JSLHR

Research Article

Development of Speech Intelligibility Between 30 and 47 Months in Typically Developing Children: A Cross-Sectional Study of Growth

Katherine C. Hustad,[a,b] Tristan Mahr,[b] Phoebe E. M. Natzke,[b] and Paul J. Rathouz[c]

Purpose: We sought to establish normative growth curves for intelligibility development for the speech of typically developing children as revealed by objectively based orthographic transcription of elicited single-word and multiword utterances by naïve listeners. We also examined sex differences, and we compared differences between single-word and multiword intelligibility growth.
Method: One hundred sixty-four typically developing children (92 girls, 72 boys) contributed speech samples for this study. Children were between the ages of 30 and 47 months, and analyses examined 1-month age increments between these ages. Two different naïve listeners heard each child and made orthographic transcriptions of child-produced words and sentences (n = 328 listeners). Average intelligibility scores for single-word productions and multiword productions were modeled using linear regression, which estimated normal-model quantile age trajectories for single- and multiword utterances.
Results: We present growth curves showing steady linear change over time in 1-month increments from 30 to 47 months for 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Results showed that boys did not differ from girls and that, prior to 35 months of age, single words were more intelligible than multiword productions. Starting at 41 months of age, the reverse was true. Multiword intelligibility grew at a faster rate than single-word intelligibility.
Conclusions: Children make steady progress in intelligibility development through 47 months, and only a small number of children approach 100% intelligibility by this age. Intelligibility continues to develop past the fourth year of life. There is considerable variability among children with regard to intelligibility development.
Supplemental Material: https://doi.org/10.23641/asha.12330956

Speech intelligibility is a complex construct that plays a pivotal role in spoken communication. In one widely regarded definition, intelligibility describes the extent to which an acoustic signal, generated by a speaker, can be correctly recovered by a listener (Kent et al., 1989; Yorkston & Beukelman, 1980). In this framing, intelligibility is dyadic, with both listener and speaker making joint contributions. Intelligibility can be influenced by a wide range of variables including the length and nature of speech being produced (single words, individual sentences, narrative discourse, conversational discourse), familiarity of the listener with the speaker, expertise of the listener, and availability of visual information (Kent et al., 1994), to name a few. Kent et al. (1994) have suggested that "a particular talker has a range of intelligibility potentials, depending on listener familiarity, nature of the linguistic message, physical setting, motivation, effort level, and so on" (p. 81). Thus, any given measure of intelligibility is best considered a snapshot of performance under a specific set of circumstances, which must be interpreted cautiously. Nonetheless, there is a clinical need—particularly in pediatric and adult motor speech disorders where intelligibility is often a key focus of intervention—to provide an objective and reliable measure that can be used for clinical decision-making. Intelligibility measures are often used as a basis of comparison for documenting and monitoring change in

[a]Department of Communication Sciences and Disorders, University of Wisconsin–Madison
[b]Waisman Center, University of Wisconsin–Madison
[c]Department of Population Health, Dell Medical School, The University of Texas at Austin

Correspondence to Katherine C. Hustad: kchustad@wisc.edu

Editor-in-Chief: Bharath Chandrasekaran
Editor: Stephanie Anna Borrie

Received January 9, 2020
Revision received February 7, 2020
Accepted February 12, 2020
https://doi.org/10.1044/2020_JSLHR-20-00008

Disclosure: The authors have declared that no competing interests existed at the time of publication.

speech performance (Yorkston et al., 1999), as an index of severity of the speech disorder (Weismer & Martin, 1992), and as an indicator of functional ability (or disability) relative to "normal" performance (Yorkston et al., 1999).

Adult speakers without communication disorders are generally assumed to be fully intelligible (measured at or near 100%). However, for children, acquisition of intelligible speech is a protracted developmental process, beginning early in the first year of life with vocal play, babbling, and word approximations and continuing through childhood. Although it is clear that children acquire intelligible speech gradually, the precise course of development of intelligibility in typically developing children and the range of expected variability over the full course of development is not well understood. Existing studies have important methodological differences, including whether intelligibility was measured objectively or subjectively, whether listeners were "experts" (e.g., speech language pathologists or phoneticians), and the nature of speech material (elicited vs. spontaneous; single words vs. sentences vs. discourse or conversation). In a recent study examining relationships among different measures of intelligibility obtained from the same children, Natzke et al. (2020) found weak associations, suggesting that different measures of intelligibility are not reflective of one another. Here, we review extant studies examining intelligibility development in typically developing children to highlight the gaps in our current knowledge and the need for large-scale systematic studies that can form an evidence base for clinical decision-making regarding intelligibility.

Subjective Ratings of Intelligibility

Perhaps the most widely referenced screening guidelines and the most widely used clinical tools are based on subjective ratings of intelligibility made by parents or other familiar communication partners. Subjective ratings involve the assignment of some type of numerical rating (ordinal, interval, or categorical) representing personal perception of a speaker's intelligibility. For example, Coplan and Gleason (1988) identified cut-points for typical intelligibility development in children between 12 months and 5 years of age by asking parents of 235 children to make a forced-choice categorical rating regarding how much of their child's speech they thought a stranger would be able to understand. Response categories were: less than 50%, 50%, 75%, or almost all. From these parental responses, children were determined to be "passing" or "failing" at each of three intelligibility levels: 50%, 75%, and 100%. Logistic regression was used to develop curves for percentage of children passing each categorical intelligibility level as a function of age. Intelligibility guidelines, based on cut-points for 90% of the parental ratings in the sample, suggest the following: (a) At 22 months of age, nearly all children are 50% intelligible per parent report; (b) at 37 months of age, nearly all children are 75% intelligible per parent report; (c) at 47 months of age, nearly all children are 100% intelligible per parent report. It should be noted that these subjective parent ratings, which form the basis for the recommendations from this study, have not been validated with objective measures of intelligibility; thus, the extent to which parents rate their child accurately relative to some objective standard is unknown.

More recently, McLeod et al. (2015, 2012) developed the Intelligibility in Context Scale (ICS) to characterize intelligibility across different communication partners and contexts as revealed by parent ratings. The ICS asks parents to rate their perception of their child's intelligibility on a 5-point scale across seven different contexts. The one published normative study on the ICS examined 803 children from New South Wales, Australia, who were between the ages of 4;0 and 5;5 (years;months). Results showed that parents tended to rate their child's intelligibility as highest for themselves and lowest for strangers. In addition, girls had significantly higher ICS scores than boys. The effects of age on ICS scores was examined with children separated into 5-month bands. Results showed that children in the oldest age band had significantly better ICS scores than children in the two younger age bands. The authors note that effect sizes were very small for both sex and age. Studies of the ICS have examined its relationship with segmental measures such as percent consonants correct, percent vowels correct, and percent phonemes correct as scored on standardized tests. Results generally indicate that ICS ratings have weak correlations with these measures. The ICS has not been examined relative to other measures of intelligibility to our knowledge. It is widely used and has been translated to more than 60 languages (McLeod et al., 2015); however, normative data are very limited, and growth curves have not been developed.

Objective Ratings of Intelligibility

Objective ratings of intelligibility commonly involve transcription of speech, either using traditional orthography, broad phonetic transcription, or narrow phonetic transcription by one or more listeners. Objective ratings can be obtained and scored in different ways, but the primary commonality among measures is that a specific speech sample is directly analyzed to yield a quantitative score.

Intelligibility of Spontaneous Speech

One objective approach to quantifying intelligibility that has been reported in the literature involves the use of language transcripts, usually gathered for the purpose of characterizing expressive language. Specifically, the number of complete and intelligible utterances divided by the total number of utterances in a transcript yields a percentage of intelligible utterances (Binger et al., 2016; Rice et al., 2010; Yoder et al., 2016). In this context, the unit of measure is the full utterance, which is scored as intelligible or unintelligible. An intelligible utterance is one where the expert transcriber, who is typically allowed to listen to the sample or segments of it several times, is able to assign words to all of the units in an utterance. An unintelligible utterance is one where the expert transcriber is unable to assign words to at least one unit in the utterance. In a study designed to examine mean length of utterance development between 3 and 9 years of age in children with specific language impairment relative to typically developing peers, Rice et al. (2010) reported data on percentage intelligible utterances for the language transcripts of their study participants. Data from 136 typically developing children revealed that the percentage intelligible utterances had a mean of 86% at 2.5 years of age with a linear increase to about 95% by 9 years of age. Standard deviations were reported in this article and generally showed a trend of decreased variability with age.

Similarly, Flipsen (2006) examined intelligibility of conversational speech of 320 children between 3 and 8 years of age using speech samples that were transcribed by experts using narrow phonetic transcription for the purposes of characterizing segmental integrity. Intelligibility was determined by counting the number of words that were phonetically transcribed as fully intelligible divided by the total number of words produced. Results indicated that, at 3 years of age, mean intelligibility was 96%; by 8 years of age, mean intelligibility was 99%. As with utterance-level intelligibility obtained from language samples, variability decreased with age as observed by smaller ranges and smaller standard deviations.

In general, spontaneous speech and language samples are ecologically valid representations of children's functional communication, but common analysis procedures have some disadvantages for estimating intelligibility. First, typically only one individual transcribes a child's speech, and this person is an expert in child speech and/or language and is thus not representative of an everyday communication partner that a child might encounter. Second, transcribers are usually allowed to play back recorded speech samples multiple times—a convenience unavailable in real-life listening situations. Third, there is typically a communication partner (clinician or parent) who interacts with the child during a speech and language sample. The partner provides considerable contextual information and who may even gloss the child's utterances, aiding the transcriber in making sense of the child's speech. Finally, in a spontaneous speech sample context, the child's intended message is not definitively known a priori because it is spontaneously generated and the content of the speech/language sample is accepted as accurate if the transcriber assigned words (possibly the wrong words) to the child's spoken message. For these reasons, intelligibility measures obtained from speech and language samples may provide an inflated estimate of intelligibility.

Intelligibility of Elicited Utterances

In motor speech disorders research and clinical practice, intelligibility is commonly measured by having speakers produce a known corpus (words, utterances, read narrative passages) and then having naïve listeners write down what they thought the speaker said. Individual words produced by the speaker are scored as correct or incorrect based on whether orthographic representations by naïve listeners match the target words produced by the speaker. Intelligibility is calculated by dividing the number of words identified correctly by the number of words possible, multiplied by 100 (Tikofsky & Tikofsky, 1964; Yorkston & Beukelman, 1978, 1980). Standard clinical tools such as the Sentence Intelligibility Test (Yorkston et al., 1996), the Assessment of Intelligibility of Dysarthric Speech (Yorkston et al., 1984), and the Test of Children's Speech (TOCS; Hodge & Daniels, 2007) employ this type of transcription method. Clinically, transcription intelligibility is considered by some to provide a gold standard because listeners are forced to commit to paper their perceptions of what a speaker said, resulting in a quantitative measure of the integrity of the speech signal. There are several advantages to this approach including its objectivity, the use of naïve listeners who do not have expertise in listening to impaired speech (lending ecological validity to the measure), and the fact that lexical targets produced by speakers are definitively known; thus, listener transcriptions can be scored against a template. A recent small-scale cross-sectional study by Hodge and Gotzke (2014a) examining transcription intelligibility of 12 typically developing children at the ages of 3, 4, 5, and 6 years by naïve listeners suggests that intelligibility on the TOCS for single words increased linearly from 70% at age 3 years to 87% at age 6 years, and for sentences from 86% at age 3 years to 97% at age 6 years. Standard deviations were markedly reduced with age for both single-word and sentence intelligibility. These data provide an important starting point for understanding typical development, but data on additional children are needed to characterize the range of typical variability.

Although there are studies of intelligibility development in typically developing children, there is currently no solid evidence base to guide clinical decision making regarding age-appropriate expectations and benchmarks for typical versus atypical intelligibility development. This is a significant problem for early identification of children who are at risk for intelligibility impairment such as those with cerebral palsy, particularly for those children with mild or borderline intelligibility issues (Hodge & Gotzke, 2014a, 2014b; Hustad, Sakash, Natzke, et al., 2019). In other domains of development, normative growth curves allow practitioners to track development over time by centiles (e.g., 10th percentile, 50th percentile; WHO Multicentre Growth Reference Study Group & de Onis, 2006) and to specify where a child's performance falls relative to other children of the same age. This information allows quantification of whether a child is performing within the range of typically developing peers and how far from typical expectations a child is. It also allows determination of whether a child is maintaining a consistent rate of development over time.

In this study, our goal was to establish normative growth curves for intelligibility development for the speech of typically developing children as revealed by objectively based orthographic transcription of elicited sentences by naïve listeners. A key goal was to lay the foundation for the identification of age-specific cut-points for disordered versus typical intelligibility development. We focused on typically developing children between the ages of 30 and 47 months because this is a time frame of rapid development when children are acquiring speech sounds along with the ability to produce multiword utterances. We used a corpus of elicited words and sentences that were the same for each child so that we could make direct comparisons between children at different ages without the confounding variable of expressive language differences among children within and between age points. This corpus also allowed us to compare listener orthographic transcriptions against known target responses, ensuring that intelligibility scores were an accurate reflection of which target words were perceived correctly by listeners. We used two different naïve listeners for each child to reduce the potential of listener bias and listener learning associated with having the same listeners hear more than one child. This type of measurement is consistent with longstanding approaches to characterizing intelligibility in motor speech disorders, yielding functional measures of both single-word and multiword performance. We addressed the following specific research questions:

1. How does intelligibility grow between 30 and 47 months in typically developing children for single-word and for multiword utterances? What is the range of typical development?
2. Is there a difference between boys and girls in intelligibility development for single-words and/or multiword utterances?
3. Is growth in intelligibility different for single words versus multiword utterances?

We hypothesized that intelligibility would show steady growth through the age of 47 months. We expected that the range of variability among children would be wide but that this variability would gradually decrease with age as children's speech motor control and speech sound development matured. We had no reason to expect that boys and girls would differ from one another in their speech intelligibility, other than the finding by McLeod et al. (2015), which had a small effect size. We therefore sought to examine whether we would find similar results with a transcription intelligibility paradigm. We did not have an a priori hypothesis regarding intelligibility differences between single words and connected speech (multiword utterances). On the one hand, at earlier ages, it might be expected that single-word productions, involving fewer sound segments and therefore reduced speech motor coordination, would be more intelligible than multiword productions because they are simpler to produce. On the other hand, there is precedent in the literature for multiword utterances to have an intelligibility advantage over single words because of linguistic contextual information available in multiword utterances.

Method

Participants

A total of 164 typically developing children (92 girls, 72 boys) contributed speech samples for this study. Child participants were recruited through three methods: the placement of flyers at community locations, such as public libraries and coffee shops; postings and paid advertisements on social media; and through use of a K–12 Registry maintained by the Waisman Center Clinical Translation Core. The K–12 Registry comprises a database of families with children who attend local schools and who have opted in to being contacted about research projects. Recruitment materials for ongoing studies conducted by our research group were sent to parents of children from the K–12 Registry within the target age range. Recruitment materials indicated that younger siblings of school-age children (age 2;6 and above) were eligible for this particular study. The majority of our participants were recruited through this registry. Families of children were compensated for their participation.

All child participants, regardless of method or source of recruitment, were screened by study team members to ensure the following inclusion criteria were met: (a) American English as the primary language in the home, (b) hearing within normal limits as indicated by parent report and passing a pure-tone hearing screening or distortion product otoacoustic emission screening bilaterally,[1] (c) speech within normal limits as indicated by articulation scores on the Arizona Articulatory Proficiency Scale–Third Edition (Fudala, 2001), and (d) language within normal limits as indicated by the Preschool Language Scale–Fourth Edition screening test (Zimmerman et al., 2012). Children receiving intervention services for any educational or developmental concern were excluded as were those with any medical diagnoses related to development. This study was part of a larger data collection effort, extending through 8 years of age; children in this article comprise all participants between 30 and 47 months of age.

Chronological age of children in 1-month increments was distributed across the ages from 30 to 47 months with four to 13 children per 1-month increment. Both boys and girls were represented in each 1-month age increment, although the sample had more girls than boys. See Figure 1 for the distribution of children by age and sex. Children in this sample represented the local community, which is skewed toward middle-class and upper middle-class families who are caucasian. Demographic information on the children in this study is presented in Table 1.

A total of 328 nondisabled adults served as listeners in this study. Two different listeners were quasirandomly assigned to hear the speech of each child (164 children × 2 listeners = 328 listeners); each listener heard only one child producing all stimulus material. Listeners were recruited from the university community via public postings and were primarily undergraduate students. Listeners were compensated monetarily for their participation.

[1]Some children under 36 months of age did not tolerate hearing screening. For those children, we relied on parent report of hearing history and a negative report of parental concern to meet hearing criteria.

Figure 1. Distribution of children by sex and age.

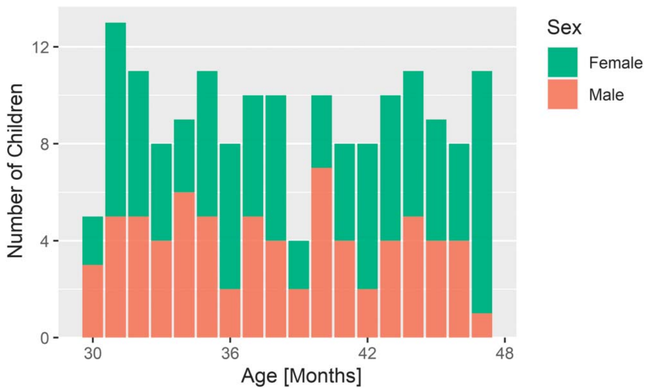


Table 1. Demographic characteristics of children (*n* = 164).

Characteristic	Male (*n* = 72)	Female (*n* = 92)
Race		
White	61 [1]	74
Black		2
Asian	1	
American Indian		1
Native Hawaiian/Pacific Islander		
More than 1 race	5	6 [1]
Other		
Not reported	4	8
2-Factor Hollingshead Social Index mean	56.76 (7.43)	56.42 (7.35)
Maternal education		
Graduate degree or graduate professional training	36	42
Standard college or university degree	32	40
Partial college or specialized training	2	4
High school graduate	1	
Not reported	1	6

Note. Number of additional children in this racial category whose parents identified them as having Hispanic ethnicity are indicated in [ ]. All other children were identified as non-Hispanic. Standard deviations are indicated by ( ).

Inclusion criteria for listeners were as follows: (a) hearing within normal limits as indicated by passing pure-tone hearing screening at 25 dB HL for 250 Hz, 500 Hz, 1 kHz, 4 kHz, and 6 kHz bilaterally; (b) age between 18 and 45 years; (c) no more than incidental experience listening to or communicating with persons having communication disorders[2]; (d) native speaker of American English; and (e) no identified language, learning, or cognitive disabilities per self-report. Listeners were 90 men and 238 women. The mean age of listeners was 20.5 years (*SD* = 3.6).

This study was reviewed and approved by the University of Wisconsin–Madison Institutional Review Board (Social and Behavioral Sciences). Informed consent was obtained by a parent or legal guardian for all child participants.

Materials and Procedures

Acquisition of Speech Samples From Children

Children participated in a standard speech protocol that was administered by a research speech-language pathologist in a sound-attenuating suite. Children produced a standard set of speech stimuli from the TOCS+ (Hodge & Daniels, 2007), a measure that has been used frequently in speech development research (Hodge & Gotzke, 2014a, 2014b; Hustad, Sakash, Broman, et al., 2019; Hustad, Sakash, Natzke, et al., 2019). Children completed a single-word production task and a multiword production task. We used an iPad to present each child with an image and a prerecorded auditory model, which was immediately repeated by the child. Child productions were monitored online by a research assistant. Productions were required to be (a) free of overlap with examiner speech, (b) free of extraneous noises, and (c) composed of all constituent words in the stimulus sentences. Children were asked to repeat productions when these three criteria were not met. Children were recorded using a digital audio recorder (Marantz PMD 570) at a 44.1-kHz sampling rate (16-bit quantization). A condenser studio microphone (Audio-Technica AT4040) was positioned next to each child using a floor stand, located approximately 18 in. from the child's mouth. The level of the signal was monitored and adjusted on a mixer (Mackie 1202 VLZ) to obtain optimized recordings and to avoid peak clipping.

Single-word stimuli were 38 individual words, including all items from the TOCS-30 word probe (Hodge & Daniels, 2007), as well as eight additional words included to ensure adequate representation of corner vowels. Multiword stimuli were 60 sentences ranging from two to seven words (10 items of each sentence length). Lexical, phonetic, syntactic, and morphological features of all stimuli were developed to be appropriate for children. Note that not all children were able to produce utterances of each sentence length due to developmental constraints. The multiword protocol started with the 10 two-word utterances and advanced to the 10 three-word utterances and so on, stopping when the child was not able to produce all 10 utterances of the target length. See Table 2 for a distribution of utterance lengths produced by 1-month age intervals.

Acquisition of Intelligibility Data

Digital recordings of children's speech were separated into single audio files and edited to remove any extraneous noises. We peak-amplitude normalized the files to ensure that maximum loudness levels were the same across children and productions, while preserving the amplitude contours of the original productions.

Audio files were played back to listeners using in-house software that presented audio samples in a self-paced experimental task and collected typed orthographic

[2]We did not obtain information regarding listener exposure to the speech of children. The majority of listeners were college students, spanning a range of majors.

Table 2. Number of children who produced each sentence length by age.

Age (months)	2-wd[a]	3-wd	4-wd	5-wd	6-wd	7-wd
30	5	5	1	0	0	0
31	13	12	3	0	0	0
32	11	11	1	0	0	0
33	8	8	2	0	0	0
34	9	9	3	1	0	0
35	11	11	4	1	0	0
36	8	8	6	2	2	0
37	10	9	2	0	0	0
38	10	10	7	2	0	0
39	4	4	4	1	1	1
40	10	10	10	3	1	1
41	8	8	7	1	1	0
42	8	8	7	5	2	2
43	10	10	8	5	1	1
44	11	11	10	8	7	5
45	9	9	9	6	6	6
46	8	8	8	4	2	2
47	11	11	10	8	7	6

[a]All children completed the two-word utterances. wd = word.

transcriptions. Listeners completed two orthographic transcription tasks, one involving single words and the other involving multiword utterances. The order of presentation of the two tasks was counterbalanced across the two listeners for each child; individual utterances within each task were randomized for each listener.

During the listening task, listeners were seated individually in a sound-attenuating suite in front of a 19-in. flat-panel screen with an external speaker placed directly beneath the screen. The peak audio output level was calibrated to approximately 75 dB SPL from where listeners were seated and was checked periodically to ensure that all listeners heard stimuli at the same output level. Listeners were told that the purpose of the study was to determine how understandable children were to unfamiliar listeners like themselves. They were instructed that children would be producing real words and to take their best guess if they were unsure as to what the child said. Listeners were provided with instructions on how to use the experimental software to advance through the experiment and how to type in their orthographic response for each child utterance. In addition, they heard four familiarization sample productions to acclimate them to the experimental task. Data from the sample productions were excluded from analyses. Listeners heard each production one time. Completion of the listening task took approximately 30 min per listener.

Listeners' typed orthographic transcriptions of children's productions were scored based on whether they were an exact phonemic match to target productions. This is a standard approach in speech motor disorders research (Hodge & Gotzke, 2014a, 2014b; Yorkston & Beukelman, 1978, 1980), and it allowed us to consider intelligibility from a lexical perspective. Intelligibility scores were obtained by counting the number of words transcribed correctly by each listener relative to the target words that children were attempting to produce. Listener transcriptions that were an exact phonemic match to the target word were counted as correct. Misspellings and homophones were accepted as correct, as long as all phonemes in the listener transcription matched the target words.

Speech Intelligibility Data: Preprocessing, Reliability, and Standardization

The total number of words transcribed correctly by each of the two listeners per child was divided by the number of words produced for the single-word intelligibility task and the total number of words produced across all utterance lengths for the multiword intelligibility task. This yielded four separate mean intelligibility scores per child (2 listeners × 2 tasks). Scores were multiplied by 100 to yield a percentage intelligibility score. All children in the sample produced all the single-word stimuli and all the two-word utterances. Table 2 reports the number of children completing each utterance length by age.

For each child, we computed the difference in average intelligibility between the two listeners for single-word and for multiword intelligibility scores following Lee et al. (2014). If the two listeners' averages differed by more than 10 percentage points (for either intelligibility measure), data from a third listener was obtained, and data from the listener who differed from the other two by more than 10 percentage points were discarded. Of the 164 child visits, this occurred in 27 instances. These instances appeared to be random and were not associated with specific children or age groups. For the final data set used in this study, the average difference between the two listeners was 5.8 percentage points (4.5 *SD*) for single-word utterances and 5.6 percentage points (4.2 *SD*) for multiword utterances.

We also calculated interrater reliability of the average single-word intelligibility scores and multiword intelligibility scores for the two listeners of each child and each visit using the intraclass correlation coefficient (ICC) estimated with the irr R package (Version 0.84.1; Gamer et al., 2019). We used an average score, consistency-based, one-way random effects model, and we observed strong agreement between raters on intelligibility scores, $ICC_{(\text{single word})} = .94$, 95% confidence interval [.92, .96]; $ICC_{(\text{multiword})} = .97$, 95% confidence interval [.96, .98]. Finally, intelligibility scores for single-word intelligibility were averaged across the two listeners per child, and the same was done for multiword intelligibility scores for each child.

A methodological challenge with multiword intelligibility measures in early childhood is that all children of a given age did not necessarily produce all utterances of each length. As such, for one child, we may be computing intelligibility averaging across, say, two- to four-word utterances, whereas, for another child, that average may span two to six word utterances. At face value, these two computed intelligibility measures are not directly comparable. To address this issue, we used the following approach (with technical details in Supplemental Material S1) to compute an "integrated weighted intelligibility response." First, for each missing utterance length, we imputed the intelligibility value using least squares regression as a function of (a) all the lower length intelligibility scores (including single words) and (b) the observed length of longest utterance (LOLU) value for that child. We imputed the lower word length intelligibilities first and worked our way up to the higher ones, using imputed lower values as predictors. Inclusion of the observed LOLU values does mean that the imputation involves a mild extrapolation and it will tend to correct missing scores up or down relative to what would be estimated based solely on lower word length values. This is appropriate here, owing to the fact that scores could be correlated with LOLU. Second, for any given age, we estimated the proportions of children whose LOLU was two words, three words, and so forth, up to seven words. We then based our integrated weighted intelligibility response on a weighted average of the two- to seven-word intelligibility scores (including the imputed ones). In effect, for the older children, all two- to seven-word scores were more equally weighted, but for younger children, the shorter utterances were given greater weight. This downplays the role of the longer utterances for the ages at which those lengths are relatively rare, allowing them to influence the final intelligibility score, but not to the same degree as the shorter utterances. In summary, our integrated weight intelligibility response approach therefore accomplished two goals: First, imputation filled in missing intelligibility scores so that all children had intelligibility scores for all utterance lengths, and as a result, the overall multiword intelligibilities were comparable between children. Second, by weighting the intelligibility scores based on which utterance lengths were typical at each age, we limited the contribution of imputed scores in longer utterance lengths to the overall average.

Statistical Analysis of Intelligibility Growth Trajectories

Our main goal was to develop a model for the mean and for the quantiles of the distribution for the single-word intelligibility scores and for the multiword integrated intelligibility response as a function of age in this typically developing population. We modeled single- and multiword intelligibility separately (except for Question 3). We used flexible linear regression models for mean intelligibility score as a function of age by modeling the age effect with a 3-degree-of-freedom (*df*) natural cubic spline for age. Natural splines are superior to polynomial functions because they provide more statistical stability in the tails of the predictor (age) distribution (Harrell, 2013). Simultaneous with the mean regression model, we also modeled the error variance using a log link and a 2-*df* natural spline function in age. The regression model is fitted via a weighted least squares algorithm with weights inversely proportional to the fitted error variance. Whereas this approach will not yield vastly different results for mean intelligibility than a model assuming constant error variance, it will, crucially, allow for flexible modeling of the more extreme quantiles of intelligibility as a function of age. Because of our strong interest in the quantiles, we are also relying more heavily on the normal assumption for the error distribution than would be the case if we were only interested in the mean regression. As such, we assessed the adequacy of the normality assumption on the standardized residuals using "worm plots" (van Buuren & Fredriks, 2001) across a variety of models and did not find any deviations from normality not explainable by ordinary sampling variance. This modeling was implemented in the *gamlss* package (Version 5.1.5; Rigby & Stasinopoulos, 2005) in the R programming environment (Version 3.6.1; R Core Team, 2019).

Questions 1 and 2: To examine intelligibility growth for single- and multiword utterances, we used the foregoing modeling approach, which directly yields estimated normal-model quantile age trajectories for each of single- and multiword utterances. These are presented graphically in Figures 2 and 3. To test for differences in the growth trajectories for boys versus girls, we maintained the 3-*df* natural spline model for the mean to capture the main effect of age, as well as the 2-*df* natural spline for the error variance. We augmented this model with a main effect for sex (whether group averages differ) and a Sex × Linear Age effect (whether the linear rate of change between the groups differ). We jointly tested these two parameters to determine whether the two sexes differed in trajectory (i.e., in level or slope or both).

Question 3: To compare growth in mean single-word versus mean multiword intelligibility, we modeled the within-child difference in intelligibility scores as a function of age. Specifically, for each child, we computed $\gamma_{\text{diff}} = \gamma_{\text{multi}} - \gamma_{\text{single}}$. We fit a baseline model with a fixed mean of 0 and a 2-*df* natural cubic spline for the error variance. This baseline model allowed the variance in the differences to change with age but assumed that the difference in intelligibility scores was 0 at all ages. We augmented this model with an intercept term and a linear age term, centering age at 38.5 months so that the intercept term estimated the average difference at the middle of the sample's age

Figure 2. Growth curve and quantiles for single-word intelligibility.

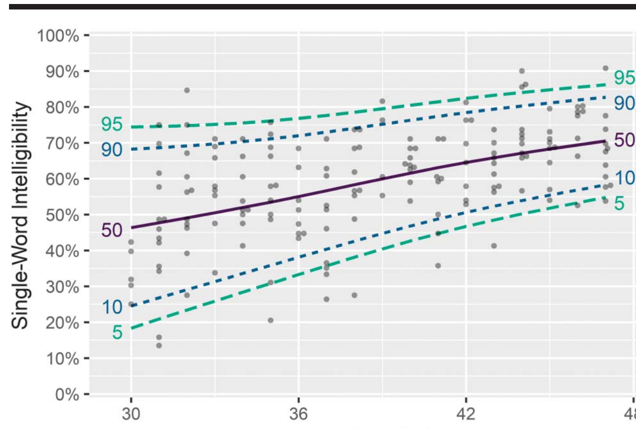
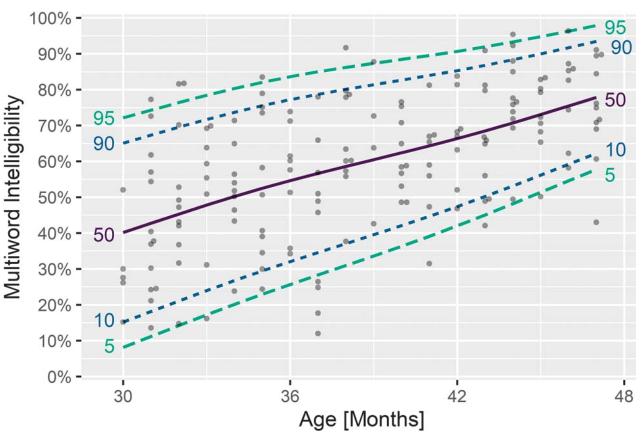


Figure 3. Growth curve and quantiles for multiword intelligibility.



range. These two effects jointly tested whether the within-children difference in intelligibility scores was nonzero.

Results

Question 1: How does intelligibility grow between 30 and 47 months in typically developing children for single-word and multiword utterances? What is the range of typical development?

Children showed steady growth in both their single-word intelligibility and their multiword intelligibility. Figures 2 and 3 show the observed intelligibility scores and the estimated growth curve quantiles for each measure. Tables 3 and 4 provide these percentiles as normative references.

For single words (see Figure 2), the estimated mean intelligibility increased with age: 46% at 30 months, 55% at 36 months, 65% at 42 months, and 70% at 47 months. In this 18-month window, children's single-word intelligibility grew from just under 50% intelligibility to 70%. Variability in intelligibility scores decreased with age; the estimated standard deviation for intelligibility scores was 17.0 at 30 months, 13.2 at 36 months, 10.9 at 42 months, and 9.5 at 47 months. Thus, the range for typical performance—the difference between 95th and 5th percentiles—decreased from 56 percentage points at 30 months to 31 percentage points at 47 months. Another result of this decreasing variability was that the floor for normal performance (the 5th percentile) increased from 18% at 30 months to 55% at 47 months. That is, the average intelligibility (50th percentile) at 36 months was the 5th percentile score at 47 months.

Table 3. Normative percentiles for single-word intelligibility.

Age (months)	Single-word intelligibility percentiles 5	10	25	50	75	90	95
30	18	25	35	46	58	68	74
31	21	27	37	48	59	69	75
32	23	29	39	49	60	69	75
33	26	31	40	51	61	70	75
34	28	34	42	52	62	70	76
35	31	36	44	53	63	71	76
36	33	38	46	55	64	72	77
37	36	40	48	57	65	73	78
38	38	43	50	58	67	74	79
39	40	45	52	60	68	75	79
40	43	47	54	62	69	76	80
41	45	49	56	63	71	77	81
42	47	51	57	65	72	78	82
43	48	52	59	66	73	79	83
44	50	54	60	67	74	80	84
45	52	55	62	68	75	81	85
46	53	57	63	69	76	82	85
47	55	58	64	70	77	83	86

Table 4. Normative percentiles for multiword intelligibility.

Age (months)	Multiword intelligibility percentiles 5	10	25	50	75	90	95
30	8	15	27	40	53	65	72
31	11	18	30	43	56	67	74
32	14	21	33	45	58	70	76
33	17	24	35	48	60	72	78
34	20	27	38	50	63	74	80
35	23	29	40	52	65	76	82
36	26	32	43	55	67	77	84
37	28	35	45	57	68	79	85
38	31	37	47	59	70	80	86
39	34	40	49	60	71	81	87
40	36	42	52	62	73	83	88
41	39	45	54	64	75	84	90
42	42	47	56	66	76	85	91
43	45	50	59	68	78	87	92
44	48	53	61	71	80	88	93
45	51	56	64	73	82	90	95
46	55	59	67	75	84	92	96
47	58	62	70	78	86	93	98

Intelligibility in multiword utterances (see Figure 3) showed a similar pattern of growth as the single words. The estimated mean intelligibility increased with age: 40% at 30 months, 55% at 36 months, 66% at 42 months, and 78% at 47 months. The average for multiword intelligibility started lower than single words at 30 months (40% vs. 46%) and ended a few points higher at 47 months (78% vs. 70%). Variability in intelligibility scores also decreased with age; the estimated standard deviation for intelligibility scores was 19.5 at 30 months, 17.6 at 36 months, 14.8 at 42 months, and 12.2 at 47 months. These standard deviations were approximately 2–4 percentage points larger than those for single-word intelligibility. Thus, the range of typical performance was wider for multiword intelligibility: The difference between the 95th and 5th percentiles was 64 percentage points at 30 months and reduced to 40 percentage points at 47 months. The floor for normal performance (the 5th percentile) increased from 8% at 30 months to 58% at 47 months. In this case, the average score (50th percentile) at 38 months was the 5th percentile at 47 months.

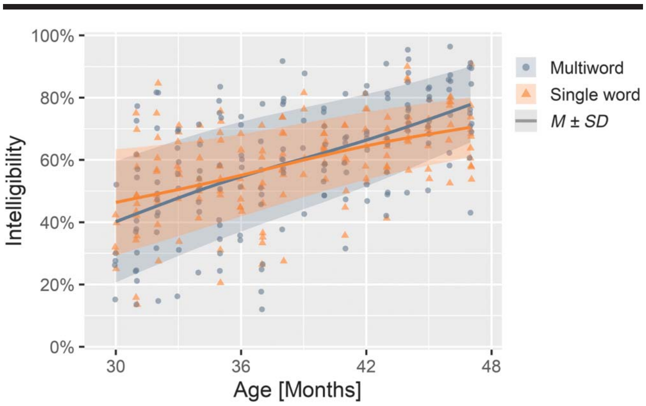
Question 2: Is there a difference between boys and girls in intelligibility development for single words and for multiword utterances?

We tested for differences in the growth trajectories for boys versus girls by augmenting our growth curve models with a main effect of sex and a sex × linear age effect. These effects did not improve model fit for single-word intelligibility, $\chi^2(2) = 0.24$, $p = .88$, or for multiword intelligibility, $\chi^2(2) = 0.46$, $p = .80$. The average intelligibility at each age did not differ between the two groups. The estimated group difference (girl growth curve − boy growth curve) at each age for single-word intelligibility ranged from −0.7 to 1.6 percentage points, and for multiword intelligibility, it ranged from −0.9 to 3.6 percentage points. Furthermore, we tested the group difference at ages when the estimated difference was greatest in magnitude (47 months for single-word intelligibility, 30 months for multiword intelligibility) and did not observe any significant group differences (single: $t = 0.67$, $p = .51$; multi: $t = 0.49$, $p = .63$).

Question 3: Is growth in intelligibility different for single words versus multiword utterances?

Figure 4 visualizes the estimated mean and standard deviation from the single-word and multiword growth curve models. Informal visual comparison suggests that multiword intelligibility grew more quickly than single-word intelligibility. As described above, we formally tested for differences in the growth trajectories by modeling whether the within-child difference in intelligibility scores—that is, multiword minus single-word intelligibility—changed with age. Our baseline model included no predictors (assuming a difference of 0 at all ages), and our augmented model included an intercept effect and linear age effect. These effects significantly improved model fit over the baseline model, $\chi^2(2) = 19.15$, $p < .001$. The intercept term—the estimated average difference at 38.5 months—did not differ significantly from 0, $p = .51$. However, the linear age effect was significant such that a 1-month change in age predicted an increase in the average intelligibility difference of 0.67 percentage points, $SE = 0.15$, $t = 4.55$, $p < .001$. Following the age range of the sample, children started with a single-word advantage at 30 months of 5.1 percentage points, $SE = 1.6$, and ended with a multiword advantage at 47 months of 6.3 percentage points, $SE = 1.4$.

Figure 4. Estimated growth trajectories for the mean and standard deviation of each intelligibility type. The standard deviation band is wider for multiword intelligibility than for single words (reported in Question 1). There is a single-word advantage at 30 months but a multiword advantage at 47 months (reported in Question 1 and estimated in Question 3).



Discussion

In this study, our main goal was to create growth curves for intelligibility development in children that could be used to determine whether children at risk for speech motor impairment were developing within the range of typical age expectations. To do this, we collected single-word and multiword elicited speech samples from typically developing children between the ages of 30 and 47 months and had naïve listeners orthographically transcribe these speech samples. Key findings from this study were as follows. First, children showed steady growth in intelligibility of both single words and multiword utterances through the age of 47 months. Only the most advanced children (95th percentile) approached 100% intelligibility by 47 months; even children performing at the 90th percentile were well under this threshold. There was a wide range of variability among typically developing children in intelligibility development, especially at younger ages, but this variability reduced with age. We also found that boys and girls did not differ in their intelligibility. Finally, there were differences in growth of single-word intelligibility and multiword intelligibility, with single-word utterances being more intelligible at the earliest ages and multiword utterances being more intelligible at later ages. These findings are discussed in detail below.

Growth and Variability in Intelligibility

A key contribution of this article is the creation of growth curves for intelligibility development in typically developing children based on advanced statistical modeling of objectively derived speech intelligibility data from unfamiliar listeners. Other studies have examined average intelligibility by age in 6-month or 12-month intervals, but neither percentile-level data on speech intelligibility development nor age intervals of 1 month have been examined previously. A main finding from this study is that children showed steady progress in intelligibility development for single words and multiword utterances all the way through 47 months; however, there was tremendous variability among children at each 1-month age increment. The average child, performing at the 50th percentile, had 70% intelligibility for single words and 78% intelligibility for multiword utterances at 47 months. Children showed considerable growth from 30 to 47 months, with the average child improving by 24 and 38 percentage points for single words and multiword utterances, respectively. Children in the higher percentiles seemed to make slower gains in intelligibility development than those in the lower percentiles, in part, because those in the higher percentiles started with higher intelligibility and did not have as far to go prior to approaching the ceiling of development. It should be noted that none of the children reached 100% intelligibility by 47 months of age for single words or for connected speech. Children in the 95th percentile came close, with 98%

intelligibility at 47 months of age. These findings indicate that, for the vast majority of children, intelligibility is still developing beyond 4 years of age, and for many children, a considerable amount of growth remains prior to approaching mature adult-level intelligibility.

There was a very wide range of variability among typically developing children for both single-word intelligibility and multiword intelligibility at any given age based on the difference between scores at the 5th versus 95th percentile. The most extreme example of this can be seen for multiword intelligibility at 30 months of age where the range spanned from 8% to 72%. This variability was slightly higher for multiword intelligibility than for single-word intelligibility. In Tables 3 and 4, we present percentiles by 1-month age increments for intelligibility growth. Several observations are noteworthy. In particular, the range of scores between the 5th and 95th percentiles within age increments was reduced by about 20 percentage points at 47 months of age (relative to variability at 30 months of age). However, even at 47 months, there was still a 30–40 percentage point range in intelligibility scores across children in the different percentile bands. The finding of reduced variability with age is consistent with other domains of development, particularly within speech and language. In the time frame between 30 and 47 months, children are making critical advances in their articulatory, speech motor, and phonological development. In particular, children are rapidly acquiring speech sounds, with the vast majority of consonants emerging or reaching mastery in this window of time (Sander, 1972).

Interestingly, the magnitude of variability within age across 5th–95th percentiles was greater than the magnitude of change across age from 30 to 47 months within each percentile. Thus, within-age differences among children were greater than between-ages development over time. For example, at 36 months, for single words, intelligibility ranged from 33% to 77% across percentiles (difference of 44). However, for the 5th percentile, intelligibility change from 30 to 47 months was 37 percentage points; for the 95th percentile, intelligibility change in the same time frame was 12 percentage points. We can think of this finding in developmental terms: Intelligible speech follows a rapid developmental trajectory (on average), and if intelligible speech emerges early in some children and late in some children, then we can expect large within-age differences between children early on. These findings clearly highlight that there is a wide range of typical performance in intelligibility development, which is a key reason why early differential diagnosis of speech deficits can be challenging in children.

Findings from this study show both similarities and differences to those of previous studies. Methods for this study were most consistent with those of Hodge and Gotzke (2014a), using elicited words and sentences from the TOCS+ (Hodge & Daniels, 2007) with naïve listeners who orthographically transcribed children's productions. Importantly, Hodge and colleagues pooled data from children into 1-year age bands and presented means along with standard deviations. Results showed that mean intelligibility for single words at 3 years was 70% and at 4 years was 80%. In this study, scores were generally lower, with median scores (50th percentile) of 55%–63% in the Year 3 age range and 65%–70% in the Year 4 age range. One explanation for the higher scores in the Hodge and Gotzke study may relate to the smaller sample size (n = 12 per 1-year age group), and the fact that the specific age in months for the children in the study was not specified; thus, the composition of older versus younger children within the two age bands is unknown, making findings difficult to compare directly to results of this study. For multiword utterances, results from Hodge and Gotzke showed mean intelligibility at 3 years to be 86% and at 4 years to be 92%. In this study, our results were, again, considerably lower, with median scores (50th percentile) of 55%–64% in the Year 3 age range and 66%–78% in the Year 4 age range. A key methodological difference between this study and the Hodge and Gotzke study for multiword utterances is that listeners were allowed to hear multiword utterances up to 2 times, whereas in this study, they heard each utterance only once. The opportunity to hear each stimulus item a second time may have bolstered intelligibility scores. Such a finding would be consistent with studies of perceptual learning of speech (Borrie et al., 2012).

Results from this study are very discrepant with the parent report–based findings of Coplan and Gleason (1988), who suggested that 90% of children should be at least 50% intelligible at 22 months, 75% intelligible at 37 months, and 100% intelligible at 47 months. In this study, our youngest children were 30 months of age. However, relative to multiword intelligibility percentile data from this study, only children in the 75th percentile were near 50% intelligible to unfamiliar listeners at 30 months. At 37 months of age, only children in the 90th percentile were approximately 75% intelligible, and at 47 months, only children in the 95th percentile approached 100% intelligible. Thus, our findings suggest that the results of Coplan and Gleason overestimate typical intelligibility development and are most reflective of high-performing children in the upper percentiles.

Similarly, findings of Rice et al. (2010) showed that intelligibility as measured by percentage intelligible utterances was an average of 86% between 30 and 35 months, 91% between 36 and 41 months, and 92% between 42 and 47 months. Average (50th percentile) children in these same age bands for this study were about 30–40 percentage points lower for multiword intelligibility. Data from Rice et al. are most consistent with children performing in the 95th percentile for each age group and thus represent an overestimate of transcription-based intelligibility scores for the average child. Note that we used a different methodology for obtaining intelligibility scores in this study, and therefore, direct comparison of findings is not entirely appropriate, though it does provide useful context for considering results. It should also be noted that this discrepancy is not surprising; we would expect intelligibility to be higher when a rich linguistic context is available to an expert transcriber.

Finally, results from Flipsen (2006) reveal that intelligibility ranged from 88% to 100% at both 3 and 4 years of age for children with typical speech, with averages at about 95% and 97%, respectively. Again, these results are far above those obtained in this study and, as with studies by Coplan and Gleason (1988) and Rice et al. (2010), are likely most reflective of methodological differences in how intelligibility was measured.

Differences in Growth for Single- Versus Multiword Intelligibility

In this study, we found that single words were reliably more intelligible than multiword utterances before 35 months of age. However, the rate of change was higher by 0.67 percentage points per month for multiword utterances than for single-word utterances, ultimately resulting in higher intelligibility for multiword utterances starting at 40 months. It is well established that intelligibility of sentences is generally higher than intelligibility of single words, owing to the linguistic context provided by sentences in adult speakers (Miller et al., 1951) and in typically developing children at 4 years of age (Hustad et al., 2012). However, previous studies have not examined age effects for intelligibility of single words versus multiword utterances in early talkers; thus, our findings present a novel contribution. In early development until the third year of life, our results suggest that there is an advantage for single words over multiword productions. This is most likely the result of immature segmental articulation and reduced motor control such that production of multiword utterances disproportionately taxes the speech production system resulting in speech that is diminished in intelligibility. Simpler, single-word productions that require less motor coordination thus have an advantage in terms of intelligibility. In children with dysarthria, we have observed this effect, particularly for those with severe dysarthria (Hustad et al., 2012). Young typically developing children may bear some resemblance to those with dysarthria in that speech production may be characterized by a variety of different articulatory errors, coordination difficulties relative to mature speakers, and different perceptual features of speech. However, as children's articulation, phonology, and morphosyntax develop, their ability to produce increasingly complex and longer utterances with greater precision increases, eventually leading to the advantage for multiword utterances observed in adult speech. Results from this study suggest a developmental transition from an intelligibility advantage for single-word productions to similar performance at 35 months and a transition from similar performance to an intelligibility advantage for multiword productions around 41 months of age.

Limitations and Future Directions

There are a number of limitations to this study. Most critically, although our sample was representative of the geographic region in which the study was conducted, it is not representative of the United States as a whole. Children were primarily caucasian and from middle-class and upper middle-class families where parents were highly educated. These children were likely advanced in their speech and language abilities, and thus, data from their speech intelligibility samples may provide an overestimate of intelligibility relative to the population of typically developing children in the United States. Similar sampling biases also occur in other studies of children's speech referenced in this article. Research is needed to investigate a wider range of children, with a focus on children from diverse backgrounds, to extend our understanding of intelligibility development.

Children and their listeners primarily spoke American English from the North Central or Inland North regions of the United States (Labov et al., 1997), with all having been recruited from Madison, Wisconsin, and surrounding areas. We do not know whether our findings would generalize to different dialect groups. While we do not have reason to believe that different dialects in the United States would have different patterns of intelligibility development, it is possible that socioeconomic differences may interact with dialect to yield different intelligibility growth patterns. Studies should examine intelligibility development in different dialects to determine how dialect may relate to intelligibility development.

Children participated in highly structured tasks for this study. We used an elicitation paradigm where children heard a model of each target utterance and then repeated the model. Hearing a model prior to producing utterances may have impacted how children produced target utterances, likely in a positive manner, such that our samples may have captured children's best productions. Studies should examine speech intelligibility in spontaneous contexts with unfamiliar listeners to determine how spontaneous speech may differ from elicited speech. It is likely that spontaneous speech, particularly in a contextually rich environment where real communication occurs, may have higher intelligibility than our results have revealed for words and for sentences.

Listeners heard children in an ideal listening environment that was quiet. Studies are needed that examine intelligibility in real environments that include noise to extend our understanding of the many complex contributors to intelligibility.

Perhaps most importantly, we examined only typically developing children in this study. We did not compare these children with same-age peers with known disorders or risk factors for disorders. Therefore, we are only able to provide descriptive information regarding age expectations for typically developing children across a range of quantiles. Development of definitive cut-points for intelligibility by age requires analyses of intelligibility data from both typical and atypical children using response operator curves that allow characterization of sensitivity and specificity of intelligibility scores for differentiating between groups of children. Such studies are urgently needed.

Clinical Implications

Intelligibility can be measured in a variety of different ways. Characterization of intelligibility using objective, reliable, and replicable methods is critically important, particularly if clinical decisions will be made based on intelligibility data. In this article, we report normative growth curves for the development of speech intelligibility in typically developing children. Data reported here are likely to overestimate transcription intelligibility scores via naïve listeners because of the population characteristics of children from which we sampled. Thus, our findings may provide a best case scenario for intelligibility development. Nevertheless, there are several key clinical implications from this study. First, even the most intelligible children should not be expected to be 100% intelligible by 4 years of age. Second, the range of typical viability is very wide, particularly for young children; this range narrows with age. Thus, "typical" performance is a fast moving target during this age range. Third, multiword intelligibility should be higher than single-word intelligibility by 42 months of age. Finally, intelligibility development continues to advance into the fourth year of life for most typically developing children.

Acknowledgments

This study 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. The authors thank the children and their families who participated in this research and the graduate and undergraduate students at the University of Wisconsin–Madison who assisted with data collection and data reduction, especially Michael Molinaro, who collected data from many children.

References

Binger, C., Ragsdale, J., & Bustos, A. (2016). Language sampling for preschoolers with severe speech impairments. *American Journal of Speech-Language Pathology, 25*(4), 493–507. https://doi.org/10.1044/2016_AJSLP-15-0100

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech, Language, and Hearing Research, 55*(1), 290–305. https://doi.org/10.1044/1092-4388(2011/10-0349)

Coplan, J., & Gleason, J. R. (1988). Unclear speech: Recognition and significance of unintelligible speech in preschool children. *Pediatrics, 82*(3), 447–452.

Flipsen, P. (2006). Measuring the intelligibility of conversational speech in children. *Clinical Linguistics & Phonetics, 20*(4), 303–312. https://doi.org/10.1080/02699200400024863

Fudala, J. B. (2001). *Arizona Articulatory Proficiency Scale–Third Edition*. Western Psychological Services.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement*. https://CRAN.R-project.org/package=irr

Harrell, F. E. (2013). *Regression modeling strategies with application to linear models, logistic regression, and survival analysis*. Springer Science & Business Media.

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

Hodge, M., & Gotzke, C. L. (2014a). 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. (2014b). 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

Hustad, K. C., Sakash, A., Broman, A. T., & Rathouz, P. J. (2019). Differentiating typical from atypical speech production in 5 year old children with cerebral palsy: A comparative analysis. *American Journal of Speech-Language Pathology, 28*(2S), 807–817. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0108

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

Hustad, K. C., Schueler, B., Schultz, L., & DuHadway, C. (2012). Intelligibility of 4-year-old children with and without cerebral palsy. *Journal of Speech, Language, and Hearing Research, 55*(4), 1177–1189. https://doi.org/10.1044/1092-4388(2011/11-0083)

Kent, R., Kent, J., Weismer, G., Martin, R., Sufit, R., Brooks, B., & Rosenbek, J. (1989). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics, 3*(4), 347–358. https://doi.org/10.3109/02699208908985295

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

Labov, W., Ash, S., & Boberg, C. (1997). *A national map of the regional dialects of American English*. https://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html#Heading3

Lee, J., Hustad, K. C., & Weismer, G. (2014). Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy. *Journal of Speech, Language, and Hearing Research, 57*(5), 1666–1678. https://doi.org/10.1044/2014_JSLHR-S-13-0292

McLeod, S., Crowe, K., & Shahaeian, A. (2015). Intelligibility in Context Scale: Normative and validation data for English-speaking preschoolers. *Language, Speech, and Hearing Services in Schools, 46*(3), 266–276. https://doi.org/10.1044/2015_LSHSS-14-0120

McLeod, S., Harrison, L. J., & McCormack, J. (2012). The Intelligibility in Context Scale: Validity and reliability of a subjective rating measure. *Journal of Speech, Language, and Hearing Research, 55*(2), 648–656. https://doi.org/10.1044/1092-4388(2011/10-0130)

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*(5), 329–335. https://doi.org/10.1037/h0062491

Natzke, P., Sakash, A., Mahr, T. J., & Hustad, K. C. (2020). *Measuring speech production development in children with cerebral palsy between 6 and 8 years of age: Relationships among measures*. Manuscript submitted for publication.

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

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Sander, E. K. (1972). When are speech sounds learned? *Journal of Speech and Hearing Disorders, 37*(1), 55–63. https://doi.org/10.1044/jshd.3701.55

Tikofsky, R. S., & Tikofsky, R. P. (1964). Intelligibility as a measure of dysarthric speech. *Journal of Speech and Hearing Research, 7*(4), 325–333. https://doi.org/10.1044/jshr.0704.325

van Buuren, S., & Fredriks, M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine, 20*(8), 1259–1277. https://doi.org/10.1002/sim.746

Weismer, G., & Martin, R. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R. Kent (Ed.), *Intelligibility in speech disorders* (pp. 67–118). John Benjamins.

WHO Multicentre Growth Reference Study Group, & de Onis, M. (2006). WHO child growth standards based on length/height, weight and age. *Acta Paediatrica, 95*(S450), 76–85. https://doi.org/10.1111/j.1651-2227.2006.tb02378.x

Yoder, P. J., Woynaroski, T., & Camarata, S. (2016). Measuring speech comprehensibility in students with Down syndrome. *Journal of Speech, Language, and Hearing Research, 59*(3), 460–467. https://doi.org/10.1044/2015_JSLHR-S-15-0149

Yorkston, K., & Beukelman, D. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders, 11*(6), 499–512. https://doi.org/10.1016/0021-9924(78)90024-2

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

Yorkston, K., Beukelman, D., Strand, E., & Bell, K. (1999). *Management of motor speech disorders in children and adults* (2nd ed.). Pro-Ed.

Yorkston, K., Beukelman, D., & Tice, R. (1996). *Sentence Intelligibility Test*. Madonna Rehabilitation Hospital.

Yorkston, K., Beukelman, D., & Traynor, C. (1984). *Assessment of intelligibility of dysarthric speech*. Pro-Ed.

Zimmerman, I., Steiner, V., & Pond, R. (2012). *Preschool Language Scale–Fifth Edition (PLS-5)*. The Psychological Corporation. https://doi.org/10.1037/t15141-000