The Effects of Clinical Experience in the Rating of Intelligibility of Phonetically Contrasted Words

by

Emily Elizabeth Hanner

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 3, 2019

Keywords: intelligibility, experienced vs. novice listeners,
perception, speech sound disorders,
direct magnitude estimation, crowdsourcing

Approved by

Marisha Speights Atkins, Chair, Assistant Professor of Communication Disorders
Lawrence Molt, Professor of Communication Disorders
Mary Sandage, Associate Professor of Communication Disorders
Aurora Weaver, Assistant Professor of Communication Disorders

Abstract

Purpose: The purpose of this study was to investigate if crowdsourced lay listeners with minimal exposure to child speech and certified clinicians in the field of speech-language pathology experienced in treating speech sound disorders rate speech intelligibility by direct magnitude estimation (DME) differently.

Method: Speech recordings of 9 preschool children producing phonetic contrasts were rated by 30 listeners, 15 inexperienced listeners and 15 experienced clinicians, to compare perceptual ratings of child speech using Direct Magnitude Estimation (DME) to determine whether a listener bias exists resulting in the inconsistent subjective rating of intelligibility. Listening judgments were recruited through two crowdsourcing methods, Amazon Mechanical Turk and the ASHA Community sites.

Results: The results of this study reinforce the correlation between measures of whole-word accuracy and ratings of intelligibility. It was found that listeners, both inexperienced and experienced with child speech productions, distinguish differences in intelligibility categorically when compared to word production accuracy. A significant difference was not found between DME intelligibility ratings of inexperienced and experienced listeners.

Conclusions: Online crowdsourcing for the perceptual rating of child speech intelligibility provides high-quality data consistent with measures of whole word accuracy. Additionally, in this study there was evidence that indicates inexperienced and experienced listeners ratings are in concordance. This novel approach to rating child speech intelligibility increases the ability to obtain laypersons ratings using an ecologically valid approach.

Acknowledgments

I want to sincerely thank my committee chair, Dr. Speights Atkins, for being an excellent leader and friend during this process. She has been a constant source of encouragement to me while also pushing me to go the extra mile. I would also like to thank my committee members: Dr. Lawrence Molt, Dr. Mary Sandage, and Dr. Aurora Weaver. You have each provided invaluable knowledge, resources, and clarity to me throughout this important work. Javier Livio, thank you for training me, answering all of my questions, and being the 'brains' behind the Intelli-turk© platform without which this project would have been impossible. To the members of the Technologies for Speech-Language Research Lab, I appreciate the laborious jobs you have done behind the scenes for the completion of this project. Lastly, I want to express my thankfulness to my husband, Kevin Hanner, for supporting my dreams, cooking supper for us nearly every night while I am away from home, and loving me fully.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DME	Direct Magnitude Estimation
SG	Speaker Group
PWC	Percentage of Whole-Words Correct
PCC	Percentage of Consonants Correct
SLP	Speech-Language Pathologist
WRS	Word Recognition Score
PCE	Phonetic Contrast Errors
TE	Target Errors
Non-TE	Non-Target Errors
WRE	Word Recognition Error

I. Introduction

Speech-language pathologists (SLPs) are the professionals responsible for diagnosing and treating both articulation and phonological disorders in children. Because children with speech sound disorders are often described as unintelligible to inexperienced listeners, improving speech intelligibility within this population is necessary for successful oral communication between them and the people interacting with them on a daily basis (Edition & Bauman-Waengler, 2012). The effectiveness of treatment is dependent on a definitive and accurate evaluation of intelligibility (Miller, 2013; Hustad, Oakes, & Allison, 2015).

Perfect articulation is not required for intelligibility but must be able to be mapped by the listener. Research has shown that a number of variables may influence intelligibility including, but not limited to, age (Hodge & Gotzke, 2014a), utterance length and complexity (Allison & Hustad, 2014), listener familiarity with the speaker and experience with listening to child or dysarthric speech (D'Innocenzo, Tjaden, & Greenman, 2006; Liss, Spitzer, Caviness, & Adler, 2002), and quantitative characteristics of the speech signal (Allison & Hustad, 2018). While there are some standardized protocols for the rating of intelligibility available, auditory-perceptual judgments, which are vulnerable to biases, are generally used as the final clinical decision-making tool (Kent, 1996). Relying on intelligibility estimation can present unreliable and inconsistent measurements across the child population because of the limitations of perception and listener variability for clinical diagnoses. It has been postulated that listeners who have gained experience in listening to and understanding speech sound disorders or other types of disordered speech may perceive speech samples from a particular population to be more intelligible than naïve listeners due to their capacity to habituate to the disordered speech (Flipsen, 1995; Kent, 1996).

Kent, Weismer, Kent & Rosenbeck (1989) employed a method of speech intelligibility assessment that distinguished specific phonetic attributes that play a role in determining intelligibility in dysarthric adults. Various phonological categories have been attributed to decreased intelligibility in child speakers as well. Decreased word accuracy has also been shown to correlate with intelligibility ratings. Nevertheless, it is still undetermined whether listeners have different types of biases based on listening experience when both contrast categories and listening experience are controlled (Willoughby, 2019; Speights Atkins, Willoughby, Weaver, Sandage, Bailey & Livio, 2019). This study utilizes acoustic-phonetic contrast word sets (Kent et. al,1989) found to be sensitive to decreased speech intelligibility in adults with dysarthria. Selected categories that represent common phonological errors found to predict decreased speech intelligibility in children with and without speech disorders (Willoughby, 2019; Speights et al, 2019) were used to compare the ratings of intelligibility between two listener groups, inexperienced listeners and experienced clinicians.

II. Review of Literature

Speech Intelligibility

In the field of speech-language pathology, a consensus on the measurement, assessment, and determination of speech intelligibility, especially in young children, has yet to be reached. Due to the nature and complexity of the speech signal and the numerous factors influencing speech intelligibility, little agreement has been made regarding the process of evaluating it. Of primary importance, speech intelligibility is a collaborative product of both the speaker and the listener. Intelligibility not only depends on characteristics of the spoken message such as linguistic structure, familiarity, and length of utterance, but also on contextual information including auditory signal quality and contextual and visual cues provided by the speaker. The listener's competence also plays a significant role in determining intelligibility. For example, the listener's familiarity with the speaker, ability to discern contextual cues, and comprehension of visual and acoustic speech signals aid in determining whether a message is fully understood (Kent, 1992). A child's speech intelligibility level is often based on the amount of speech understood by people in his or her environment (Flipsen, 1995). Therefore, a critical distinction between the two components of intelligibility measurement, signal-dependent intelligibility and signal-independent intelligibility, should be made when considering intelligibility within clinical and social settings (Miller, 2013).

On average, eight to nine percent of young children are diagnosed with articulation or phonological disorders (NIDCD, 2016). Therefore, the need to devise an intelligibility assessment measure that is reliable and specific is essential for the efficacy of treatment (Miller, 2013; Hustad, Oakes, & Allison, 2015). While signal dependent listening tasks such as word recognition, direct magnitude estimation, and interval scales are common approaches for objective measurement of intelligibility, time constraints and listener availability in clinical settings prevent these methods from being used consistently (Gordon-Brannan & Hodson, 2000; Miller, 2013). In fact, seventy-five percent of speech-language pathologists (SLPs) rate intelligibility without the application of any standardized protocol (Skahan, Watson, & Lof, 2007). According to Kent (1996), auditory-perceptual judgments are most often the final determiner in the clinical decision-making process and provide the standards against which objective measures are evaluated. For a number of communication disorders, auditory-perceptual judgment is the chief means for assessing the outcome of an intervention program. Discerning subtle differences between speech productions can be labor intensive for clinicians, particularly when evaluating extensive speech samples. While convenient, the use of auditory-perceptual judgments alone in clinical practice is susceptible to various errors and biases (Kent, 1996). Researchers have found that among even the most experienced clinicians perceptual severity ratings have varied greatly (Dale et al, 2019).

Goldstone and Henrickson (2010) defined categorical perception (CP) as "the phenomenon by which the categories possessed by an observer influences the observers' perception (p. 1)." According to the concept of categorical perception, individuals tend to perceive the world around them according to the categories formed by the individual. Differences between objects belonging in contrasting categories are emphasized and differences between objects in the same categories are masked. The effect of categorical perception has been best shown through studies involving speech phoneme categories. Although there is evidence that some of the effects of categorical perception are innate or a property of the sound signal itself, recent evidence suggests that categorical perception is subject to learning. For example, talented musicians exhibit a notable categorical perceptional effect for relative pitch differences, which proposes that training is instrumental in sensitizing boundaries between semitones (Burns & Ward, 1978; Zatorre & Halpern, 1979).

Variability in auditory-perceptual judgments is inevitable because listeners are prone to occasionally hear what is *not there* while at times neglecting to hear what *is there*. Human perceptions, in general, are susceptible to a number of errors; thus, limitations in the perception of typical speech may be even more magnified when attending to atypical or disordered speech. For example, factors such as the availability of visual information, the quality of the acoustic signal, the physical features and history of the client, and listener characteristics can play a role in perceptual judgments of speech. The listener's linguistic experience and familiarity with the speaker is especially important when considering auditory-perceptual judgment because some listeners may be familiar with certain types of disordered speech, whereas others' linguistic backgrounds significantly affect their perceptual judgments. Titles such as experienced clinicians, certified speech-language pathologists, and those experienced in listening to speech disorders are used to describe individuals considered competent in diagnostic perceptual tasks. Although experience is valuable, it does not guarantee interjudge agreement unless the experience holds fundamental commonalities (Kent, 1996).

Experienced vs. Inexperienced Listeners

Speech characteristics can be studied from two perspectives: that of the speaker or of the listener. The perspective of the speaker focuses on the articulation of speech sounds and acoustics of speech signals, while the perspective of the listener is based on perceptual judgment. Controlling for the listening group is of high importance in perception studies because the experience of the listeners has been shown to play an influential role in perceptual judgments. For example, listeners' familiarity with a specific type of speech has been found to influence the way speech samples are judged (Boonen, Kloots, Verhoeven, & Gillis, 2019). Individuals who are skillful in understanding a specific population's disordered speech may conclude that a speech sample of that population is more intelligible compared to inexperienced listeners (Flipsen, 1995). Experienced listeners are more exposed to particular types of speech; so, it is assumed that sensitivity and proficiency in noticing subtle differences are heightened (Beukelman & Yorkston, 1980; Munson, Johnson, and Edwards, 2012). In a study carried out by McGarr (1983), experienced listeners systematically supplied higher intelligibility ratings than listeners who were unfamiliar with deaf speech. Audiologists and primary schoolteachers more accurately recognized normal hearing (NH) children's speech, while inexperienced listeners more accurately recognized hearing impaired (HI) children. Of the three listener groups, inexperienced listeners were more likely to incorrectly label the utterance as that of a child with cochlear implants (CI) or hearing aids (HA). This may be explained by the idea that variability is very common in child speech, and both audiologists and primary schoolteachers are familiar with this variability as well as aware of the normal deviations in children's developing speech. This could possibly guide them to be more tolerant toward differences in speech compared to listeners who are less experienced and compel them to demonstrate more hesitance in using these labels (Boonen, Kloots, Verhoeven, & Gillis, 2019).

Experience in listening to less intelligible speech has repeatedly been shown to improve listeners' recognition and comprehension of speech, when speech is produced by an individual with dysarthria (Tjaden & Liss, 1995), a hearing impairment (McGarr, 1983), or a foreign accent (Verhoeven, 2013). However, there is still much to be learned about the cognitive mechanisms that underlie these improvements (Francis, Nusbaum, & Fenn, 2007). Clinical judgments are likely to be to be influenced by individuals' level of clinical experience.

Wolf et al. (2003) found that students lacking clinical experience had significant difficulty perceiving essential acoustic cues for /r/ and /w/ compared to students with some clinical experience.

A limited number of studies have investigated perceptual differences between clinically-trained listeners and inexperienced ones. Significant implications can result from differing perceptions of children's speech such as contradictory feedback on the accuracy of speech productions as well as varying diagnoses. Studies involving phonetic transcriptions of children's speech have provided the current information known about differentiating between typical and atypical speech development; however, the accuracy of the gathered data lies on the ability of the listeners to perceive and identify children's speech reliably. Munson et al. (2012) investigated how clinical training affects SLPs assessment and ongoing observation of children's speech. Experienced speech-language pathologists (SLPs) and inexperienced perceptions of child speech were investigated. As predicted, experienced listeners exhibited higher intra-rater reliability than the inexperienced listeners, showing that clinical experience causes listeners to obtain a more systematic approach in making judgments of speech than inexperienced listeners. Inexperienced listeners were more likely to label a child's productions as a sound that occurs more frequently in real words than the experienced listeners. This was considered to likely be due to the fact that experienced listeners work with clients on the less commonly occurring sounds. Another possible explanation is that experienced listeners have overt awareness of children's substitution patterns; therefore, their responses reflected perceptual compensation. Evidence that experienced listeners and inexperienced listeners weigh acoustic measures differently during the rating of children's speech was found which implies more reliability and validity in judgment of speech sounds by experienced listeners. The limitations of this particular study include the asymmetry of age and gender between the two listener groups as well as the use of word fragments rather than actual words. Clinicians are more experienced in hearing fragmented words than novice listeners, so this fact possibly may have attributed to the final results (Munson, Johnson, & Edwards, 2012). Similar results were shown in a study in which a listening panel comprised of individuals with various degrees of experience with foreign-accented speech was used to assess the degree of accentedness in speakers with Foreign Accent Syndrome. Expert teachers of Dutch as a foreign language were the most lenient toward foreign-accented speech shown by willingness to consider speakers as native speakers of Dutch more often than inexperienced listeners (Verhoeven et al., 2013).

It has also been found that auditory language processing is modified by the previous experience of a listener. A listener's personal experience with the activities or message being linguistically communicated appears to control the neural processes at work during comprehension. One study in which ice-hockey experts and novices listened to sentences detailing hockey and sentences containing information about everyday situations showed significantly higher brain activation in the left inferior frontal gyrus (IFG) and bilateral caudate nuclei during language processing of hockey-related content (Lyons et al., 2010). While hockey experts were experienced with both hockey and everyday situations, personal relevance of the linguistic material may impact meaning processing to an even greater extent than personal experience. In other words, one's experiences with linguistic content, in addition to the degree to which one considers this content personally relevant, affects semantic-level language processing. Listeners' backgrounds and experiences affect perceptual strategy used when making perceptual judgments. In one particular study that compared the perception of voice quality between naïve listeners and experienced clinicians, data proposed that clinical training and experience result in listeners differing more in how they perceive voice quality (Kreiman, Gerratt, & Precoda, 1990). Regardless of the characteristic being measured, the perceptual differences between listener groups have been shown throughout multiple studies.

Although variance within studies controlling for familiarity has been shown, familiarity has been a topic of extensive study in identifying variations in listener performance of the rating of intelligibility. (King & Gallegos-Santillan, 1999; Tjaden & Liss, 1995a; Yorkston & Beukelman, 1983). Dagenais, Watts, Tarnage, and Kennedy (1999) found practicing SLPs rated the intelligibility of two dysarthric speakers higher than untrained listeners. This proposes the idea that contextual familiarity of the listener with that which is spoken by the speaker may also play a role in determining intelligibility within a clinical setting. For instance, heightened levels of linguistic context (e.g. connected speech) have been found to lead to increased intelligibility compared to single-word productions for adults (Hustad, 2007; Yorkston & Beukelman, 1981). Evidence has shown that transparent effects are present for listener familiarity with a distinct speaker, with disordered speech, and with the test material (Liss et al., 2002; D'Innocenzo et al., 2006; Utianski et al., 2011; Borrie et al., 2011). Because experience and contextual familiarity matters in the measuring of intelligibility, these potential listener effects should be considered throughout the assessment of speech sound disorders.

Current Assessment Practices of Speech Intelligibility

There are two types of features to consider regarding intelligibility: signal-dependent and signal-independent. When measuring intelligibility, a number of methods are currently used to pinpoint signal-dependent features of intelligibility, attributes based solely on the sound signal itself, whereas signal-independent characteristics employ the immediate acoustic signal as well as cues and clues from any additional verbal (e.g. syntax, semantics) or non-verbal sources (e.g.

facial expression, gestures, contextual setting, Miller, 2013). The signal independent features are not mutually exclusive but rather complementary due to the speaker and listener dyad utilizing visual and listening strategies for the purpose of maintaining intelligibility (Mattys et al., 2012; Smiljanic and Chandrasekaran, 2013). A number of methods are currently used to measure intelligibility including: phonetic contrast analysis (e.g. CID Word SPINE, Mosen, 1981), phonological process analysis (e.g. HAPP-3, Hodson, 2004), word identification without phonetic or phonological analysis (e.g. AIDS, Yorkston, Beukelman & Traynor, 1984), analysis of data from continuous speech (e.g. PCC, Shriberg, Austin, Lewis, McSweeny & Wilson, 1997), the Likert scale (McLeod, Harrison, & McCormack, 2012), visual analog scaling (Abur, Enos, & Stepp, 2019), and direct magnitude estimation (Weismer, & Laures, 2002).

Each method's effectiveness is dependent on the individual and the overall purpose of intelligibility testing as well as characteristics of the listener (Kent, Miolo, & Bloedel, 1994). Rating scales are easy to complete and efficient, especially in a clinic setting, however assigning ratings using a visual analogue scale can be biased due to listener rating disagreement of mild-moderate-severe, as well as offer no specificity for therapy targets to improve overall intelligibility. One listener may focus attention on one output feature, whereas another listener may base judgment on another feature's distortion to determine rating. Furthermore, listeners' perceptions of the severity rating of intelligibility will differ (Miller, 2013).

Direct Magnitude Estimation

As direct magnitude estimation (DME) has been determined to be a more accurate and functional scale for the measurement of intelligibility, the DME method requires listeners to assign a number along a continuous medium that corresponds with their perception of previously heard samples rather than along linear intervals (Schiavetti et al., 1981). Stevens (1951) found that ratio measurements allow for greater statistical functions compared to interval measurements. Additionally, perceptual judgements of intelligibility, much like vocal characteristics such as pitch and loudness, are difficult to make using linear intervals (Stevens, 1986; Stevens & Glanter, 1957). In research settings, DME has been frequently used in the scaling of speech intelligibility (Schiavetti, 1992). Due to the time and resources needed to carry out these approaches, they have not been employed in the majority of clinical settings (Ertmer, 2010).

Tasks involving speech intelligibility scaling for which DME is applied can use either standard value or free value scaling. According to Schiavetti (1992), a standard, or sample of speech chosen by the experimenter to constitute low, middle, or high intelligibility, is pre-assigned a value beforehand, most often 10 or 100, and is then used as a measurement to rate other stimuli against (Poulton, 1968). In contrast, there is no standard used in free modulus scaling; instead, listeners may assign any value to the first stimuli and rate subsequent samples relative to preceding stimuli (Schiavetti, 1992). Employing DME with a standard is often the preferred method because it alleviates discomfort of the listeners and avoids data complications (Engel, 1971). As a step toward developing a standard methodology for intelligibility assessment, the use of standard value DME is needed. (Willougby, 2019; Speights Atkins et. al, 2019).

Phonetic Contrasts

Although several methods for intelligibility assessment in children have been proposed, a widely adopted stimulus set that quantifies and explains the functional impact of decreased intelligibility has yet to be established. Studies involving phonetic contrast pairs have shown that particular error profiles contribute to intelligibility more so than supplementary acoustic parameters (Kent et al, 1990; Weismer, Kent & Rosenbeck, 1989). By investigating an explanatory approach to assessment of speech intelligibility for adult speakers with dysarthria, intelligibility across various phonetic contrast categories was measured to acquire an error profile disclosing the most frequently occurring. The speech stimuli utilized were single words representing nineteen different phonetic contrasts including paired phonemic variations with subtle differences based on the category. Intelligibility was judged by the listeners' ability to recognize the intended word or perceive it as the phonetic contrast pair; however, listener experience was not reported. The detectable phonetic contrast productions were then determined to cause intelligible or unintelligible speech (Kent, Weismer, Kent, & Rosenbek, 1989). Kent, Kent, Weismer, Sufit, Rosenbek, Martin, & Brooks (1990) explored this explanatory model further in a study in which phonetic contrast pairs were recorded by twenty-five speakers with amyotrophic lateral sclerosis (ALS) and scored by listeners through closed-set word recognition. Intelligibility was analyzed between all nineteen contrast groups, and results showed that the stop-nasal and initial glottal null contrasts were the highest contributors to unintelligible speech.

Explanatory intelligibility studies have been explored across the lifespan (see Table 1). As shown, inexperienced listeners and experienced listeners have been included as participants but not consistently controlled as variables. Evidence has shown that subjective rating of intelligibility is susceptible to variability between raters and experienced listeners are often able to habituate to disordered speech patterns in order to understand disordered speech to a greater degree than novice listeners (Ertmer, 2010; Gordon-Brannan &Hodson, 2000; Kent, Miolo, & Bloedel, 1994; Kent, 1996; Klein & Flint, 2006; Miller, 2013). Therefore, considering this listener bias is vital in determining whether intelligibility ratings within a clinical setting are transferable to real-world situations.

Table 1:
Previous Intelligibility Studies

Study	Year	Participants	Stimuli	Listener
Hodson & Paden	1981	Child Intelligible and Unintelligible	Single words	Trained graduate students
Billman	1986	Child Disordered	Unknown	Unknown
Kent et al.	1990	ALS	Phonetic contrasts	Unknown
Ansel & Kent	1992	Dysarthric with Cerebral Palsy	Phonetic contrasts	Trained listeners with varying experience levels
Turner et al.	1995	Dysarthric with ALS	Reading passage	Graduate students
Weismer et al.	2001	Dysarthic with ALS and Parkinson's disease	Sentence list	Undergraduate or graduate students with no extensive clinical experience
Klein & Flint	2006	Adults controlled phonological processing errors mimicking child speech	Sentence list	College students

Note: Several studies have investigated explanatory models for assessment of intelligibility with ratings from experienced and inexperienced listeners.

Justification

To further explore an explanatory model for speech intelligibility in children that controlled for listener experience, a preliminary study employed twenty-one inexperienced listeners as participants (Willoughby, 2019). Measures of whole-word accuracy and specific error types correlated with intelligibility when using phonetic contrast pairs. These results, however, did not explain whether listeners familiar with child speech provide ratings phonological categories were lower based on theories of listening bias in experts. Additional research is needed to determine if there is a listening bias for listeners with experience that results in the inconsistent subjective rating of intelligibility.

The purpose of this study was to determine if listening experience influences intelligibility rating of words categorized by phonetic contrast type. Intelligibility ratings of child speech productions that varied in levels of whole word accuracy were rated using direct magnitude estimation (DME) by two groups: listeners with no more than incidental experience with child speech and expert listeners, speech language pathologists who treat children with speech sound disorders. Recorded words were selected according to features reflecting eight phonetic categories. Listening experience was controlled by operationally defining inexperienced listeners as those who had more than ten hours per week of incidental child speech exposure. This study will investigate the perceptual response of expert listeners operationally defined as clinically certified SLPs with three or more years of clinical-based experience with child speech sound disorders to determine if findings differ based upon listening experience.

To investigate the influence of listening experience on the rating of intelligibility of phonetically-contrasted words in preschool age children, we addressed the following research questions:

(a) Is there a relationship between speaker accuracy group (high, mid, and low) and mean DME provided by experienced listeners? The null hypothesis was that there is no relationship. We hypothesized a correlation between rating of intelligibility and word accuracy (Kent, 1992; Speights Atkins et al., 2019; Willoughby, 2019).

(b) Is there a difference in DME and speaker group when inexperienced and experienced listeners are compared? We hypothesized that experienced listeners' DME will differ from inexperienced listeners due to listener bias intrinsic to clinical training.

Method

Speech Samples

Speech samples were retrieved from the Speech Evaluation and Exemplars Database (Speights, Boyce, & Willoughby, 2018). The speech samples, consisting of recordings of children recruited from a local early education center, were approved through an IRB protocol allowing speech samples to be conserved in a public speech database and retrieved for research use at a later time. Speech samples were recorded in a quiet room in which sound levels were measured beforehand to deduce that the environmental noise level was below 40 dBA SPL (Williams, Zhou, Stewart, & Knott, 2016). Each sample was recorded at a 44K sampling rate with 24-bit depth using a handheld H6N recorder with cardioid XLR MOVO LV402 microphones.

Speech samples consisted of words within eight different phonetic contrast categories: (1) stop-fricatives, (2) stop-affricates, (3) final cluster-final singletons, (4) fricative-affricates, (5) alveolar-palatals, (6) front-back vowels, (7) high-low vowels, and (8) initial cluster-initial singletons (Kent et al., 1989). These phonetic contrasts have been affiliated with reduced intelligibility in children with phonological based disorders (Bankson et al., 2013; DuHadway & Hustad, 2012; Skahan, Watson, & Lof, 2007).

Nine child speakers, varying in levels of speech sound development, were selected from the database. The sample included 5 males and 4 females. Ages of the children ranged from 3 years 4 months to 5 years 5 months old. Each child was assessed for the presence of a speech sound disorder using the Diagnostic Evaluation of Articulation and Phonology in which scores are based on a scale of 10 and a standard deviation of 3 (Dodd, Hua, Crosbie, & Ozanne, 2002). A score of 7 is one standard deviation below the mean and was used as the criterion for determining the presence of a speech sound disorder. Six children were determined to exhibit non-disordered speech, while three children were determined to exhibit disordered speech. Based on a preliminary analysis of phonetic contrast categories, the phonological processes observed included the stopping of affricates, final consonant deletion, velar fronting, cluster reduction, and backing. All child speakers demonstrated the following characteristics: (1) bilateral hearing at 20dB for 0.5 kHz, 1 kHz, 2 kHz, and 4 kHz; (2) American English as their primary language; (3) ability to orally communicate at least one-word utterances. A Beltone Audio Scout portable audiometer with fitted headphone cups was used to test hearing at all four frequencies.

Child speakers were categorized into three groups based on whole word production accuracy related to Proportion of Whole-Word Correctness (PWC) measures: high accuracy, medium accuracy, and low accuracy. PWC is a measure used to determine the proportion of words produced correctly out of an entire sample set. PWC was calculated from transcripts orthographically transcribed by two trained, graduate students. Each student independently completed the transcriptions with no prior knowledge of the stimulus list items. Binary scoring of transcriptions was used to determine correctness of each word, coding "0" for incorrect transcriptions (those not matching the intended word) and "1" for correct transcriptions (those matching the intended word) (Ingram, 2002). Interrater reliability was .84. The agreed upon transcriptions provided each speaker with a PWC score calculated by comparing the number of words produced correctly to the total number of words produced. PWC percentages were then compared to percentages of consonants correct (PCC) categories to inform classification of intelligibility level: mild (>85%), mild-moderate (65%-85%), moderate-severe (50%-65%), and severe (<50%) (Shriberg et al., 1997). Children exhibiting a PWC above 85% were placed in the high accuracy speaker group (SG), the children with a PWC between 50% and 84% were assigned to the mid accuracy SG, and those whose PWC fell below 50% were considered to be part of the low SG (See Table 2).

Table 2 Speaker Groups (SG)

	Speaker	**% Whole Word Correct (PWC)**	**Age**	**Sex**	**Disorder**
High accuracy= 100%-85%	H-1	88%	4;4	F	ND
	H-2	87%	4;2	M	ND
	H-3	85%	4;10	M	ND
Mid accuracy= 50%-84%	M-1	83%	4;1	M	ND
	M-2	70%	5;7	M	SSD
	M-3	55%	3;8	F	ND
Low accuracy= 0%-50%	L-1	38%	3;10	F	SSD
	L-2	24%	3;4	M	ND
	L-3	6%	5;5	F	SSD

Notes. Speakers categorized based on percentage of whole words correct.

Preparation of Speech Samples: Materials and Procedure

Sound file sets comprised of the entire stimulus word list were created for each of the speaker groups (high, mid, and low). The speaker group sets were then counterbalanced and randomized in order to create one stimulus list later presented to the listeners. Counterbalancing guaranteed that sound files from all speakers and all words in the stimulus list were distributed evenly. A minimum of 10 sound files produced by each of the nine child speakers were included in each list which served to reduce learning effects of stable child speech patterns. Lastly, the sound files were randomized individually in order to control for order effects. Each stimulus list began with the same nine sound files consisting of single syllable word productions from the Clinical Assessment of Articulation and Phonology produced by nine of the child speakers (Secord & Wayne, 2013). The purpose of the uniformity of these initial speech samples was two-fold: they provided additional practice opportunities for the listeners as well as allocated consistent stimulus items for each listener. Following these first nine sound files, each list included its counterbalanced and randomized sound files consisting of 91 phonetic contrasts. Every sound file included in the stimulus list was analyzed across 15 experienced clinicians and 15 inexperienced listeners for intelligibility measurement averages.

Adult Listeners

Two groups of adult listeners were recruited for this study, novice listeners (G1) and experienced listeners (G2). Adult inexperienced listeners (n=16; age *M=35.31, SD=10.27*) were recruited through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Crowdsourcing is a method of gathering information, most often through online recruitment, allowing large datasets to be completed simultaneously without the time necessity and inconvenience of scheduling each participant. Because of the substantial and diverse listener population that crowdsourcing provides, the results have been shown to be comparable to those obtained in natural environments resulting in a more reliable and ecologically valid measure of intelligibility (Byun, Halpin, & Szeredi, 2015). AMT enlists workers, or internet users, to complete jobs referred to as Human Intelligence Tasks (HITs). All AMT workers were provided access to our research experiment link, Intelli-turk©, and those who selected the link and agreed to complete the HIT were assigned specific token numbers and associated confirmation codes for de-identified administrator task review and compensation. Inclusion criteria entailed listeners to be at least 19 years of age, non-hearing-impaired, inexperienced with child speech, and speakers of American English as their primary language. After completing the informed consent, participants self-identified as being inexperienced with child speech by answering “no” to two questions: (1) Do you have a child who is currently 2-7 years old? (2) Do any of the following apply to you? *Pre-school or elementary faculty*, *a child instructor of any kind*, *a nanny/caretaker or babysitter*, *spend more than 10 hours a week listening to children ages 2-7 talk*. AMT workers were all residents of the United States, representing Western, North Central, Northern, Midland, and Southern dialectical regions.

The second group (G2), adult experienced listeners, were recruited through traditional recruitment methods: word-of-mouth, flyers, social media, the American Speech-Language-Hearing Association (ASHA) community, and the Speech and Hearing Association of Alabama (SHAA). Inclusion criteria required listeners to be at least 19 years of age (*M =42.62, SD* = 11.87), non-hearing-impaired, experienced with child speech, and speakers of American English as their primary language. All recruited G2 listeners were practicing clinicians in the field of Speech-Language Pathology with clinical experience ranging from 3 years to 38 years (*M* = 13.93, *SD* = 8.49). To determine experience the following three questions were asked: (1) How many years of clinical experience do you have? (2) What population do you primarily work within your clinical practice? *Adults or children?* (3) What is your area of expertise? All experienced listeners were residents of the United States, representing Western, North Central, Midland, and Southern dialectical regions.

Listening Experiment

The listening experiment included 100 words from eight different phonetic contrast categories and one non-contrast category. Listeners were asked to type the word they heard and rate the intelligibility using a sliding bar. Listener progress was tracked through the individual de-identified token numbers and confirmation codes generated through the Intelli-turk© administrator platform. Succeeding verification of complete participation of the experiment, listeners were compensated through AMT.

All listeners were instructed to complete the listening experiment in a quiet place while using headphones with the volume set at a comfortable listening level. Listeners were required to verify that their computers and headphones were functioning properly before advancing. Speech recognition ability was screened within the Intelli-turk© web application using the Word Intelligibility Picture Identification (WIPI) Test (Ross & Lerman, 1971). This word recognition task was initially designed for children but has been used to evaluate listener performance in adults for experimental purposes (Bradley & Sato, 2004; Ishikawa et al., 2017; Lenhardt, Skellet, Wang, & Clark, 1991; Papso & Blood, 1989).

Statistical Analysis

Descriptive statistics were used to evaluate the distribution of DME scores. Analyses were completed using R (Version 1.2.5033© 2009-2019 RStudio, Inc.) to determine significant differences in DME across listener Experience Groups for each speaker group (SG). A Linear Mixed Effects Regression (LEMR) with lme4 package was used for the analysis (Bates, 2014). This allowed participants to be entered as a random effect within the model to compare the interaction among the Listener Experience group designation and Speaker Groups for DME (Goldstein, Browne, & Rasbash, 2002; Jaeger, 2008; Nagaraj, 2017; Peng & Lu, 2012; Raudenbush & Bryk, 2002). Significance was set at alpha = 0.05.

Results

Descriptive Statistics

Descriptive Statistics for the DME reported (produced) across listener groups met normality assumptions following visual inspections of histogram and Q-Q Plots. The mean DME was 51.90 (±29.26) with the entire range of the scale used across participants and speaker groups. The median of 47.80 reflected the slight kurtosis to the lower anchor of the scale found in other studies that used speakers for the low PWC speaker group (Willoughby et al., 2019). Kurtosis was eliminated when DME values for each speaker groups were averaged, Table 3 provides the descriptive statistics for each Listener Experience group controlling for PWC speaker group.

Table 3. Descriptive Statistics DME (N=31)

	Experienced Listeners			**Inexperienced Listeners**		
	Low PWC	**Mid PWC**	**High PWC**	**Low PWC**	**Mid PWC**	**High PWC**
Mean	32.21	52.99	60.93	39.20	59.37	66.69
Median	30.21	51.95	62.77	38.19	62.19	69.38
Std. Dev.	9.09	8.28	11.03	10.52	10.08	11.48
Range	31.71	35.46	41.58	41.43	36.43	40.94

Note. Overall DME was normally distributed with kurtosis = -0.74, skewness = -0.24 and Shapiro-Wilk = .979 (93) *p* =.129.

Linear Mixed Effects Regression

A linear mixed effect regression (LMER) model was completed with the *DME* as the dependent variable, participants as the random variables, and Speaker Group entered as a repeated measure, listener Experience level, and the interaction (Speaker group x Experience

Group) set as fixed effects. For the linear effect, denoted ".L" in the speaker group were contrast coded (−1, 0, 1) such that negative beta values were associated with the Low PWC speaker group and positive values were associated with the High PWC SG. For the Quadratic effect, denoted with ".Q" contrast coding (1, -2, 1) for the three groups where negative beta values were associated with Mid PWC speaker group. Satterthwaite's method for a type III ANOVA indicated significant differences for PWC speaker group, *F*(2, 58) = 230.65; *p* < .001. No other significant differences were identified across Experience level, *F*(1, 29) = 3.74; *p* = .063. Author note the p-value for Experience level approached our alpha level. No significant difference was identified for the interaction among Experience level and Speaker groups, *F*(2,58) = 0.10; *p* = .904. Table 4 provided the contrast estimates and confidence intervals (CI) for the LMER model (*p* < .001).

Table 4. Results of linear mixed-effects regression on DME (N=31)

Fixed Effects

					95% CI	
	Estimate	**SE**	**Df**	**t Value**	**Lower**	**Upper**
Intercept	51.90	1.65	29	31.45***	48.68	55.13
Experience.L	–4.51	2.33	29	–1.93	–9.08	0.06
Speaker Group.L	19.87	0.96	58	20.77***	17.99	21.75
Speaker Group.Q	–5.24	0.96	58	–5.48***	–7.12	-3.30
Experience x Speaker Group.L	0.61	1.35	58	0.44	–2.04	3.26
Experience x Speaker Group.Q	<0.01	1.35	58	0.01	–2.65	2.65

Random Effects

	Variance	**Std.**
Participant (Intercept)	74.87	8.65
Error (Residuals)	28.35	5.33

Note. CI- confidence interval; estimate for the Intercept indicates the estimate DME grand mean; L designate contrast coding for Linear effects and Q designates the contrast coding for Quadratic effects. See R Script in Appendix.
*** p-value ≤ .001

The linear contrast for speaker contrast indicated that the listeners in both groups produced higher DME scores for the High PWC group compares to the Low PWC group (contrast estimate = 19.87; *p* < .001). The significant quadratic effect for speaker group indicates that all listeners produced significantly different DME values for the Mid PWC speaker group when compared to mean across Low and High PWC groups (contrast estimate = -5.24; *p* < .001). Figure 1 provides box plots for the box plots for each speaker group DME controlling for listener experience level (ExpLvl). The horizontal line within the box indicates the median. Boundaries of the box indicate the 25th and 75th percentiles. Lines indicate the 10th and 90th percentiles.

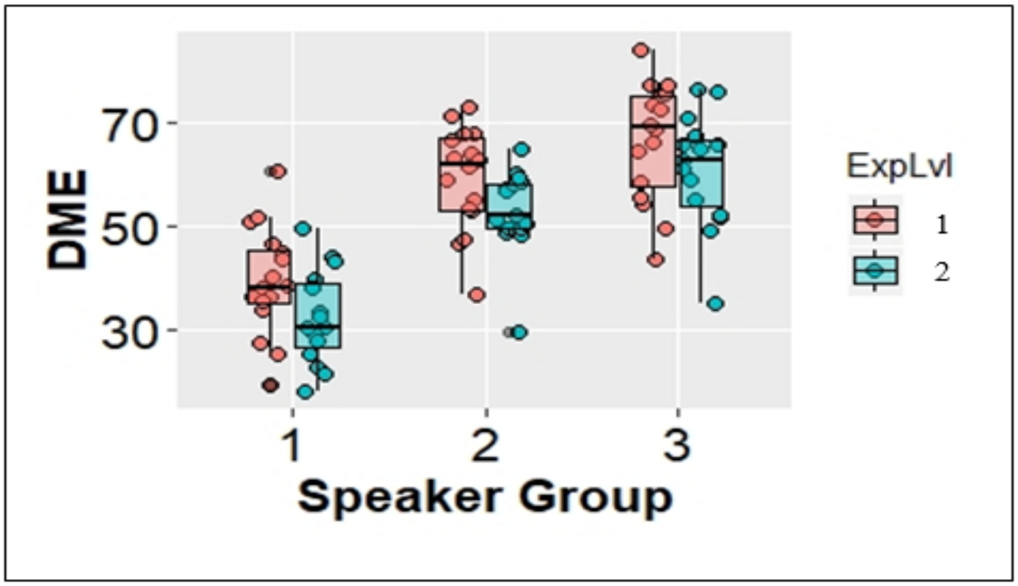

Figure 1: Direct Magnitude Estimation (DME) for low (1), mid (2) and high (3) speaker groups rated by inexperienced (1) and experienced (2) groups.

Discussion

In this study, intelligibility ratings by listeners inexperienced with child speech were compared to experienced speech-language pathologists to investigate whether a difference exists between how speech-sound disordered speech is perceived by clinicians in a clinic setting versus novice listeners in functional, real-world situations. Another aim of the study was to first determine if a relationship exists between speaker accuracy groups (high, mid, and low) and mean DME provided by both inexperienced and experienced listeners. In a previous research study, Willoughby (2019) found a correlation between speaker accuracy groups and word accuracy by inexperienced listeners, so it was important to note whether the same was found within an additional group of inexperienced listeners and experienced clinicians. Secondly, we explored if a difference exists between DME ratings and speaker groups when inexperienced listeners and experienced clinicians are compared. The underlying hypothesis of this study was that listeners inexperienced with child speech would rate intelligibility lower than experienced clinicians due to listener bias.

Speakers with Lower Proportions of Whole Word Correctness are Rated as Less Intelligible

The results of this study reinforce a correlation between measures of whole-word accuracy and overall intelligibility (Ingram & Ingram, 2001; Willoughby, 2019). A significant difference between the high, mid, and low speaker groups' word production accuracy was seen in the DME scores entered by both inexperienced and experienced listeners. This finding endorses the relationship between decreased word accuracy and decreased intelligibility. Shriberg and team (1997) established that clinical perception and severity measures can serve as primary tools that help gauge intelligibility concerns, and our finding endorses this notion.

Continuity in the Ratings of Inexperienced Listeners and Experienced Clinicians

The results of this study revealed no significant difference between DME ratings and speaker groups when inexperienced listeners and experienced clinicians were compared. Although this relationship is not what we hypothesized, our findings support a previous systematic review of crowdsourcing that revealed continuity in the ratings of inexperienced and experienced listeners (Willoughby, 2019). Sescleifer and colleagues (2018) completed a systematic review analyzing the advantage of crowdsourcing to evaluate perceptual speech outcomes in which 376 disordered speakers were given over 700,000 distinctive ratings by online workers through AMT. Within this review, five studies deliberately explored the relationship with an established measure (e.g. expert rating or an acoustic gold standard) and found that online workers' subjective ratings were highly consistent with current accepted measures of assessment (Sescleifer, Francoisse, & Alexander, 2018).

Unexpected Findings

The results of this study showed that speech-language pathologists (SLPs) rated intelligibility lower than inexperienced listeners. Although these results are opposite of what was originally hypothesized, a possible explanation of the lower DME ratings by SLPs is that, depending on work setting and years of experience, SLPs may hold professional biases in order for children to quality for services. For example, in school settings children have to score a minimum of two standard deviations below the average in order to quality for speech-language services.

Clinical Implications

This study was designed as the next step toward understanding challenges that listeners may experience in unknown contexts with unfamiliar speakers. Knowing if a listener bias exists is a vital precursor to the creation of a standardized measurement of intelligibility. At the current time, the perceptual rating of intelligibility is considered the typical measure of speech-sound disorder severity due to the lack of an objective method in place. The results of this study suggest that speech experts and outside listeners rate child speech in a similar fashion. In other words, clinicians in the clinic setting and listeners in the outside world may be experiencing the same difficulties with comprehending children with speech-sound disorders. Both listener groups identified categorical differences in intelligibility consistent with clinical impressions of severity measures of determined by the Proportion of Whole-Word Correctness. Because early intervention within this population is essential to their general success, this information can serve as valuable during the consideration of outside ratings and measurements of intelligibility in addition to those made within a speech-language pathology clinic setting. This also provides evidence to for the development of community-based tools to screen speech in young children and advancement of early identification of disordered speech by healthcare professionals in a manner similar to newborn hearing screening.

Limitations and Further Directions

This study was completed as an investigation to determine if there is a significant listener bias when perceptually rating child speech intelligibility. While speakers were divided in categories based on whole-word production accuracy, age and disorder types were not controlled. Enlarging the speaker sample size would accommodate for possible effects of age and disorder type. The small speaker sample size could have impacted overall intelligibility if mid-accuracy group measurements exceeded or were similar to those in the high speaker group. Future consideration of classification of children within the mid-range group is warranted. No group differences between the high-accuracy and low-accuracy groups and greater differences in the mid-accuracy group can be hypothesized; therefore, future studies may focus on larger mid-group samples. Furthermore, ratings of intelligibility using single words may not be representative of what a listener is experiencing during multiple word utterances or conversation; however, single word intelligibility scores were obtained to evaluate DME.

Future studies exploring differences between inexperienced and experienced perception of speaker intelligibility would profit from larger listener groups. There was no significant difference found in the ratings of inexperienced and experienced listeners; however, it would be important to note whether the trend for experienced clinicians to rate intelligibility lower than inexperienced listeners continued. While mild versus severe disorders are typically easier to categorize, mid-accuracy group studies may be more compelling if listener experience differences are considered. It is important for future studies to evidence no listening bias for the particular word set used. Due to the absence of a listening bias in this study, future studies using DME scores for building machine-learning models are warranted.

Lastly, incorporating a third listening group comprised of experienced listeners other than SLPs, such as pediatric nurses or elementary school teachers, would be beneficial for future studies. This would allow for a greater understanding how speech intelligibility is rated and possibly support the incorporation of speech screenings within medical clinics and/or classrooms.

Acknowledgements

The research was supported by funding from the College of Liberal Arts and the Department of Communication Disorders at Auburn University. This project was supported by numerous undergraduate and graduate students within the Technologies for Speech-Language Research Lab. Bayley Smith and Bailey Evans provided transcriptions of words for calculation of whole-word accuracy. Caroline Willis helped prepare and organize the speech samples used in the

study.

References

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research*, *35*(2), 296–308.

Bankson, N. W., Bernthal, J. E., & Flipsen, P. (2013). Speech sound assessment procedures. *Articulation and phonological disorders: Speech sound disorders in children*, 177-211.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beukelman, D. R., & Yorkston, K. M. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders, 13*(1), 33-41.

Boonen, N., Kloots, H., Verhoeven, J., & Gillis, S. (2019). Can listeners hear the difference between children with normal hearing and children with a hearing impairment? *Clinical Linguistics & Phonetics, 33*(4), 316-333.

Borrie, S. A., McAuliffe, M. J., Liss, J. M., Kirk, C., O'Beirne, G. A. and Anderson, T. (2011). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes, 27*(7-8), 1039-1055.

Bradley, J. S., & Sato, H. (2004). Speech intelligibility test results for grades 1, 3, and 6 children in real classrooms. *Proceedings 18th International Congress on Acoustics,* Kyoto.

Burns, E. M., & Ward, W. D. (1978). Categorical perception—phenomenon or epiphenonmenon: evidence from experiments in the perception of melodic musical intervals. *The Journal of Acoustical Society of America, 63*(2), 456-468.

Byun, T. M., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders*, *53*, 70–83.

Connolly, J. H. (1986). Intelligibility: a linguistic view. *International Journal of Language & Communication Disorders*, *21*(3), 371–376.

D'Innocenzo, J., Tjaden, K., & Greenman, G. (2006). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical Linguistics & Phonetics, 20*(9). 659-675.

Dale, E. W., Plumb, A.M., Sandage, M.J., & Plexico, L.W. (2019). Speech-language pathologists' knowledge and competence regarding percentage of consonants correct. *Communication Disorders Quarterly*, 00(0), 1-9.

Dodd, B., (1995). *The differential diagnosis and treatment of children with speech disorder.* San Diego, CA: Singular Publishing Group, Inc.

Dodd, B., Hua, Z., Crosbie, S., & Ozanne, A. (2002). *Diagnostic Evaluation of Articulation and Phonology: DEAP*. London, England: Psychological Corporation Ltd.

DuHadway, C. M., & Hustad, K. C. (2012). Contributors to intelligibility in preschool-aged children with cerebral palsy. *Journal of Medical Speech-Language Pathology*, *20*(4).

Edition, F., & Bauman-Waengler, J. (2012). *Articulatory and phonological impairments: A clinical focus*. Oxnard, CA: Pearson.

Engel, T. (1971). Psychophysics II. Scaling methods. In J. W. Kling & L. Riggs (Eds.), *Woodworth and Schlossberg's experimental psychology* (pp. 63-80). New York: Thieme.

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.

Flipsen, P. (1995). Speaker-listener familiarity: Parents as judges of delayed speech intelligibility. *Journal of Communication Disorders, 28*(1), 3-19.

Francis, A. L., Nusbaum, H. C., & Fenn, K. (2007). Effects of training on the acoustic-phonetic representation of synthetic speech. *Journal of Speech, Language & Hearing Research, 50*(6), 1445-1465.

Goldman, R., & Fristoe, M. (2015). *Goldman Fristoe Test of Articulation. 3.* Minneapolis MN: Pearson Education, Inc., & PsychCorp (Firm).

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*(1), 585-612.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews. Cognitive Science, 1*(1), 69-78.

Gordon-Brannan, M. (1994). Assessing intelligibility: Children's expressive phonologies. *Topics in Language Disorders*, *14*(2), 17–25.

Hustad, K. C., Oakes, A., & Allison, K. (2015). Variability and diagnostic accuracy of speech intelligibility scores in children. *Journal of Speech, Language, and Hearing Research*, *58*(6).

Ingram, D., & Ingram, K. D. (2001). A whole-word approach to phonological analysis and intervention. *Language, speech, and hearing services in schools*.

Ishikawa, K., MacAuslan, J., & Boyce, S. (2017). Toward clinical application of landmark-based speech analysis: Landmark expression on normal adult speech. *The Journal of the Acoustical Society of America, 142*(5), EL441-EL447.

Kent, R.D. (1992). *Intelligibility in speech disorders* Philadelphia, PA: John Benjamins Publishing Company.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7-23.

Kent, R. D., Kent, J. F., Weismer, G., Sufit, R. L., Rosenbek, J. C., Martin, R. E., & Brooks, B. R. (1990). Impairment of speech intelligibility in men with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Disorders*, *55*(4), 721–728.

Kent, R. D., 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.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499.

Klein, H. B., Grigos, M. I., McAllister Byun, T., & Davidson, L. (2012). The relationship between inexperienced listeners’ perceptions and acoustic correlates of children’s /r/ productions. *Clinical Linguistics & Phonetics, 26*(7), 628-645.

Klimacka, L., Patterson, A., & Patterson, R. (2001). Listening to deaf speech: Does experience count? *International Journal of Language & Communication Disorders, 36,* 210-215.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33*(1). 103-115.

Kwiatkowski, J., & Shriberg, L. D. (1992). Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. *Journal of Speech, Language, and Hearing Research*, *35*(5), 1095-1104.

Lansford, K. L., Borrie, S. A., & Bystricky, L. (2016). Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *American Journal of Speech-Language Pathology*, *25*(2), 233-239.

Lenhardt, M. L., Skellet, R., Wang, P., & Clark, A. M. (1991). Human electronic speech perception. *Science, 253( ), 82-85.*

Logan, N. R. (2010). *Methods used to assess intelligibility in children with phonological* *disorders: Results of a national survey*. Retrieved from the University of Central Missouri.

Lyons, I. M., Mattarella-Micke, A., Cieslak, M., Nusbaum, H. C., Small, S. L., & Beilock, S. L. (2010). The role of personal experience in the neural processing of action-related language. *Brain & Language, 112*(3), 214-222.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7/8), 953-978.

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Thirteenth Annual Conference of the International Speech Communication Association*.

McGarr, N. S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research, 26*(3), 451-458.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders*, *48*(6), 601–612.

Munson, B., Johnson, J. M., & Edwards, J. (2012). The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology, 21*(2), 124-139.

National Institute on Deafness and Other Communication Disorders (2016). *Quick Statistics about voice, speech, language, and swallowing.* Retrieved from: https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

Papso, C. F., & Blood, I. M. (1989). Word recognition skills of children and adults in background noise. *Ear and Hearing, 10*(4), 235-236.

Parson, J., Braga, D., Tjalve, M., & Oh, J. (2013, September). Evaluating voice quality and speech synthesis using crowdsourcing. In *International Conference on Text, Speech and Dialogue* (pp. 233-240). Springer, Berlin, Heidelberg.

Ross, M., & Lerman, J. (1970). Word Intelligibility by Picture Identification. *Journal of Speech and Hearing Research, 13,* 44-53.

Secord, W., Donohue, J. A. S. (2002). CAAP: Clinical Assessment of Articulation and Phonology. Greenville, S.C: Super Duper Publications.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement, and management* (pp. 11-34). Amsterdam: John Benjamins Publishing Company.

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research*, *24*(3), 441–445.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, *40*(4), 708–722.

Smiljanic, R., & Chandrasekaran, B. (2013). Processing speech of varying intelligibility. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 060102). ASA.

Speights Atkins, M., Boyce, S. E., & Willoughby, K. E., (2018). SEED- Speech Exemplars and Evaluation Database. Auburn University Technologies for Speech-Language Research Lab. Permanent URL: http://hdl.handle.net/11200/49140

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. *American Journal of Speech-Language Pathology*, *16*(3), 246–259.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). Oxford, England: Wiley.

Stevens, S. S. (1986). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New Brunswick, NJ: Transaction Publishers.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of experimental psychology*, *54*(6), 377.

Tjaden, K., & Liss, J. M. (1995). The influence of familiarity on judgments of treated speech. *American Journal of Speech-Language Pathology, 4*(1), 39-48.

Utianski, R., Lansford, K., Liss, J., & Azuma, T. (2011). Effects of topic knowledge on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *Journal of Medical Speech-Language Pathology, 19*(4), 25-36.

Verhoeven, J., De Pauw, G., Pettinato, M., Hirson, A.,Van Borsel, J., & Mariën, P. (2013). Accent attribution in speakers with Foreign Accent Syndrome. *Journal of Communication Disorders, 46*(2), 156-168.

Weismer, G., Martin, R., & Kent, R. D. (1992). Acoustic and perceptual approaches to the study of intelligibility. *Intelligibility in Speech Disorders*, 67–118.

Williams, W., Zhou, D., Stewart, G., & Knott, P. (2016) The practicality of using a smart phone 'App' as an SLM and person noise exposure meter. *Proceedings of ACOUSTICS* Brisbane, Australia.

Willoughby, K. (2019). Exploring an Explanatory Child Speech Intelligibility Model Using Phonetically Contrasted Word Productions.

Zatorre, R. J., & Halpern, A. R. (1979). Identification, discrimination, and selective adaption of simultaneous musical intervals. *Perception & Psychophysics, 26*(5), 384-395.