

Not surprisingly, we found a strong positive association between single-word intelligibility scores and articulatory goodness ratings ($r = .61$). As indicated in the methods, intelligibility and goodness ratings were obtained during the same listening task in close temporal proximity by the same listener (i.e. listeners heard the word, transcribed it for intelligibility and then made ratings of articulatory goodness). We would clearly expect these variables to be intercorrelated because of the nature of the task. However, because listeners did not have explicit knowledge of whether they correctly understood the child, we consider the impact of intelligibility on goodness ratings in a separate analysis, as discussed below.

Prediction of goodness ratings

With regard to prediction of perceptual articulatory goodness ratings, we found that age was a larger contributor than clinical articulation scores. The predictive nature of age is not surprising given that development is a critical variable relative to speech for children in the age range of this study. Children are rapidly acquiring speech sounds, refining their speech motor control, and improving their speech intelligibility; thus we would expect articulatory goodness to improve reliably with age. This finding was confirmed in our analysis, revealing that overall, goodness ratings improved by .32 points on a 1–7 point scale per year of age across intelligible and unintelligible words.

We found that clinical articulation scores resulted in a very small improvement in the prediction of articulatory goodness ratings over age, but the magnitude of that improvement, although statistically significant, was not clinically meaningful. This finding was counter to our expectation that articulatory goodness ratings would be closely tied to overall clinical articulation test results as indicated by standardised articulation test scores. One explanation for the weak predictive value of clinical articulation scores (and also the weak correlation between goodness ratings and clinical articulation scores) is that listeners likely focused their ratings on goodness at the word level rather than at the speech sound level measured by clinical articulation scores. Another possible explanation is that the listener's goodness rating of a word was biased by their perception of whether they thought they understood the child. Future studies should explore how this relationship might differ when the listener is not naïve to the actual word the child is saying. A third possible explanation is that listeners may not have penalised children for typical developmental speech sound errors when making their ratings of articulatory goodness. For example, a child might make a common substitution error but still be rated well for articulatory goodness. It is important to note that although all children in the current study had articulation scores that were within an age-appropriate range as determined by the AAPS-3, they still may have had developmental or even non-developmental speech sound errors. Future research should explore the nature of speech sound errors and their impact on articulatory goodness ratings.

In this analysis, we considered aggregate scores for articulatory goodness over all single words and aggregate scores across speech sounds on a clinical articulation test. Examination of articulatory goodness ratings at the individual speech sound level along with measures of individual speech sound development may yield different results and indicate a stronger relationship between variables.

Effects of intelligible vs. unintelligible words on articulatory goodness ratings

Listener ratings of articulatory goodness in words that were intelligible were consistently higher than ratings of words that were unintelligible. The magnitude of this difference was 1.33 points on a 1–7 point scale. Thus, articulation of words that were intelligible were consistently rated 20% higher than those that were not intelligible regardless of the child's age. Listeners were not given explicit feedback regarding their performance on intelligibility tasks, but it is likely that they did have some insight into whether they had identified words correctly for many productions.

A key finding of this study was that articulatory goodness ratings increased with age for unintelligible words at the same rate as for intelligible words (.32 articulatory goodness points per year of age: 4.5% increase in articulatory goodness ratings per year). That is, listeners were sensitive to changes in goodness that occurred with age and rated older children as having higher articulatory goodness than younger children, even for words that they did not understand. The fact that even unintelligible words showed this constant articulatory goodness gain indicates that goodness ratings were not fully dependent on intelligibility. It also suggests that ratings of articulatory goodness are sensitive to fine grained developmental change in speech.

Limitations and future directions

There are several limitations to the current study. First, all participants were typically developing and had speech sound production within normal limits. We chose this population because it is a logical first step in the examination of perceptual ratings of articulatory goodness and their relationship to other measures of speech development in children. Future studies should examine articulatory goodness ratings in concert with other speech measures in disordered populations of children with a broader range of speech abilities to determine how well listener ratings of articulatory goodness may differentiate between disordered and typical populations and among children with different disorder conditions. Additionally, there are other variables besides the ones included in the current study that could affect listener's ratings of articulatory goodness. Future studies should examine other factors that may be related to ratings of articulatory goodness, such as type of error (distortion versus substitution versus omission), speaking rate, neighbourhood density of the target word, and degree of familiarity of the words to the speaker. Second, productions of children's speech were elicited using an imitation paradigm with an adult model. This type of speech elicitation task may have inadvertently influenced the child's production of words. Future studies should examine articulatory goodness ratings for different speech elicitation tasks. Lastly, ratings of articulatory goodness and intelligibility were obtained by the same

listener during the same listening task. As a result, ratings of articulatory goodness and intelligibility scores were not independent of one another. Future research should separate these tasks and use multiple listeners with a between-subjects design to determine if the same findings are replicated to further validate listener ratings of articulatory goodness.

In summary, findings of this study suggest that listener ratings of articulatory goodness are developmentally sensitive, regardless of a child's speech intelligibility. However, these ratings are not strongly related to clinical articulation scores and appear to capture change in speech development irrespective of intelligibility. Listener ratings of this kind are inexpensive, do not require professional training/expertise to make, and may provide useful information regarding the quality of a child's speech. These data may be a useful complement to articulation test scores and speech intelligibility for describing the speech of children. Perceptual articulatory goodness ratings may also have clinical utility in measuring advances in the speech of children with neuromotor impairment who are often difficult to assess using traditional standardised measures.

Acknowledgements

We thank the children and their families who participated in this research, and the research staff, graduate students, and undergraduate students at the University of Wisconsin – Madison who assisted with data collection and data reduction.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by Grant R01DC015653 from the National Institute on Deafness and Other Communication Disorders, awarded to Katherine C. Hustad. Support was also provided by a core grant to the Waisman Center, U54 HD090256, from the National Institute of Child Health and Human Development.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bishop, D. V. (2010). Which neurodevelopmental disorders get researched and why? *PLoS One*, *5*(11), e15112. https://doi.org/10.1371/journal.pone.0015112

Connolly, J. H. (1986). Intelligibility: A linguistic view. *The British Journal of Disorders of Communication*, *21*(3), 371–376. https://doi.org/10.3109/13682828609019848

Conradi, E. (1904). Psychology and pathology of speech-development in the child. *The Pedagogical Seminary*, *11*(3), 328–380. https://doi.org/10.1080/08919402.1904.10534103

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology*, *29*(4), 2155–2169. https://doi.org/10.1044/2020_AJSLP-19-00168

Ertmer, D. J. (2010). Relationships between speech intelligibility and word articulation scores in children with hearing loss. *Journal of Speech, Language, and Hearing Research*, *53*(5), 1075–1086. https://doi.org/10.1044/1092-4388(2010/09-0250)

Fudala, J. B. (2001). *Arizona articulatory proficiency Scale-3* (3rd ed.). Western Psychological Services.

Hodge, M., & Gotzke, C. L. (2014). Construct-related validity of the TOCS measures: Comparison of intelligibility and speaking rate scores in children with and without speech disorders. *Journal of Communication Disorders*, *51*, 51–63. https://doi.org/10.1016/j.jcomdis.2014.06.007

Hodge, M., & Gotzke, C. L. (2014). Criterion-related validity of the Test of Children's Speech sentence intelligibility measure for children with cerebral palsy and dysarthria. *International Journal of Speech Language Pathology*, *16*(4), 417–426. https://doi.org/10.3109/17549507.2014.930174

Hodge, M., Gotzke, C. L., & Daniels, J. (2007). *TOCS+ intelligibility measures*. University of Alberta.

Hustad, K. C., Mahr, T., Natzke, P. E., & Rathouz, P. J. (2020). Development of speech intelligibility between 30 and 47 months in typically developing children: A cross-sectional study of growth. *Journal of Speech, Language, and Hearing Research*, *63*(6), 1675–1687. https://doi.org/10.1044/2020_JSLHR-20-00008

Hustad, K. C., Mahr, T. J., Natzke, P., & Rathouz, P. J. (2021). Speech development between 30 and 119 months in typical children I: Intelligibility growth curves for single-word and multiword productions. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3707–3719. https://doi.org/10.1044/2021_JSLHR-21-00142

Hustad, K. C., Sakash, A., Natzke, P., Broman, A. T., & Rathouz, P. J. (2019). Longitudinal growth in single word intelligibility in children with cerebral palsy from 24 to 96 months of age: Predicting later outcomes from early speech production. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1599–1613. https://doi.org/10.1044/2018_JSLHR-S-18-0319

Kent, R. (1993). Speech intelligibility and communicative competence in children. In A. P. Kaiser & D. B. Gray (Eds.), *Enhancing children's communication: Foundations for Intervention* (Vol. 2, pp. 223–239). Paul H. Brookes.

Kent, R., Miolo, G., & Bloedel, S. (1994). The intelligibility of children's speech: A review of evaluation procedures. *American Journal of Speech-Language Pathology*, *3*(2), 81–95. https://doi.org/10.1044/1058-0360.0302.81

Kent, R., Weismer, G., Kent, J., & Rosenbek, J. (1989). Toward phonetic intelligibility testing in dysarthria. *The Journal of Speech and Hearing Disorders*, *54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

Lagerberg, T. B., Åsberg, J., Hartelius, L., & Persson, C. (2014). Assessment of intelligibility using children's spontaneous speech: Methodological aspects. *International Journal of Language & Communication Disorders*, *49*(2), 228–239. https://doi.org/10.1111/1460-6984.12067

Lenth, R. (2019). *Emmeans: Estimated marginal means, a.k.a. Least-squares means*. (Version 1.4.0) https://CRAN.R-project.org/package=emmeans

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, *27*(4), 1546–1571. https://doi.org/10.1044/2018_AJSLP-17-0100

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics & Phonetics*, *31*(1), 56–79. https://doi.org/10.1080/02699206.2016.1233292

Munson, B., & Urberg Carlson, K. (2016). An exploration of methods for rating children's productions of sibilant fricatives. *Speech, Language and Hearing*, *19*(1), 36–45. https://doi.org/10.1080/2050571X.2015.1116154

Natzke, P., Sakash, A., Mahr, T., & Hustad, K. C. (2020). Measuring speech production development in children with cerebral palsy between 6 and 8 years of age: Relationships among measures. *Language, Speech, and Hearing Services in Schools*, *51*(3), 882–896. https://doi.org/10.1044/2020_LSHSS-19-00102

Poole, I. (1934). Genetic development of articulation of consonant sounds in speech. *The Elementary English Review*, *11*(6), 159–161.

R Core Team. (2019). *R: A language and environment for statistical computing*. R foundation for statistical computing. http://www.R-project.org/

Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of/s/and/θ: A comparative study of rating methods. *Clinical Linguistics & Phonetics*, *31*(1), 80–103. https://doi.org/10.1080/02699206.2016.1205665

Yorkston, K., & Beukelman, D. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communication Disorders*, *13*(1), 15–31. https://doi.org/10.1016/0021-9924(80)90018-0

Zimmerman, I., Steiner, V., & Pond, R. (2012). *Preschool Language Scale-5 Screening Test*. Psychological Corporation.