COMPARISON OF INTELLIGIBILITY MEASURES FOR ADULTS WITH PARKINSON'S DISEASE, MULTIPLE SCLEROSIS AND HEALTHY CONTROLS

By

Kaila L. Stipancic
June 15, 2015

A thesis submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfillment of the requirements for the
degree of

Master of Arts

Department of Communicative Disorders and Sciences

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Acknowledgement

I would like to thank my committee members Dr. Kris Tjaden and Dr. Elaine Stathopoulos for all of their help and guidance throughout this project. I would also like to thank Dr. Gregory Wilding for his help with the statistical analysis for this project and Jennifer Lam for her editing eye and willingness to answer all of my questions.

I would especially like to thank Dr. Tjaden for her countless hours spent meeting with me, answering my questions, and reading over this document. This project would not have been possible with her. Dr. Tjaden, thank you for investing in me over the past two years. This experience has been invaluable and has expanded my excitement for research and heightened my skills in a multitude of areas. I feel prepared to take on any challenge that comes my way and am so thankful for your impact on me throughout my MA career.

Lastly, I would like to thank my family and friends for their support. To my MA classmates: thanks for your willing participation in this project and for always asking about the progress; we did it, SLPs! To my friends: thanks for understanding that "I have to write" is a legitimate excuse for postponing hangouts, that Friday nights sometimes have to be spent in the lab, and for celebrating the little victories with me. To my family- my momma, dad, brother, and sister-in-love: thank you for always asking how the work was coming, for acting interested even when you had no idea what I was talking about, and for prayerfully and lovingly supporting me through it all. I am eternally grateful.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Purpose: The current study sought to investigate the relationship between two metrics of sentence intelligibility in adults with Parkinson's Disease (PD), Multiple Sclerosis (MS), and healthy controls. An objective measure of intelligibility, orthographic transcription, and a subjective measure of intelligibility, Visual Analog Scaling (VAS), were the two metrics of intelligibility examined. Areas of interest included 1) comparisons of the pattern of intelligibility change in transcription and VAS, 2) strength of the relationship between these two types of intelligibility measures, and 3) differences in intralistener and interlistener reliability between the two metrics.

Methods: 78 speakers and the speech samples reported in Tjaden, Sussman, and Wilding (2014) and Kuo, Tjaden, and Sussman (2014) were used in the current study. The pool of 78 speakers consisted of 32 healthy control speakers, 16 speakers with PD, and 30 speakers with MS. Speakers read Harvard Psychoacoustic Sentences in habitual, clear, fast, loud, and slow conditions. In Tjaden et al. (2014) and Kuo et al. (2014), 50 naive listeners used a VAS on a computer to estimate how much of the speaker's message was understood (e.g., from 'didn't understand anything' to 'understand everything'). In the current study, 50 naive listeners heard the same stimuli, but were instructed to type exactly what they heard. Responses were scored to obtain a percentage of key words transcribed correctly for each stimulus. Results from the current study were compared to results from the VAS task studies (Tjaden et al., 2014; Kuo et al., 2014) using descriptive statistics (e.g., mean, standard deviation, etc.), parametric statistics (e.g., multivariate linear model fit to the data in this repeated measured design), correlation analyses (e.g., between the two metrics), and metrics of reliability.

Results and Discussion: Results revealed that the pattern of transcription intelligibility scores was very similar to scaled intelligibility derived from VAS. However, transcription scores were higher in magnitude than the VAS scores. In addition, correlation analyses showed the two intelligibility measures were highly correlated. Last, both interlistener and intralistener reliability were marginally higher for the VAS reported in Tjaden et al. (2014) and Kuo et al. (2014) than for the transcription data in the current study. These results suggest that a less time-consuming task, such as the VAS task, may be a viable substitute for a more time-consuming transcription task when documenting intelligibility in a clinical population to obtain an overall metric of severity for tracking disease progression and/or treatment progress.

Introduction

Parkinson's Disease

Parkinson's disease (PD) is a degenerative neurologic disease that involves a depletion of dopamine in the brain. PD is often characterized by resting tremors, muscular rigidity, a delay in movement initiation (i.e., akinesia), a slow and reduced range of motion (i.e., bradykinesia), and postural and gait disturbances (Yorkston, Beukelman, Strand, & Hakel, 2010). These motor symptoms are also often present in the speech subsystems. The most common types of speech production impairments due to motor symptoms in individuals with PD are disturbances in prosody (e.g. monopitch, monoloud, reduced stress), articulation (e.g. imprecise consonants), and rate (e.g. variable rate) (Duffy, 2007). These speech production impairments have been perceptually described as a dysarthria. It has been estimated that between 60% and 90% of individuals with PD also present with dysarthria (Duffy, 2013; Logemann, Fisher, Boshes, & Blonsky, 1978; Mackenzie, 2011). The speech impairments of dysarthria secondary to PD often lead to decreased intelligibility in these individuals (Cannito et al., 2012; Kent & Kim, 2011).

Multiple Sclerosis

Multiple Sclerosis (MS) is a progressive neurologic disease that involves degeneration of the myelin that covers axons in the central nervous system (Benedict et al., 2004). This myelin sheath typically allows for quick and efficient transmission of nerve impulses that are required for precise and accurate motor movements (Duffy, 2013). Additionally, MS may be characterized by atrophy in the gray matter of the brain (Benedict et al., 2006; Benedict et al., 2004). Consequently, MS is often characterized by motor disturbances, such as weakness and spasticity, as well as visual and sensory disturbances (Duffy, 2007). These motor disturbances are often present in the speech subsystems. The most prominent, deviant speech characteristics,

which are presumed to be due to these motor disturbances in individuals with MS, involve impaired control of loudness, imprecise articulation, and harsh vocal quality (Duffy, 2007). These speech production impairments have been perceptually described as a dysarthria (Duffy, 2013). It has been estimated that between 40% and 50% of individuals with MS also present with dysarthria (Mackenzie, 2011). The speech impairments in dysarthria secondary to MS are thought to contribute to reduced intelligibility in individuals with MS (Kent & Kim, 2011).

Dysarthria

Dysarthria has been defined as a group of neurologic speech disorders that involve impairments in any of the five components of speech: respiration, phonation, resonance, articulation, and prosody (Duffy, 2007). These impairments are presumably a result of abnormalities in the central or peripheral nervous systems (Duffy 2007; 2013). According to Yorkston et al. (2010), the motor speech impairment in dysarthria "is characterized by slow, weak, imprecise, or uncoordinated movements of the speech musculature" (p. 4). Individuals with dysarthria have been reported to be difficult to understand, due to their abnormally perceived rate, precision, and coordination of speech subsystems (Duffy, 2013; Yorkston et al., 2010). Therefore, individuals with dysarthria have a motor speech impairment that commonly results in poor intelligibility.

Intelligibility

Intelligibility refers to the degree, or accuracy, with which a listener recovers the acoustic signal or message produced by a speaker (Duffy, 2013; Hustad, 2008; Kent, Weismer, Kent, & Rosenbek, 1989). Schiavetti (1992) stated that intelligibility describes "the match between the intention of the speaker and the response of the listener to the speech passed through the transmission system" (p.13). Therefore, intelligibility is a product of both speaker and listener

efforts, as well as the message itself and the communicative context in which the message is delivered (Hustad & Weismer, 2007). In other words, the intelligibility of a speaker with dysarthria reflects both the impaired speech signal, as well as the strategies used by the listener to recover the signal (Duffy, 2013). In the dysarthria literature, intelligibility has been measured in two primary ways. First, intelligibility has been measured by calculating linguistic units correct (e.g., phonemes or words). Second, intelligibility has been measured by examining a listener's global understanding of the message (Kent & Kim, 2011). In these ways, intelligibility measures seek to quantify the understandability of a speaker.

Endeavors to quantify intelligibility are both clinically and theoretically relevant, as intelligibility is often a central outcome measure of speech-language therapy (Miller, 2013). Intelligibility has also been described as how effective one is in their communication (Cannito et al., 2012). Therefore, decreased intelligibility can significantly affect quality of life. As such, treatment techniques aimed at increasing intelligibility are pivotal to speech-language therapy (Hustad & Weismer, 2007). Because intelligibility is often negatively affected in individuals with PD and MS, strategies to maximize intelligibility are often central to speech therapy protocols for these individuals. Quantifying intelligibility is necessary in order to demonstrate the effectiveness of these techniques in improving intelligibility.

Techniques to Improve Intelligibility

Global behavioral techniques are widely recommended and are aimed at improving intelligibility. These behavioral techniques include, but are not limited to, rate manipulation, increased sound pressure level, and clear speech (Tjaden, Richards, Kuo, Wilding, & Sussman, 2013; Tjaden, Sussman, & Wilding, 2014; Yorkston et al., 2010).

Rate manipulation. Rate manipulation can involve either a reduction or an increase in speech rate. Speech rate is defined as speech output divided by the amount of time and is typically measured in number of words or syllables per minute or second (Kent & Kim, 2011; Yorkston et al., 2010). A slowing of speech rate may be useful for individuals with dysarthria who exhibit a faster than normal speech rate, which constitutes a small minority of dysarthric speakers, and also for those who exhibit a slower than normal speech rate, which constitutes a much larger proportion of dysarthric speakers. A reduction in speech rate is thought to benefit intelligibility by increasing the ability of individuals with dysarthria to obtain more 'normal' positioning of articulators by allowing more time for required movements to take place (Yorkston et al., 2010). A reduced speech rate may also supply listeners with increased processing time in which to decipher the intended message of the speaker (Hustad & Weismer, 2007). Overall, a reduction in speech rate may benefit intelligibility by affording both the speaker and the listener additional time.

Increasing speech rate may be a useful technique for individuals who exhibit a slower than normal speech rate. A slow rate of speech may lead to phoneme prolongations and a greater number of silent pauses, which both contribute to listener difficulty in parsing the acoustic signal (Dagenais, Garcia, & Watts, 1998). With healthy speakers, listeners attempt to parse utterances for comprehension while the message is being spoken. However, for speakers with dysarthria, listeners may have difficulty with parsing on-line, due to the irregularities mentioned above, and may have to wait until the end of the message before attempting to decipher meaning (Dagenais et al., 1998). This additional wait time may tax the working memory of listeners and cause degradation of the message before comprehension can occur. Thus, an increase in speech rate may allow parsing after the message has been spoken to be more viable (Liss, 2007). Increasing

speech rate has seldom been reported in the context of a therapeutic program and has often been thought to contribute to a reduction in intelligibility (Tsao, Weismer, & Iqbal, 2006). However, some studies have found a faster than habitual speech rate to be correlated with improved speech naturalness (Dagenais, Brown & Moore, 2006; Logan, Roberts, Pretto, & Morey, 2002). Speech naturalness has been found to be associated with increased intelligibility (Yorkston, Hammen, Beukelman, & Traynor, 1990). Thus, an increase in speech rate may also be associated with improved perceptual outcomes for speakers with a faster than normal or near normal speech rate. However, divergent evidence is available on this subject. Some studies have found that ratings of speech acceptability, which is likely correlated with speech naturalness, is strongly correlated with intelligibility (Whitehall, Ciocca, & Yiu, 2004), while others have found that speech acceptability is not correlated with intelligibility (Hanson, Beukelman, Fager, & Ullman, 2004). At this time, it is uncertain for which individuals either increasing or decreasing speech rate will be effective (Van Nuffelen, De Bodt, Vanderwegen, Van de Heyning, & Wuyts, 2010; Van Nuffelen, De Bodt, Wuyts, & Van de Heyning, 2009). It is equally uncertain how this rate manipulation strategy may contribute to increased intelligibility.

Increased sound pressure level. Increased vocal intensity is another global therapy technique with the potential to enhance intelligibility in individuals with dysarthria (Cannito et al., 2012; Yorkston, Hakel, Beukelman, & Fager, 2007). When individuals with dysarthria increase their intensity, it simply makes it easier for listeners to hear the message being produced. Cannito et al. (2012) found significant improvement in intelligibility for six out of eight speakers with PD following Lee Silverman Voice Treatment (LSVT™), which exclusively targets an increase in intensity. Additionally, there is some evidence to suggest that increased Sound Pressure Level (SPL) may also lead to beneficial changes in articulation and rate, which

may provide further increases in intelligibility (Miller, 2013). A variety of studies that elicited an increased intensity using stimulation support the use of therapy that increases SPL in order to enhance intelligibility in this population (Cannito et al., 2012; El Sharkawi et al., 2002; Neel, 2009; Ramig, Bonati, Lemke, & Horrii, 1994; Ramig, Countryman, Thompson, & Horii, 1995; Tjaden and Wilding, 2004).

Clear speech. A clear speaking style referred to as "clear speech" has been described in the literature as involving exaggerated articulation, a reduced rate, and an increased vocal intensity (Hustad & Weismer, 2007). Several studies found that when healthy speakers used a clear speech style, intelligibility was increased by between 11 and 35 percentage points (Ferguson, 2012; Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Picheny, Durlach, & Braida, 1985; Schum, 1996; Uchanski, Choi, Braida, Reed, & Durlach, 1996). As well, increases of intelligibility by approximately eight percent were found when individuals with dysarthria utilized clear speech (Beukelman, Fager, Ullman, Hanson, & Logemann, 2002; Hanson et al., 2004; Tjaden et al., 2014). Thus, using a clear speaking style shows promise for enhancing intelligibility in individuals with dysarthria.

Methods of Measuring Intelligibility

The ability to measure intelligibility is critical for quantifying the overall degree of communication impairment. Additionally, by measuring intelligibility over time, treatment effects and disease progression can be quantified. Choosing an appropriate method to measure intelligibility can be difficult, as there are advantages and disadvantages associated with each. The following review focuses on sentence-level metrics of intelligibility. Typically in everyday conversation, speech is produced in utterances comprised of multiple words, rather than in single words or phonemes. Therefore, sentence-level metrics of intelligibility are presumed to index the

magnitude of an individual's communicative difficulty (Weismer, 2009). Primarily, transcription and scaling tasks are used to measure intelligibility at the sentence level.

Orthographic transcription. Transcription has been characterized as an objective measure of intelligibility (Hustad & Weismer, 2007; Miller, 2013; Weismer, 2009) and involves the listener writing the speaker's message word-for-word (Kent & Kim, 2011). The word-for-word transcription is then compared to the target production and the percentage of words correctly transcribed is calculated (Hustad & Weismer, 2007; Kent & Kim, 2011). Percent correct scores derived from orthographic transcription are presumed to reflect the magnitude of an individual's speech impairment and can provide a severity measure relative to normal speech (Hustad & Weismer, 2007; Weismer, 2009). Although orthographic transcription is time-consuming for both the listeners who must transcribe and for the individuals who score the transcriptions, transcription is the gold standard for quantifying intelligibility in the dysarthria literature (Yorkston, Beukelman, & Traynor, 1984). Furthermore, transcription is the only measure that allows for evaluation of the listener's role in intelligibility, or analysis of the cognitive-perceptual processes that the listener is using to recover the intended message (Liss, 2007). Transcription has been considered to yield good consistency and reliability; however, only a few studies that used transcription reported listener reliability (Bunton, Kent, Kent, & Duffy, 2001; Tjaden, Kain, & Lam, 2014; Tjaden & Wilding, 2010).

Scaling tasks. While transcription has been described as an objective measure of intelligibility, scaling tasks have been characterized as more subjective measures of the same phenomenon, as listeners are instructed to estimate how much of the speaker's message they understood or to judge the extent to which the message was understood (Hustad, 2006b; Hustad & Weismer, 2007; Miller, 2013). Several types of scaling tasks have been discussed in the

literature. These types of scaling tasks include equal-appearing interval scaling, Direct-Magnitude Estimation (with or without a modulus), Visual Analog Scaling, and other estimation procedures (Hustad, 2006b; Hustad & Weismer, 2007; Schiavetti, 1992). Equal-appearing interval scaling requires listeners to designate a number, most often on a five, seven, or nine-point scale, that represents their perception of intelligibility for a given speech sample (Hustad & Weismer, 2007; Schiavetti, 1992). While this is one of the most commonly used clinical methods for measuring intelligibility (Schiavetti, 1992), anecdotal evidence has indicated that listeners have difficulty perceiving intelligibility in a linear fashion, and as such, cannot break intelligibility into equal intervals. For this reason, interval scaling has been deemed an inappropriate way to measure intelligibility (Schiavetti, 1992). Direct-Magnitude Estimation (DME) requires listeners to assign a numerical value to speech samples to represent proportional differences between the speech samples that they hear (Hustad & Weismer, 2007). Speech samples may be scaled relative to a modulus, or a standard example. Additionally, speech samples may be scaled relative to each other, which is called modulus-free DME. In the latter case, listeners assign the first speech sample they hear any number. Listeners then scale subsequent samples proportionally to the first (Schiavetti, 1992). While this abates the problem of listeners having to break intelligibility into equal intervals, there is anecdotal evidence that listeners find the task unusual and difficult (Schiavetti, 1992; Tjaden & Wilding, 2004). Visual Analog Scaling (VAS) involves listeners choosing a point on a continuous line that does not contain any ticks or intervals other than at the endpoints to represent their perception of a speaker's intelligibility (Kent & Kim, 2011). Lastly, estimation procedures, such as one used by Hustad (2006b), simply ask listeners to estimate the percentage of words that they understand

from a speaker's message. All of the methods mentioned above constitute scaling tasks for quantifying intelligibility.

Scaling tasks for quantifying intelligibility have been criticized in the dysarthria literature. Consistency and reliability for these tasks have been questioned and, in certain cases, have been found to be poorer than is ideal for research purposes (Miller, 2013; Schiavetti, 1992). Furthermore, scaling tasks provide little, if any, insight into the listener's role in intelligibility. Listeners' internal yardsticks differ on what, for example, counts as a 3/10 or a 5/10 on a scaling task (Miller, 2013). Additionally, when using scaling tasks, listener error patterns cannot be investigated. When only a single number is used to reflect how intelligible a listener deems a speaker's message to be, this data has no explanatory capacity. That is to say, a single number is simply a marker of speech disorder severity, rather than a way to evaluate what caused the reduction or increase in intelligibility (Liss, 2007). Scaling tasks, however, provide some attractive benefits, as they are less time-consuming and labor-intensive than orthographic transcription (Miller, 2013).

Comparison of intelligibility measures. Few studies have directly compared intelligibility metrics for the same speech materials or stimuli. In the literature for other populations, such as hearing-impaired individuals, the comparison between VAS and transcription has been examined (Huttunen & Sorri, 2004; Samar & Metz, 1988). In the dysarthria literature, several studies have reported multiple intelligibility measures (i.e., DME vs. transcription) for different types of speech stimuli (e.g., Metz, Schiavetti, Samar, & Sitler, 1990; Sussman & Tjaden, 2012; Yunusova, Weismer, Kent, & Rusche, 2005). Tjaden and Wilding (2010), for example, found a significant correlation between percent correct scores from transcription and scaled estimates of intelligibility for a reading task. This result may suggest that

transcription and scaled estimates tap into the same perceptual phenomenon. However, in the same study, it was found that there was no relationship between percent correct transcription scores for a reading task and scaled estimates of intelligibility for extemporaneous speech (Tjaden & Wilding, 2010). This suggests that results from scaling tasks and transcription may become more distinct when applied to different types of speech stimuli.

In one of the few studies directly comparing intelligibility metrics for the same stimuli, Hustad (2006b) found that for four speakers with dysarthria, overall, transcription scores were higher than when scores were obtained from listeners estimating the percentage of words that they understood from a speaker's message. However, the magnitude of difference between transcription and estimation scores varied from speaker to speaker. As well, it can be inferred that scaling tasks provide a more conservative estimate of intelligibility relative to orthographic transcription. Hustad (2006b) suggested that subjective scaling tasks may not be as reliable as an objective transcription measure. While previous studies have compared different types of intelligibility metrics, to date orthographic transcription and VAS have not been directly compared in previous dysarthria literature.

The Current Study

Historically, intelligibility has been measured using transcription. The original Assessment of Intelligibility of Dysarthric Speech (Yorkston et al., 1984) is undoubtedly the most widely used clinical tool for quantifying intelligibility. In the sentence portion of this tool, intelligibility is measured by calculating the number of words correctly transcribed and comparing this to total number of words in each sentence. In more recent years, computerized scoring for the Assessment of Intelligibility of Dysarthric Speech (1984) has become available. However, scoring accuracy is still susceptible to typing errors and homonyms, and thus, scoring

is not entirely automated. Therefore, some human scoring and, at the very least, examination and editing is required. Due to the fact that transcription is labor-intensive for both the listeners and for the people who score the accuracy of responses, less time-consuming methods of calculating intelligibility are attractive. Although transcription is the gold standard in the field for measuring intelligibility, other methods, such as the VAS task described previously, have been proposed for quantifying intelligibility.

Limited knowledge is available for how objective and subjective metrics of intelligibility compare for the same stimuli. Additionally, since the Assessment of Intelligibility of Dysarthric Speech (Yorkston et al., 1984) is the most widely used clinical tool for quantifying intelligibility in dysarthria, it is interesting to note that few, if any studies, have compared metrics of sentence intelligibility. If the two types of intelligibility metrics that are of interest here, namely orthographic transcription and VAS, are found to yield equivalent levels of severity, then there may be instances when the less time and labor-intensive scaling task could be used. This would support using a scaling measure to quantify intelligibility in an efficient way in both research and clinical settings, assuming that listener error patterns were not of interest.

Purpose. The purpose of the current study was to compare an objective intelligibility metric, specifically orthographic transcription, with a subjective intelligibility metric, specifically a VAS task. Orthographic transcription results from the current study will be compared to VAS data reported in Tjaden et al. (2014) and Kuo et al. (2014). In Tjaden et al. (2014), 50 listeners used a VAS to judge the sentence intelligibility of adult speakers with MS, PD, and healthy controls in five speaking conditions (habitual, clear, fast, loud, and slow). A detailed explanation regarding the elicitation of these conditions has been provided in the methods section. After listening to a sentence, listeners were instructed to indicate their intelligibility judgment by using

a mouse on the computerized scale. The scale was a 150 mm vertical line that had endpoints labeled with "Understand everything" to "Cannot understand anything". Software converted responses to numerical values ranging from 0 (i.e., *Understand everything*) to 1.0 (i.e., *Cannot understand anything*). Results showed that the loud and clear conditions improved intelligibility relative to the habitual condition. Rate manipulation (i.e., both slow and fast conditions) did not improve intelligibility relative to the habitual condition. This pattern of results held for all speaker groups (i.e., PD, MS, and healthy controls) (Kuo et al., 2014; Tjaden et al., 2014).

Research questions. Using the same speakers, speech materials (i.e., lists of Harvard sentences), and procedures as Tjaden et al. (2014) and Kuo et al. (2014), the current study quantified intelligibility using orthographic transcription. The overall goal of this project was to examine how percent correct intelligibility scores derived from orthographic transcription compare to intelligibility judgments derived from a scaling task. Specifically, the following research questions were of interest:

1. Is the pattern of intelligibility change the same between speaking conditions for orthographic transcription and VAS?
2. What is the strength of the relationship between percent correct scores from orthographic transcription and scale values from VAS?
3. Are there differences in intralistener and interlistener reliability between orthographic transcription and VAS?

Methods

Speakers

The 78 speakers and the speech samples reported in studies by Tjaden et al. (2014) and Kuo et al. (2014) were used in the current study. The pool of 78 speakers consisted of healthy

control speakers, speakers with PD, and speakers with MS. The 32 control speakers included 10 men (M= 57 years, range= 25-70 years) and 22 women (M= 57 years, range= 27-77 years). Healthy control speakers reported the absence of neurological disease. The 16 speakers with PD included eight men (M= 67 years, range= 55-78 years) and eight women (M= 69 years, range= 48-78 years). All of the speakers with PD had a medical diagnosis of idiopathic PD. The 30 speakers with MS included 10 men (M= 51 years, range= 29-60 years) and 20 women (M= 50 years, range= 27-66 years). All of the speakers with MS had a medical diagnosis of MS.

The control speakers were recruited through posted flyers and advertisements and the speakers with PD and MS were recruited through patient support groups and newsletters for individuals with PD or MS. All participants were required to be native speakers of standard American English, to have received a high school diploma or equivalent, and to have visual acuity or corrected acuity adequate for reading printed materials. The use of hearing aids was an exclusionary criterion. An audiologist at the University at Buffalo Speech-Language and Hearing Clinic obtained pure tone thresholds for each participant. This was done to provide a picture of overall auditory status. No speakers were excluded on the basis of pure tone thresholds.

Speakers with PD and MS were taking a variety of symptomatic medications at the time of data collection, but none had undergone neurosurgical treatment for their disease. Participants with PD ranged from two to 32 years post diagnosis (M= 9 years, SD= 7.8 years). Two of the female speakers with PD had completed LSVT™ more than two years prior to the recordings, and two had completed the treatment approximately six months prior to recordings. One of these latter two females with PD was enrolled in twice-monthly LSVT™ refresher sessions. Participants with MS ranged from two to 47 years post diagnosis (M= 14 years, SD= 11 years). Five of the speakers with MS had a primary progressive disease course, 18 had a relapsing

remitting disease course, and seven had a secondary progressive disease course. None of the speakers with MS had received LSVT™ or other treatments targeting loudness. At the time of data collection, six of the speakers with MS had received therapy for dysarthria within the past five years, and one had received treatment a year prior. All of the speakers scored a minimum of 26/30 on the Standardized Mini-Mental State Examination (Molloy, 1999), except one male with MS who scored 25/30. These scores indicated that the speakers had normal cognition (Molloy, Standish, & Lewis, 2005). Speakers were paid a modest participation fee.

Clinical metrics of single word intelligibility, sentence intelligibility, and scaled estimates of speech severity for the Grandfather Passage (Duffy, 2013) were the topic of a paper by Sussman and Tjaden (2012) and are summarized below for the purpose of describing the participants' speech. Stimuli were pooled across the 78 speakers and 42 naive listeners were blinded to the speakers' neurological diagnoses and identities. Stimuli were presented in quiet through headphones at the same sound pressure level at which they were naturally produced by the speakers. Single-word intelligibility was obtained using the single word test of Kent, Weismer, Kent, and Rosenbek (1989). Single-word intelligibility for control speakers was 97% (SD= .01), for speakers with MS was 96% (SD= .03), and for speakers with PD was 95% (SD= .03). Sentence intelligibility scores were obtained using the Sentence Intelligibility Test (SIT; Yorkston, Beukelman, & Tice, 1996). The mean sentence intelligibility for control speakers was 94% (SD= 2.7), for speakers with MS was 93% (SD= 4.5), and for speakers with PD was 85% (SD= 10). Perceptual judgments of speech severity for the Grandfather Passage were also reported in Sussman and Tjaden (2012). Speech severity is an operationally defined perceptual construct that aims to tap into speech naturalness and prosodic adequacy (Feenaughty, Tjaden, & Sussman, 2014; Kuo et al., 2014; Sussman & Tjaden, 2012). Listeners used a computerized VAS

to judge speech severity, with scale endpoints of 0 ("no impairment") and 1.0 ("severe impairment"). The mean scale values for 10 inexperienced listeners were reported in Sussman and Tjaden (2012) which reflect scaled estimates of speech severity for the Grandfather Passage. The scaled speech severity for control speakers was 0.18 (SD= .08), for speakers with MS was 0.44 (SD= .25), and for speakers with PD was 0.46 (SD= .21). Additionally, anecdotal perceptual observations were reported by Sussman and Tjaden (2012). The authors noted that many of the speakers with MS had reduced segmental precision and some prosodic (e.g., slow speech rate, excess stress) and voice deficits (e.g., harshness, hoarseness). It was also noted that many of the speakers with PD had reduced segmental precision and a breathy, monotonous voice (Sussman & Tjaden, 2012). Many of the speakers with MS and PD had relatively high intelligibility (e.g., high SIT scores: MS= 93%, PD= 85%), but a noticeable speech impairment, as reflected in the higher scaled speech severity scores relative to control speakers. The combination of the clinical metrics of intelligibility, scaled severity for the Grandfather Passage, and anecdotal perceptual judgments demonstrated that many of the speakers presented with mild dysarthria (Yorkston et al., 2010).

Experimental Speech Stimuli and Speech Tasks

Speakers read 25 Harvard Psychoacoustic Sentences (IEEE, 1969) in habitual, clear, fast, loud, and slow conditions. For the purposes of the current paper, the term non-habitual conditions will be used to refer to the clear, fast, loud, and slow conditions. Harvard sentences were semantically and syntactically normal and included declaratives and imperatives. Each sentence contained between seven and nine words, and five key words (e.g., nouns, verbs, adjectives, and adverbs). Audio recording of speakers took place in a quiet or sound-treated room. The acoustic signal was transduced using an AKG C410 head mounted microphone

positioned 10 cm and 45 to 50 degrees from the left oral angle. The signal was preamplified, low pass-filtered at 9.8 kHz, and digitized directly to computer hard disk at a sampling rate of 22 kHz using TF32 (Milenkovic, 2005). A calibration tone was also recorded to allow for offline measure of vocal intensity (see Lam, Tjaden, & Wilding, 2012).

A unique random ordering of the 25 Harvard sentences was recorded for each speaker and condition. Non-habitual conditions were elicited using a magnitude production paradigm and all speakers were given the same standard instructions that were read from a printed script. For the habitual condition, speakers were asked to produce the sentences using their normal, comfortable speech. For the clear condition, speakers were given instructions similar to those used in other clear speech studies to ensure that they would exaggerate articulation, increase vocal intensity, and reduce rate (Smiljanić & Bradlow, 2009). More specifically, speakers were instructed to say each sentence twice as clearly as their typical speech by pretending they were speaking to someone in a noisy environment or to someone with a hearing loss. Speakers were told to exaggerate the movements of their mouth and that their speech may be slower and louder than usual. For the fast condition, speakers were instructed to use a rate twice as fast as their typical rate. For the loud condition, speakers were instructed to produce sentences using speech twice as loud as their regular speaking voice. For the slow condition, speakers were instructed to produce the sentences at a rate half as fast as their regular rate and to stretch out words rather than solely inserting pauses. Also for the slow condition, speakers were instructed to produce each sentence on a single breath, as speakers have been instructed in other studies (e.g., McHenry, 2003).

All speakers produced the sentences in the habitual condition first, followed by randomly assigned orderings of the remaining conditions. Engaging the speakers in conversation for a few

minutes between conditions addressed potential carry-over effects. Prior to recording, speakers were familiarized with the stimuli and, for non-habitual conditions, were given a brief period to practice. Using a sentence from the Sentence Intelligibility Test (SIT; Yorkston et al., 2007), an investigator modeled the desired speaking condition and told the participants that their clear (or fast, loud, or slow) speech might differ from that of the investigator. After this model, the speakers practiced using the specified speech style using a different sentence and were given general feedback by the investigator. Speakers with PD were recorded one hour prior to taking PD medications. By recording speakers with PD an hour prior to taking medications, the presence of dysarthria was maximized, as any potential benefits of medication on speech would likely be at their lowest. The recording timing of speakers with MS was not a concern, as there are no documented effects of MS medication on speech.

For the purposes of the current study, as well as the perceptual studies of Tjaden et al. (2014) and Kuo et al. (2014), a subset of sentences was selected from the larger body of 25 Harvard Sentences (IEEE, 1969) to allow listeners to complete the perceptual task in one listening session. For each speaker, a random sample of the same 10 sentences produced in each of the five conditions was of interest. For example, for the first male speaker with PD (PDM01), sentences 1, 13, 14, 16, 17, 18, 19, 22, 23, and 25 in each of the conditions (habitual, clear, fast, loud, and slow) were selected, but for the first female speaker with MS (MSF01), sentences 1, 3, 5, 6, 19, 20, 22, 23, 24, and 25 were selected. For the remainder of this paper, any reference to the Harvard sentences refers to this subset of 10 sentences per speaker per condition.

Acoustic measures of SPL and articulatory rate were obtained using TF32 to verify the presence of production differences between the speaking conditions. These measures and the procedures for obtaining them are reported in Tjaden et al. (2014) and Kuo et al. (2014).

Sentences were first segmented into runs, defined as a stretch of speech bounded by silent periods or pauses between words of at least 200 ms (Turner & Weismer, 1993). Conventional acoustic criteria were used to identify run onsets and offsets. Articulatory rate was computed by dividing the number of syllables produced by run duration in milliseconds and multiplying by 1,000. Mean articulatory rate for each speaker and condition was calculated by averaging the articulatory rate for all runs. Mean SPL was also calculated for each speech run. Root-Mean Squared (RMS) traces were generated in TF32 and voltages were converted to dB SPL in Excel with reference to each speaker's calibration tone. The loud condition for one female with MS was excluded from all analyses due to technical difficulties during recording.

Descriptive statistics for the acoustic measures were reported in Tjaden et al. (2014) and Kuo et al. (2014). The descriptive statistics in Table 1 indicate that all speaker groups increased mean SPL for the loud, clear, and fast conditions relative to the habitual condition. The average magnitude of the increase across groups was seven to 10 dB for the loud condition, three to four dB for the clear condition, and two to three dB for the fast condition.

Table 1. Mean sound pressure level (dB SPL) with standard deviations in parentheses as a function of group and condition.

	Habitual	**Clear**	**Fast**	**Loud**	**Slow**
Control	73 (2.7)	77 (4.5)	76 (4.1)	83 (4.0)	73 (4.0)
MS	72 (3.0)	75 (4.4)	75 (5.3)	80 (3.6)	72 (4.7)
PD	72 (3.2)	75 (4.0)	74 (4.7)	79 (4.0)	72 (4.6)

The descriptive statistics in Table 2 indicate a reduced rate for the slow and clear conditions relative to the habitual condition. Relative to the habitual condition, the average magnitude of the rate reduction across groups was 29% to 49% for the slow condition and 19% to 37% for the clear condition. As well, the descriptive statistics in Table 2 indicate an increased

rate for the fast condition relative to the habitual condition. The average magnitude of rate increase across groups was 16% to 28% for the fast condition relative to the habitual condition.

Table 2. Mean articulation rate (syllables/second) with the standard deviations in parentheses as a function of group and condition.

	Habitual	**Clear**	**Fast**	**Loud**	**Slow**
Control	3.7 (.44)	2.3 (.32)	5.13 (.55)	3.2 (.46)	1.9 (.48)
MS	3.6 (.60)	2.7 (.63)	4.76 (.99)	3.3 (.69)	2.4 (.60)
PD	4.1 (.58)	3.3 (.75)	4.91 (.86)	4.0 (.71)	2.9 (.75)

Overall, all groups increased mean SPL in the loud condition relative to the habitual, clear, slow, and fast conditions. All groups reduced articulatory rate in the slow condition relative to the habitual, clear, fast, and loud conditions. All groups increased articulatory rate in the fast condition relative to the habitual, clear, loud, and slow conditions. The MS and control groups also slowed articulation rate in the loud condition relative to the habitual condition, but the PD group did not. Lastly, for all groups, the clear condition was characterized by an increased mean SPL and reduced mean articulatory rate relative to habitual, but with a lower magnitude of adjustments than for the loud and slow conditions. Therefore, the clear, fast, loud, and slow conditions were produced distinctly from each other and from the habitual condition.

Listeners

As stated previously, the purpose of the current study was to compare VAS and orthographic transcription. Therefore, it was important that the transcription data be collected using the same methods that were used to collect the VAS data. Fifty listeners, therefore, judged intelligibility in the current study, as in Tjaden et al. (2014) and Kuo et al. (2014). The inclusionary criteria for listeners were similar to those used in Tjaden et al. (2014) and in Kuo et al. (2014). Listeners ranged in age from 18 to 30 years and were required to pass a hearing

screening at 20 dB HL for 250, 500, 1000, 2000, 4000, and 8000 Hertz (Hz) bilaterally. Listeners were native speakers of standard American English and had at least a high school diploma or equivalent. Listeners were also required to report no history of speech, language, or hearing problems, and have limited to no experience with disordered speech. Listeners were recruited using flyers posted at the University at Buffalo and were paid a modest participation fee.

Stimuli Preparation and Perceptual Task

While speech may be relatively intelligible in ideal settings for speakers with mild dysarthria, simulating a more challenging setting is useful to reduce the likelihood of ceiling effects. One way to create a more challenging listening environment for listeners is to introduce background noise (Bunton, 2006; Smiljanić & Bradlow, 2009; Yorkston et al., 2007). Therefore, the Harvard sentences were mixed with multi-talker babble to induce a more challenging listening environment and to reduce the likelihood of ceiling effects (Bunton, 2006). Listeners have likely experienced environments that sound similar to the multi-talker babble background noise (e.g., restaurant, shopping mall, party, etc.). This makes multi-talker babble a more ecologically valid background noise as compared other types of noise, such as white or pink noise.

Sentences were first equated for peak vowel amplitude using Goldwave Version 5 (Goldwave Inc., 2010) to minimize differences in audibility among sentences. Stimuli then were mixed with 20-talker babble using Goldwave Version 5, and a signal to noise ratio (SNR) of -3 dB was then applied to each sentence. This SNR was identified with pilot testing to not produce ceiling or floor effects and has also been used in studies that investigated the intelligibility of clear speech (Ferguson & Kewley-Port, 2002; Maniwa, Jongman, & Wade, 2008). Using procedures similar to Tjaden et al. (2014) and Kuo et al. (2014), stimuli were presented to

individual listeners at 75 dB SPL via headphones (SONY, MDR V300) in a double-walled audiometric booth. The task took between two and three hours with breaks, and was self-paced.

In the previous studies using VAS to quantify sentence intelligibility (Kuo et al., 2014; Tjaden et al., 2014), sentences for all speakers and conditions were first pooled and then divided into 10 lists. Sentence lists contained one sentence produced by each of the 78 talkers in each condition. Furthermore, sentence lists included similar numbers (N= 15 or 16) of each of the 25 Harvard sentences in all conditions. Listeners in the current study judged intelligibility for these 10 sentence lists. Five listeners were assigned to judge each list. Each listener also judged a random selection of 10% of sentences twice in order to determine intrajudge reliability. To familiarize listeners with the repetitive stimuli, listeners first heard all Harvard sentences (IEEE, 1969) produced by one healthy male and female speaker who were not part of the study. Then, listeners practiced using the computer interface and were exposed to sentences mixed with babble by transcribing six sentences produced by speakers. Both the sentences and the speakers used for practice were not part of the current study.

After hearing a sentence once, each of the 50 listeners were instructed to type exactly what they heard. Listeners had no knowledge of the speakers' neurological diagnoses or the speaking conditions. Following entering their response, listeners were given an opportunity to edit and change their response if needed. The software program saved the listeners' entered responses for later scoring.

After listeners completed the task, a key word scoring paradigm was used to score each response. Hustad (2006a) reported that content words are of more importance than function words, as they carry the most information in the sentence. According Hustad (2006a), when differences between scoring paradigms for transcription intelligibility were examined, a key

word scoring paradigm yielded a conservative estimate of intelligibility. Results showed that the precise nature of the scoring paradigms for transcription could influence overall scores. However, the differences were small and Hustad (2006a) deemed them to not be clinically meaningful. A number of previous studies have also used key word scoring paradigms to score listener transcriptions (Kain, Amano-Kusumoto, & Hosom, 2008; Miller, Schlauch, & Watson, 2010; Tjaden, Kain, & Lam, 2014). For these reasons, a key word scoring paradigm was chosen. This paradigm involved scoring only the key informational words in each Harvard sentence for a correct or incorrect match to the target. Each of the Harvard sentences contains five key words, which include the nouns, verbs, adjectives, and adverbs in the sentences. Following a similar approach to Cannito and colleagues (2012), a liberal scoring approach was taken. Homophones (e.g., gel for jell) and phonetically correct misspellings (e.g., doon for dune) were scored as correct. Additionally, the scoring paradigm disregarded word order (e.g., 'wooden square crate' for 'square wooden crate'). Other typing errors (e.g., both for booth) were scored as incorrect, as were incorrect plurals (e.g., cherry for cherries) and tense markers (e.g., dries for dried). An exception to this rule involved obvious spelling errors that did not create other words (e.g., arbupt for abrupt), which were scored as a correct match.

As previously stated, the listeners were divided among the 10 lists of stimuli and five listeners judged each of the 10 lists. Therefore, five listeners judged each individual sentence. The five listeners' responses were pooled and the number of key words correctly transcribed was tallied. This number was then divided by the total number of key words. This provided a mean percent correct score across five listeners for each individual sentence. In addition, percent correct scores for each sentence were than averaged across a given condition for every speaker. Percent correct intelligibility scores derived from orthographic transcriptions were compared to

the intelligibility judgments derived from the VAS task reported in Tjaden et al. (2014) and Kuo et al. (2014).

By way of review, Tjaden et al. (2014) and Kuo et al. (2014) used a computerized VAS task and listeners were asked to judge how well a sentence was understood. Listeners were presented with a continuous 150 mm vertical oriented scale on a computer monitor, which contained no tick marks. The endpoints of the scale were labeled with "Understand everything" and "Cannot understand anything". Listeners were instructed to use a mouse and click on the line to indicate how well a given speaker's sentence was understood (Kuo et al., 2014; Tjaden et al., 2014). Following completion of the experiment, software converted responses to numerical values ranging from 0 (i.e., *Understand everything*) to 1.0 (i.e., *Cannot understand anything*). As in the current study, five listeners judged each individual sentence. The scaled values were averaged across the five listeners who heard a specific sentence. This produced an average score for an individual sentence across five listeners.

Listener reliability. While the primary purpose of the current study was to directly compare sentence intelligibility results for transcription and VAS, a comparison of listener reliability for the two types of tasks was also of interest. It is a common expectation that listener reliability be reported in studies that use scaling tasks (Sussman & Tjaden, 2012; Tjaden et al., 2014; Tjaden & Wilding, 2010; Yunusova et al., 2005). However, listener reliability is not routinely reported in studies that use transcription (Hustad, 2006a; 2006b; McHenry, 2011), though it has been included in a few studies (Bunton et al., 2001; Tjaden, Kain, & Lam, 2014; Tjaden & Wilding, 2010).

In the current study, intra and inter reliability of listeners' responses were obtained. Methods for determining intralistener reliability were incorporated into the listening task. Forty

stimuli were repeated in each of the 10 lists (e.g., approximately 10% of the stimuli in each list). Listener responses from the 40 repeated stimuli were compared to listener responses from the same 40 stimuli presented earlier in the task. In order to determine interlistener reliability, the responses from listeners who heard the same set of stimuli (i.e., one of the 10 lists) were compared. The specific metrics for quantifying reliability are discussed in the data analysis section.

Scoring reliability. The intrascorer and interscorer reliability of scoring decisions is often reported in studies that use transcription (Hustad, 2006a; 2006b; McHenry, 2011). The intrascorer and interscorer reliability mentioned here refers to the consistency or reliability of scoring the transcription responses. Reliability of the scoring of transcriptions was based on a model used by Hustad (2008). Intrascorer reliability was determined by having the original scorer rescore five randomly selected listeners' transcriptions (or 10% of the transcription responses). Pearson product correlation coefficients for the first and second scoring of listener responses ranged from 0.98 to 1.00, with a mean of 0.99 (SD= .01). This is comparable to intrascorer reliability found by Hustad (2006a) and McHenry (2011) and represents strong intrascorer reliability. Interscorer reliability was determined by having a second scorer, who was not involved in the initial scoring rescore 10% of the listener responses. This second scorer rescored responses from five randomly selected listeners. Unit-by-unit agreement was obtained by dividing the number of agreements by the number of agreements plus disagreements. Pearson product-moment correlation coefficients for the first and second scoring of listener responses ranged from 0.92 to 1.00, with a mean of 0.98 (SD= .03). Again, this is comparable to interscorer reliability found by Hustad (2006a; 2006b) and McHenry (2011) and indicates a high level of reliability in the scoring of transcribed responses.

Data Analysis

Dependent measures were characterized using both descriptive (i.e., mean [M], standard deviation [SD]) and parametric statistics. For ease of understanding, the description of these analyses will be presented in three separate sections.

Research question 1: Pattern of findings for intelligibility. Initially, descriptive statistics (i.e., mean, standard deviation) were computed for the percent correct scores. The pattern of results for descriptive statistics was examined for both the transcription data and the VAS data (Tjaden et al., 2014; Kuo et al., 2014). This examination of descriptive statistics serves as a qualitative comparison of overall means for transcription and scaling.

Transcription data were also analyzed using parametric statistics. Initially, overall percent correct scores per speaker and condition (i.e., PDM01 speaker in the habitual condition) were calculated. Similar to Tjaden et al. (2014) and Kuo et al. (2014), a repeated measures Analysis of Variance (ANOVA) was performed. Using SAS Version 9.1.3 statistical software, a multivariate linear model was fit to each dependent measure. The percent correct scores were fit as a function of group (control, MS, PD), condition (habitual, clear, fast, loud and, slow), and a Group x Condition interaction. A covariate representing speaker sex was included in each model to account for different proportions of male and female speakers among groups. Follow-up contrasts were made in conjunction with a Bonferroni correction for multiple comparisons. All tests were evaluated at a 0.05 nominal significance level. The results of this ANOVA cannot be directly compared to the Tjaden et al. (2014) study or to the Kuo et al. (2014) study because these previous studies included subsets of the five conditions. Statistical outcomes were qualitatively compared to those reported in Tjaden et al. (2014) and Kuo et al. (2014).

Research question 2: Strength of the relationship between transcription and VAS.

Pearson product-moment correlation coefficients were obtained to examine the strength of the relationship between the transcription task scores and the VAS task scores. Two main correlation analyses were completed. Initially, a correlation analysis was completed for each speaker (total of 78 separate correlations). For each speaker, data were pooled across conditions, such that each speaker had 50 data points from the transcription task and 50 data points from the VAS task. Second, correlations were computed for each condition and group to quantify the strength of the relationship between transcription and VAS. Given the five conditions (habitual, clear, fast, loud, and slow) and three groups (control, PD, and MS), a total of 15 correlations were computed. Correlations were evaluated at a 0.05 nominal significance level.

Research question 3: Listener reliability comparison. For the transcription data, the number of exact word matches that were similar across the two presentations was compared. Listener reliability was calculated by summing the number of key words that were correctly transcribed in both presentations of the stimuli and dividing by the total number of key words. For a given sentence production, a listener may have transcribed three key words correct in the first presentation of the stimuli and three key words correct in the second presentation of the stimuli. However, out of the three key words transcribed correctly in both presentations, it was possible for only one of these exact words to be transcribed correctly in both presentations. For example, if a listener transcribed, “Glue the sheet to the background” in the first presentation, three key words were transcribed correctly (e.g., glue, sheet, background). If the listener then transcribed, “Add the sheet to the dark blue page” for the second presentation, again three key words were transcribed correctly (e.g., sheet, dark, and blue), however only one key word (e.g.,

sheet) was transcribed correctly in both presentations. Therefore, the reliability score for this item would be 0.2 (e.g., 1 key word transcribed correctly/5 possible key words= 0.2).

Following Neel (2009) and Tjaden et al. (2014), interlistener reliability was assessed using the Intraclass Correlation Coefficient (ICC). ICCs were calculated separately for all sentence lists, as the listeners assigned to judge each of these lists heard different sentences. ICCs involved a two-way mixed-effects model to determine the overall consistency of ratings among listeners. Aggregate listener performance was of interest, and therefore, average ICC metrics were considered the primary measure of agreement among listeners. ICCs for transcription were summarized using descriptives (e.g., mean, SD, and range), and compared to the ICC scores for the VAS data (Tjaden et al., 2014; Kuo et al., 2014) using a 0.05 nominal significance level.

Results

In the following sections, descriptive statistics are reviewed followed by parametric results for each research question. Van Nuffelen et al. (2009; 2010) determined that intelligibility changes of 8% or more are clinically meaningful. In a more challenging perceptual environment, such as the one in the current study with the addition of background noise, intelligibility changes of 5% or more are likely to be meaningful (Kuo et al., 2014; Tjaden et al., 2014). This means that any differences in intelligibility that are less than 5% are likely not clinically meaningful and can be considered comparable.

Research Question 1: Pattern of Findings for Intelligibility

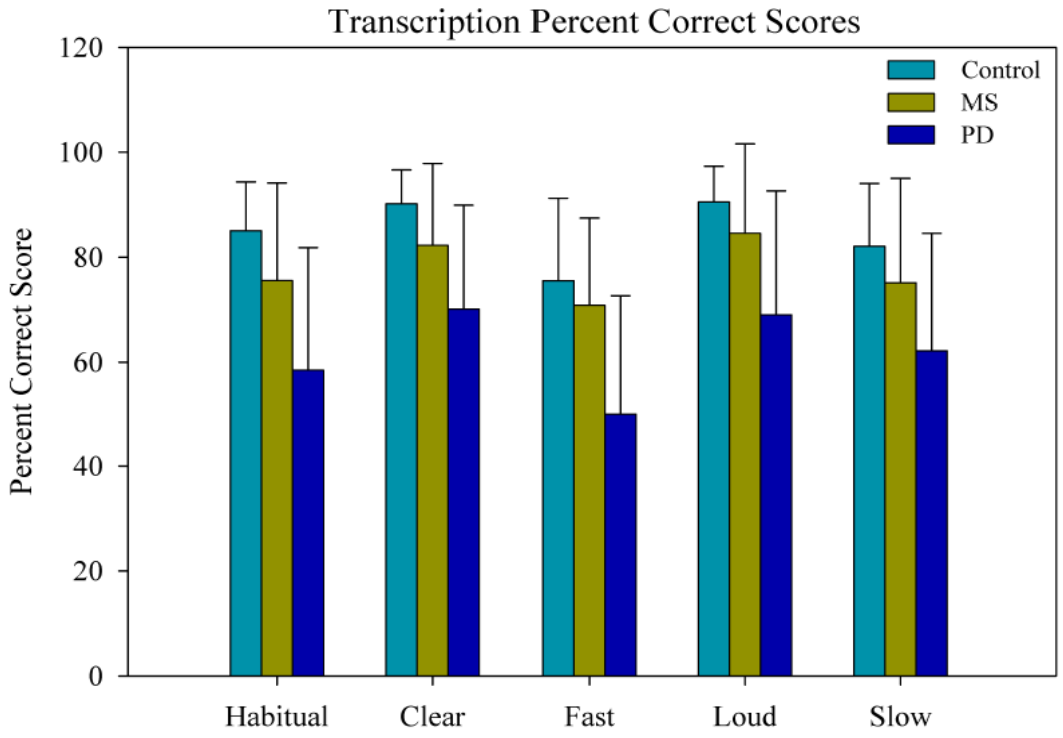
Transcription intelligibility descriptive statistics for groups. Results for transcription intelligibility, per group and condition are reported in Table 3. The transcription results, with standard deviation bars, as a function of group (turquoise- control, green- MS, dark blue- PD)

and condition are also shown in Figure 1. Figure 1 and Table 3 show the same information presented in a slightly different way in order to facilitate better understanding of the descriptive statistics.

Table 3. Transcription percent correct score means with standard deviations in parentheses as a function of group and condition.

	Habitual	Clear	Fast	Loud	Slow
Control	85.08 (9.26)	90.19 (6.43)	75.53 (15.72)	90.53 (6.81)	82.09 (11.99)
MS	75.55 (18.62)	82.29 (15.58)	70.86 (16.60)	84.55 (17.04)	75.16 (19.86)
PD	58.50 (23.33)	70.13 (19.79)	49.90 (22.77)	69.05 (23.57)	62.20 (22.34)

Figure 1. Transcription Percent Correct Scores



In the following paragraphs, the differences in intelligibility change between groups will be briefly examined. Second, the pattern of intelligibility results across conditions within each group was considered. Lastly, the pattern of intelligibility results across conditions for all groups

is summarized. First, Table 1 and Figure 3 indicate that intelligibility in each condition was always highest for the control group, followed by the MS group, and then the PD group. Percent correct scores for the control group was between five and 10 percent higher than the MS group. Subsequently, percent correct scores for the MS group was between 12 and 21 percent higher than the PD group.

On average, transcription intelligibility for the control group was best (i.e., highest percent correct score) in the loud condition, followed by the clear, habitual, slow, and fast conditions. Using the guideline that changes in intelligibility of at least 5% are clinically meaningful, for the control group, the clear and loud conditions did not differ, but both increased intelligibility relative to the habitual condition by at least 5%. Additionally, the slow and habitual conditions did not differ. The fast condition decreased intelligibility relative to the habitual condition by at least 5%.

The mean percent correct scores in Figure 1 and Table 3 also suggest that transcription intelligibility for the MS group was best in the loud condition, followed by the clear, habitual, slow, and fast conditions. Again, the clear and loud conditions were essentially the same, but increased intelligibility relative to the habitual condition by at least 5%. The habitual and slow conditions did not differ. The fast condition decreased intelligibility relative to the habitual condition by at least 5%.

Lastly, the mean percent correct scores in Figure 1 and Table 3 suggest that transcription intelligibility for the PD group was best in the clear condition, followed by the loud, slow, habitual, and fast conditions. The clear and loud conditions were essentially the same, but increased intelligibility relative to the habitual condition by at least 5%. The habitual and slow

conditions did not differ. The fast condition decreased intelligibility relative to the habitual condition by 5% or more.

To summarize, for all groups, the clear and loud conditions were essentially the same and increased intelligibility relative to the habitual condition by at least 5%. The habitual and slow conditions did not differ for any of the groups. The fast condition decreased intelligibility relative to the habitual condition for all of the groups. The increases and decreases in intelligibility relative to the habitual condition for the various conditions were of greatest magnitude for the PD group at approximately 11%, compared to approximately 7% for the MS and control groups.

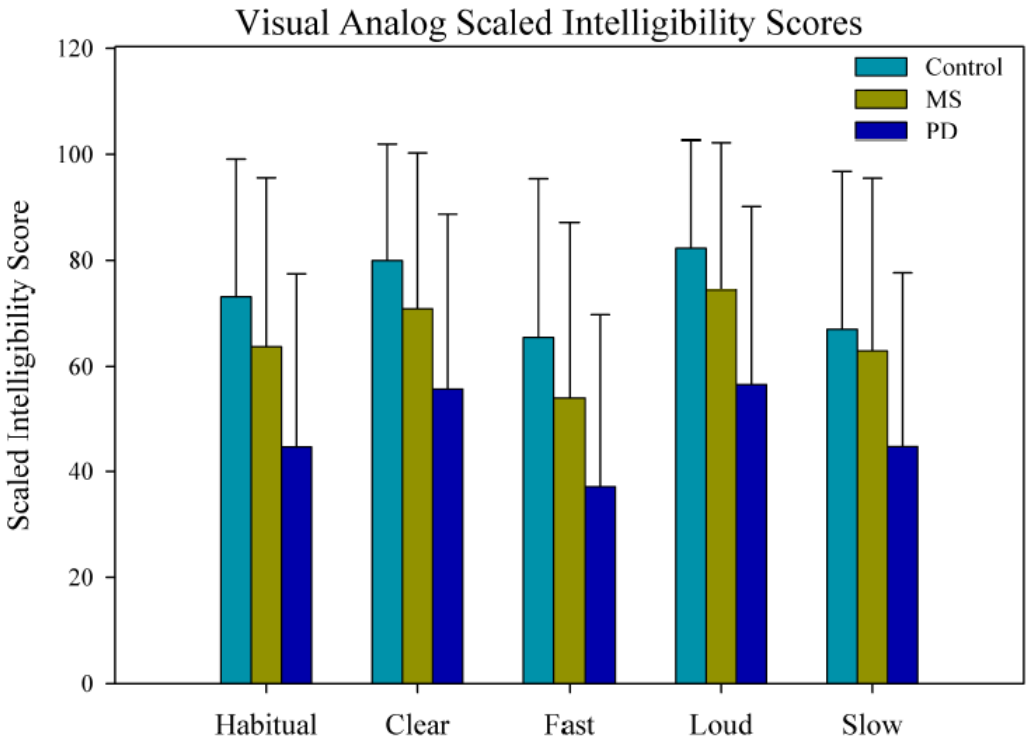
Transcription intelligibility descriptive statistics for individual speakers. Percent correct scores for individual speakers in each condition can be viewed in Appendix A. To obtain these scores, all listener responses were pooled and the correctly transcribed key words were summed for each listener and divided by the total number of key words. The pattern of results for each group as a whole shown in Table 3 and Figure 1 is considered below, followed by an examination of individual speakers trends for each of the three groups.

When looking at the overall percent correct averages per speaker in each condition (seen in Appendix A), a few observations can be made. First, group averages in Table 3 indicated that for the control group, transcription intelligibility was best in the loud and/or clear conditions, followed by the habitual and/or slow conditions, followed by the fast condition. Of the 32 control speakers, either the loud or clear condition increased intelligibility by at least 5% relative to the habitual condition for 16 speakers, or 50% of all control speakers. The fast condition decreased intelligibility by at least 5% relative to the habitual condition for 18 control speakers, or 56% of all control speakers. Group averages reported in Table 3 further indicate that for the MS group,

transcription intelligibility was best in the loud and/or clear conditions, followed by the habitual and/or slow conditions, followed by the fast condition. The loud or clear condition increased transcription intelligibility by at least 5% relative to the habitual condition for 22 of the 30 speakers with MS, or 73% of all speakers with MS. The fast condition decreased intelligibility by at least 5% relative to the habitual condition for 26 speakers with MS, or 87% of all speakers with MS. Lastly, group averages reported in Table 3 indicate that for the PD group, transcription intelligibility was best in the loud and/or clear conditions, followed by the habitual and/or slow conditions, and then the fast condition. The loud or clear condition increased transcription intelligibility by at least 5% relative to the habitual condition for 14 of the 16 speakers with PD, or 88% of all speakers with PD. The fast condition decreased intelligibility by at least 5% relative to the habitual condition for 12 out of the 16 PD speakers, or 75% of all speakers with PD. These results demonstrate the variability of individual speakers within groups in regard to how each condition affected transcription intelligibility. These results further indicated more consistency in the disordered speaker groups (MS and PD) than the control group.

Comparison of descriptive statistics. Results for the VAS intelligibility obtained by Tjaden et al. (2014) and Kuo et al. (2014) are shown in Figure 2. Standard deviations are shown via SD bars. The VAS used by listeners ranged from 0 (understand everything) to 1.0 (cannot understand anything). To allow these scaled values to be more easily compared to the percent correct scores, the scale has been reversed and was multiplied by 100, such that values closer to 100 represent greater intelligibility (e.g., all scaled scores have been subtracted from 1.0 and multiplied by 100). This reversed scale is reflected in the remainder of this paper, so that scaled scores closer to 100 and percent correct scores closer to 100 both represent greater intelligibility.

Figure 2. Visual Analog Scaled Intelligibility Scores



Using the guideline that differences in intelligibility of at least 5% are meaningful, for scaled judgments of the control group, the clear and loud conditions did not differ from each other, but both increased intelligibility relative to the habitual condition by at least 5%. The slow and fast conditions did not differ, but decreased intelligibility relative to the habitual condition by at least 5%. This pattern is similar to the one found in the current study, for transcription shown in Figure 1 and Table 3. For the control group, the clear and loud conditions increased transcription intelligibility and the fast condition decreased intelligibility relative to the habitual condition.

As shown in Figure 2, for the MS group, the clear and loud conditions did not differ, but both increased intelligibility relative to the habitual condition by at least 5%. The habitual and slow conditions did not differ. The fast condition decreased intelligibility relative to the habitual

condition by at least 5%. This pattern is similar to the one found in the current study for transcription, shown in Figure 1 and Table 3. For the MS group, the clear and loud conditions increased transcription intelligibility and the fast condition deceased intelligibility relative to the habitual condition.

As shown in Figure 2 for the PD group, the clear and loud conditions did not differ, but increased intelligibility relative to the habitual condition by at least 5%. The habitual and slow conditions did not differ. The fast condition decreased intelligibility relative to the habitual condition by at least 5%. This pattern is similar to the one found in the current study for transcription, shown in Figure 1 and Table 3. For the PD group, the clear and loud conditions increased transcription intelligibility, but the fast condition decreased intelligibility relative to the habitual condition.

A final note from the visual comparison of Figures 1 and 2 is that for both intelligibility tasks, in each condition the pattern of group results is the same. The control group was the most intelligible in each condition, followed by the MS group, and the PD group. In addition, overall percent correct scores from the transcription task were of greater magnitude than the scaled scores from the VAS task.

Comparison of parametric statistics. Statistical analysis of percent correct scores from transcription indicated a significant effect of group, $F(2, 70)= 9.78$, $p= <.001$. Mean and standard deviation scores are listed in Table 3 as a function of group and condition. Follow-up contrast tests indicated that the PD group had poorer intelligibility compared to both control ($p= <.001$) and MS groups ($p= .0064$). The MS-control contrast was not significant. There was also a main effect of condition $F(4, 70)= 36.25$, $p= < .001$. Transcription intelligibility for the clear and loud conditions was significantly better than the habitual condition ($p= <.001$), but the clear and loud

conditions did not differ. The fast condition decreased intelligibility relative to the habitual condition (p= <.001). All other comparisons were not significant. The Group x Condition interaction was not significant.

Post hoc follow-up contrasts indicated that for the control group, only the loud condition increased intelligibility relative to the habitual condition (p= .001). The clear and habitual conditions, as well as the habitual and slow conditions were not statistically different. The fast condition decreased intelligibility relative to the habitual condition (p= .001). For the MS group, transcription intelligibility for the clear and loud conditions was significantly better than the habitual condition (p= <.001), but the clear and loud conditions did not differ. In addition, the habitual and slow conditions did not differ. Lastly, for the MS group, there was no statistical difference between the habitual, slow, and fast conditions. For the PD group, transcription intelligibility was increased relative to the habitual condition in the clear and loud conditions (p= <.001), but the clear and loud conditions did not differ. The slow-habitual contrast and the fast-habitual contrast were not statistically significant for the PD group.

The parametric statistics obtained for the transcription data were compared to the parametric statistics obtained by Tjaden et al. (2014) and Kuo et al. (2014). Tjaden et al. (2014) found similar group differences in that the PD group had poorer intelligibility compared with both control and MS groups, but that the MS-control contrast did not reach significance. For the control group, scaled intelligibility for the clear and loud conditions was significantly better than the habitual condition, but the clear and loud conditions were not statistically different (Tjaden et al., 2014). This differs slightly from the results for transcription intelligibility in the current study, as only the loud condition increased intelligibility relative to the habitual condition. Scaled intelligibility for the control group was poorer in the slow condition relative to the habitual

condition (Tjaden et al., 2014), which again differs from transcription intelligibility, as the habitual and slow conditions were not statistically different. Both scaled intelligibility and transcription intelligibility were poorer in the fast condition relative to the habitual condition (Kuo et al., 2014).

For the MS group, scaled intelligibility for the clear and loud conditions was significantly better than the habitual condition, but the clear and loud conditions did not differ (Tjaden et al., 2014). Similarly, transcription intelligibility for the clear and loud conditions was significantly better than the habitual condition and the clear and loud conditions also did not differ. Both scaled and transcription intelligibility for the MS group was not statistically different in the slow and habitual conditions (Tjaden et al., 2014). Scaled intelligibility was poorer in the fast condition relative to the habitual condition (Kuo et al., 2014), but for transcription intelligibility, there was no statistical difference between the habitual and fast conditions.

For the PD group, scaled intelligibility for the clear and loud conditions was significantly better than the habitual condition, but the clear and loud conditions did not differ (Tjaden et al., 2014). Similarly, transcription intelligibility for the clear and loud conditions was significantly better than the habitual condition and the clear and loud conditions also did not differ. Both scaled and transcription intelligibility for the PD group was not statistically different in the slow and habitual conditions (Tjaden et al., 2014). Scaled intelligibility was poorer in the fast condition relative to the habitual condition (Kuo et al., 2014), but for transcription intelligibility, there was no statistical difference between the habitual and fast conditions.

Research Question 2: Strength of the Relationship between Transcription and VAS

Correlation analyses. A correlation analysis was completed to examine the strength of the relationship between the transcription task scores and the VAS task scores on a per speaker

basis. Seventy-eight correlations, one per speaker, were calculated. The correlations for each speaker can be seen in Appendix B. The correlations ranged from 0.28 to 0.86 with an average correlation of 0.62 (SD= .149). All correlations were significant ($p < 0.05$), with two exceptions. The two speakers with non-significant correlations were CSF18 (r= .355, p= .110) and MSM08 (r= .257, p= .072). When these two nonsignificant correlations were excluded, the mean correlation was 0.63 (SD= .142). The average correlation of 0.62 suggests a strong relationship between the transcription task scores and the VAS task scores (Cohen, 1988).

A second correlation analysis was completed to examine the strength of the relationship between the transcription task scores and the VAS task scores on a per condition and group basis. These correlations are presented in Table 4.

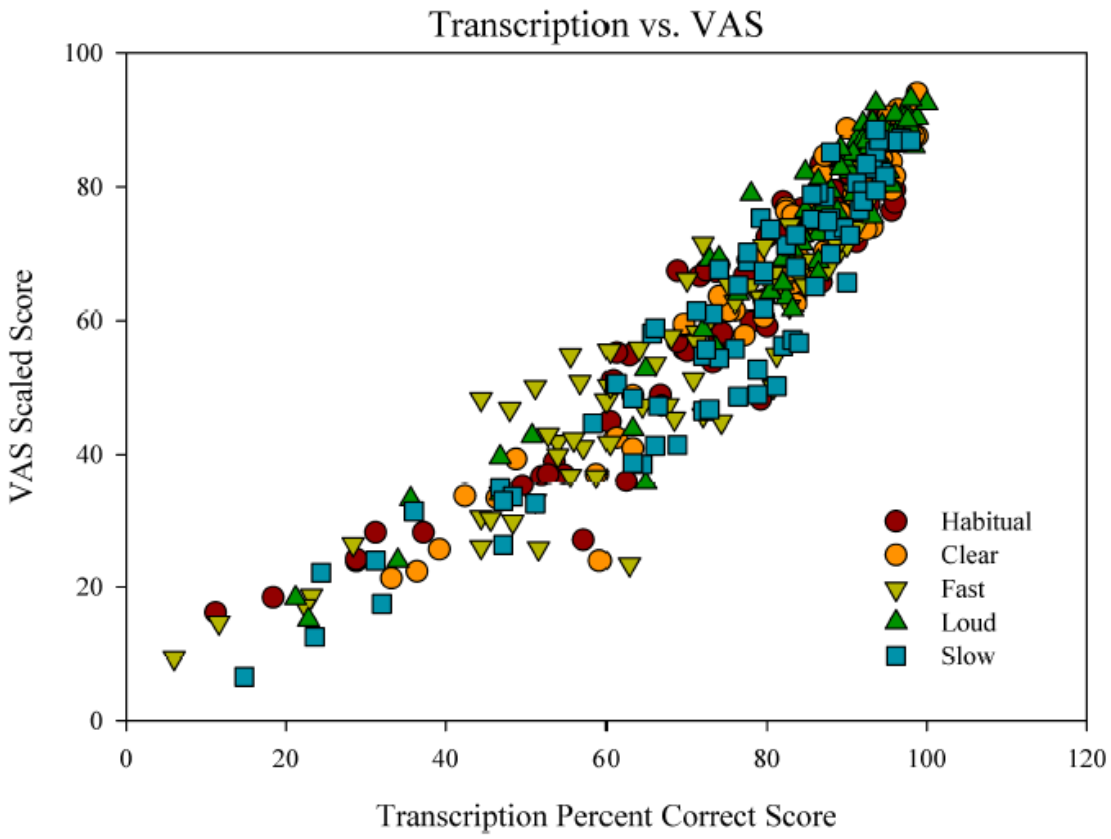
Table 4. Correlations of transcription percent correct scores and VAS scaled judgments as a function of group and condition.

	Habitual	**Clear**	**Fast**	**Loud**	**Slow**
Control	0.83	0.90	0.93	0.85	0.89
MS	0.95	0.95	0.89	0.95	0.97
PD	0.94	0.98	0.94	0.99	0.94

Figure 3 shows a scatter plot of the data with transcription percent correct scores on the x-axis and VAS intelligibility scores on the y-axis. Each point on the graph represents a single speaker in a given condition. Therefore, for a given speaker, five data points (one for each condition) is depicted in Figure 3. All correlations in Table 4 were significant ($p <.05$). On average, correlations were strongest for the PD group (mean= .96, range= .94-.99), followed by the MS group (mean= .94, range= .89-.97), and the control group (mean= .88, range= .83-.93). For each of the three groups, correlations were strongest for the clear condition (mean= .94, range= .89-.98), followed by the loud condition (mean= .93, range= .85-.95) and slow condition

(mean= .93, range= .89-.97). Correlations were weakest for the fast condition (mean= .92, range= .89-.94) and the habitual condition (mean= .91, range= .83-.95).

Figure 3. Transcription Percent Correct Scores vs. Visual Analog Scaled Intelligibility Scores by Condition



Research Question 3: Listener Reliability Comparison

Intralistener reliability. Intralistener reliability for each of the 50 listeners is available in Appendix C. The intralistener reliability analysis, performed to examine the proportion of exact matches in transcription responses, yielded Pearson product-moment correlations from 0.32 to 0.88 across the 50 listeners, with a mean of 0.66 (SD= .13). All correlations were significant (p= <.05). For comparison, in the VAS task, intralistener reliability correlation

coefficients ranged from 0.60 to 0.88 across the 50 listeners, with a mean of 0.71 (SD= .07) (Tjaden et al., 2014; Kuo et al., 2014).

Interlistener reliability. Interlistener reliability ICCs are available in Appendix D. Average ICCs for transcription intelligibility ranged from 0.78 to 0.86 (mean= .81, SD= .02), and single measure ICCs ranged from 0.33 to 0.54 (mean= .45, SD= .06). All ICC measures, both single and aggregate, were significant (p= <.05). Average ICCs across the 50 listeners for scaled intelligibility ranged from 0.85 to 0.91 (mean= .87, SD= .02), and single measure ICCS ranged from 0.54 to 0.68 (mean= .59, SD= .04) (Tjaden et al., 2014; Kuo et al., 2014).

Discussion

The overall goal of this project was to examine how percent correct intelligibility scores derived from orthographic transcription compare to intelligibility judgments derived from a scaling task (data from Tjaden et al., 2014 and Kuo et al., 2014). The pattern of intelligibility, as well as the strength of the relationship between the measures, was of interest. The difference in listener reliability for a VAS task and a transcription task was also examined. This discussion focuses on interpretation of the findings with reference to the three research questions. Clinical and theoretical implications of these findings are also considered.

Research Question 1: Pattern of Findings for Intelligibility

Results of this study pertaining to Research Question 1, endorsed two important ideas. The first is that the pattern of results for transcription intelligibility and VAS was the same with respect to group differences as well as differences among conditions. The second is that the raw scores were lower for VAS than for transcription.

Results from the current study mirror those from Hustad (2006b) who also found that subjective intelligibility scores in the form of percent estimates were lower than scores derived

from a transcription task for speakers with dysarthria. This suggests that scaled intelligibility scores are conservative, or underestimations of transcription intelligibility. The similar pattern, but difference in magnitude of intelligibility scores between these two measures has implications for clinicians and researchers. If it is true that transcription and a VAS task both measure the construct of intelligibility, then clinicians and researchers may be able to choose the less labor-intensive VAS task, with the knowledge that VAS may slightly underestimate intelligibility, but render a similar pattern of findings than if transcription was used. VAS may be used for circumstances when listener error patterns are not of interest (Liss, 2007), or when tracking changes in intelligibility related to disease progression or treatment. Additionally, since transcription and VAS yield raw intelligibility scores of different magnitudes, when the desire is to compare intelligibility findings either across time or across speakers, either transcription or VAS should be used exclusively. For example, results from a transcription task should be exclusively compared to results from the same transcription task, and results from a VAS task should be exclusively compared to results from the same VAS task.

Research Question 2: Strength of Relationship between Transcription and VAS

The first of two correlation analyses showed that, on average, the strength of the relationship between transcription percent correct scores and VAS estimates on a per speaker basis was strong (Cohen, 1988). The second correlation analysis, done on a per condition and group basis, again showed that the strength of the relationship between these two intelligibility metrics was strong (Cohen, 1988). Together, these correlation analyses demonstrate that the percent correct scores derived from transcription and judgments of intelligibility from VAS are highly correlated. Additionally, these metrics were correlated for the majority of individual speakers, although the strength of the relationship varied widely from speaker to speaker. These

correlations imply that although the magnitude of the scores may differ, the overall pattern of scores was broadly similar for the percent correct scores and the scaled judgments of intelligibility. This finding may indicate that transcription and a VAS task are tapping into the same perceptual phenomenon. Again, this finding has implications for both clinical and research purposes as discussed previously.

A study by Tjaden and Wilding (2011) used correlation analyses to examine the relationship between transcription and a DME scaling task. The authors found moderate correlations (p= .59) between scaled estimates of intelligibility from DME and percent correct scores from transcription for a reading task. However, correlations were weaker and did not reach significance when the two intelligibility metrics were utilized during different speech tasks (e.g., DME for reading vs. transcription for monologue). This result suggests the importance of using the same speech task when the desire is to compare intelligibility across time or from speaker to speaker. Intelligibility scores appear to become more distinct when the speech task itself changes, even within speakers.

Research Question 3: Listener Reliability Comparison

Miller (2013) is critical regarding the use of subjective scaling tasks to quantify intelligibility, especially when considering reliability. Miller (2013) stated that because listeners' "internal yardsticks" differ on subjective intelligibility metrics like VAS, the end result is poor inter-rater reliability (p. 603). Both interlistener and intralistener reliability (the consistency of responses between and within speakers, respectively) was slightly higher for the VAS data in Tjaden et al. (2014) and Kuo et al. (2014) than for the transcription data in the current study. The results from the current study contradict Miller's (2013) statement, as interlistener reliability was slightly better for VAS than for transcription. This finding is particularly interesting because

transcription is generally thought to have high listener reliability (Miller, 2013). There is previous dysarthria research in which transcription was utilized and reliability was not reported (e.g., Hustad, 2006a; 2006b; Liss, Spitzer, Caviness, & Adler, 2002; McHenry, 2011; Spitzer, Liss, & Mattys, 2007). However, there are a few studies that utilized transcription and did report reliability (Bunton et al., 2001; Tjaden, Kain, & Lam, 2014; Tjaden & Wilding, 2010). Tjaden, Kain, and Lam (2014) reported intrajudge correlation coefficients ranging from 0.57 to 0.99 (mean= .80, SD= .13) for a scaling task, and correlation coefficients ranging from 0.58 to 1.00 (mean= .80, SD= .13) for a transcription task. Tjaden, Kain, & Lam's (2014) results suggest that reliability for transcription and a scaling task are virtually identical; however, the current study found VAS reliability to be slightly higher than transcription reliability. While the difference found in reliability between transcription and VAS is not large, it may be a consideration when choosing a measure of intelligibility.

In the current study, reliability may have been poorer in the transcription task versus the VAS task for a variety of reasons. Of note was the length of the transcription task and the anecdotal evidence of listeners fatiguing over the course of the task. The transcription task took listeners between two and three hours, whereas the VAS task only took approximately 90 minutes (Tjaden et al., 2014). Listeners also relayed that concerted attention to the stimuli was necessary, especially due to the background noise. It is a possibility that listeners became fatigued and thus used varying levels of focus throughout the task, resulting in poorer overall reliability. Research by Liss (2007) and Choe, Liss, Azuma, & Mathy (2012) has been conducted to examine the source of this variation in listener response. Liss (2007) highlighted the challenges that listeners face specifically when encountering degraded speech signals, such as those produced by speakers with dysarthria. Choe et al. (2012) found that dysarthric speech tends

to magnify individual processing strategies and may cause listeners to 'try out' new strategies throughout a task, which may lead to decreased reliability both within and between listeners. In this way, reliability for transcription might mean something different than being able to repeat a scaling task. When listeners attempt to transcribe dysarthric speech, a variety of strategies may be at work, whereas in a scaling task, listeners consistently apply a perceptual strategy. Further research is warranted in the area of listener reliability in intelligibility metrics.

Other Considerations

Several factors should be kept in mind when interpreting the findings from this study. First, listeners heard the stimuli in the presence of multitalker babble, which is thought to produce an ecologically valid environment. In addition, authors such as Yorkston et al. (2007) have highlighted the need for investigating intelligibility of dysarthric speech in adverse listening conditions. However, because intelligibility of dysarthria in background noise has only begun to be investigated, drawing parallels from the results of this study to other populations or environments should only be done with caution.

It should also be noted that the sample of speakers with MS and PD considered in this study were highly intelligible, as discussed in the methods section. Again, this caveat speaks to the caution that should be taken when extending the current results to other populations. It is unclear how the methods of increasing intelligibility considered in this study would affect less intelligible speakers. Intelligibility results, and the difference between metrics of measuring intelligibility, may differ more for less intelligible/more severe speakers.

A lack of diversity in the listener subjects may be seen as a limitation. All listeners were recruited on the University at Buffalo campus and were of similar age, and education level (see methods section for full description of participants). However, the listeners were likely similar to

those utilized in the Tjaden et al. (2014) and Kuo et al. (2014) studies, and therefore, for our purposes, still aided in examining the relationship between the metrics, albeit for naive listeners only. In addition, the source of listener variation in intelligibility judgments, from naive and/or experienced listeners (e.g., Speech-Language Pathologists or others familiar with the speech of speakers with dysarthria), is a topic of ongoing study (McHenry, 2011). At the present time, it is uncertain whether experienced listeners or naive listeners yield better reliability in intelligibility judgments, as results have been inconsistent across studies, for scaling tasks (Neel, 2009; Van Nuffelen et al., 2009) and transcription tasks (Bunton et al., 2001; McHenry, 2011; Tjaden, Lam, & Wilding, 2013).

Last, a limitation may be that different listeners performed transcription and VAS. Although listeners from Tjaden et al. (2014) and Kuo et al. (2014) were demographically similar to those in the current study, having listeners perform both transcription and VAS may have yielded different results. A better design may be to have the same listeners do both the transcription and the scaling task, as in Tjaden, Kain, and Lam (2014) and Hustad (2006b). Because of the large number of sentences to be judged in the current study, having listeners perform transcription and VAS during the same session would not have been viable.

Clinical Implications

The present study both replicates and extends previous research, specifically that by Tjaden et al. (2014), Kuo et al. (2014), and Hustad (2006b). While Hustad's (2006b) study included only four participants with dysarthria, the current study included 46 participants with either a diagnosis of MS or PD (e.g., 30 speakers with MS and 16 speakers with PD). While a much larger sample was utilized in the current study, the results were similar to Hustad (2006b) who also found that percent estimates underestimated transcription scores. In the current study

and in Hustad's (2006b) study, scores from transcription and VAS were highly correlated, and listener reliability tended to be slightly higher in the VAS task than in the transcription task. These results suggest that a less time-consuming task, such as the VAS task, may be a viable substitute for the more time-consuming transcription task when calculating intelligibility in a clinical population in at least some circumstances. However, since there was variability among speakers with regard to the pattern of intelligibility and the strength of the relationship between the two metrics, clinicians should be cautious to use the same measure with a single patient over time, or between patients, if the purpose is to compare intelligibility from one measurement to another.

Miller (2013) and Liss (2007) bring up an important point about the use of scaling tasks. Both authors concluded that scaling tasks cannot be used to examine the factors that influence intelligibility, either within speakers or within listeners. When only a single number is obtained from an intelligibility metric, there is no explanatory capacity. This single number cannot aid in identifying therapeutic targets that may be useful for increasing speaker intelligibility (Miller, 2013), or in identifying listener strategies that that may help in increasing intelligibility (Liss, 2007). At this point in time, only sentence intelligibility metrics that involve writing the speaker's message have the type of explanatory capacity required to aid in therapy decisions or to simply better understand where breakdowns in intelligibility occur.

The purpose of the studies by Tjaden et al. (2014) and Kuo et al. (2014) was to compare the effects of rate manipulation, increased vocal intensity, and clear speech on intelligibility in an attempt to aid in therapy decisions. Results from these studies showed that listener perceptions of intelligibility improved for speakers with MS, PD, and controls in the clear and loud conditions relative to the habitual condition, and that intelligibility was not improved in the slow and fast

conditions relative the habitual condition. Overall, the studies demonstrated that clear speech and an increased vocal intensity have similar benefits on scaled intelligibility in the clinical population examined. The results of the current study corroborate the results from the previous two studies and lend support to the finding that clear speech and increased vocal intensity have the potential to improve intelligibility and that, on the whole, rate manipulation shows less promise for aiding intelligibility, at least for speakers with MS or PD with relatively mild involvement.

Directions for Future Research

Further research is warranted to examine the variables that contribute to intelligibility, such as listener error patterns and speech production characteristics. Many of these have been discussed throughout this paper, and include, but are not limited to: severity and type of dysarthria, presence of background noise, listener experience, and type of stimuli. Furthermore, future research could examine or create metrics of intelligibility that are potentially less time and labor consuming, but also offer information about the source of increased or decreased intelligibility. The current research, as well as the two related studies of Tjaden et al. (2014) and Kuo et al. (2014) could be extended upon to determine if the global techniques considered would be as beneficial for improvements in intelligibility when used in therapeutic protocols, rather than in elicited conditions as was used in these studies. Based on the current research, it appears that a scaling task may be a suitable substitute for transcription-based intelligibility measures, especially if lack of time and labor may be barriers to obtaining measures of intelligibility.

Conclusion

The current study sought to investigate the relationship between two metrics of sentence intelligibility in adults with Parkinson's Disease (PD), Multiple Sclerosis (MS), and healthy

controls. Orthographic transcription, which is considered an objective measure of intelligibility, and a VAS task, considered a subjective measure of intelligibility, were the two metrics of intelligibility examined. The primary results of this study showed the pattern of intelligibility to be very similar between scores from transcription and scaled scores from the VAS task. In addition, transcription scores were higher in magnitude than the VAS scores, but correlation analyses showed these two metrics of intelligibility to be highly correlated. Lastly, both interlistener and intralistener reliability were slightly higher for VAS than for transcription. These results suggest that although transcription is the gold standard for measuring intelligibility, there are instances when the less time and labor-consuming VAS task could be used as an alternative. Results from the current study support using a scaling measure to quantify intelligibility in an efficient way in both research and clinical settings, assuming that listener error patterns are not of interest.

Appendices

Appendix A: Individual Speaker Percent Correct Scores per Condition

Speaker	Habitual	Clear	Fast	Loud	Slow
CSF01	84.00	93.20	89.20	91.20	76.00
CSF02	71.60	69.60	48.00	72.80	65.60
CSF03	91.20	95.60	71.20	90.40	93.20
CSF04	90.00	92.40	60.40	94.80	82.40
CSF05	88.40	97.60	76.40	96.40	94.40
CSF06	93.20	96.40	96.40	97.20	93.20
CSF07	91.60	93.20	92.40	97.60	82.00
CSF08	73.60	91.20	44.40	84.80	79.20
CSF09	87.20	91.20	87.60	92.40	91.60
CSF10	82.00	83.20	70.00	87.20	64.40
CSF11	91.20	91.20	84.00	92.00	78.80
CSF12	78.40	94.00	54.00	86.40	72.00
CSF13	80.00	89.20	83.20	95.20	88.40
CSF14	94.00	98.80	89.20	98.80	96.80
CSF15	95.60	96.00	92.40	92.80	78.80
CSF16	92.00	84.40	83.20	95.60	87.60
CSF17	96.00	91.20	92.80	97.60	94.80
CSF18	95.20	92.40	85.60	93.20	91.20
CSF19	95.20	96.40	72.00	92.00	92.00
CSF20	87.20	90.80	46.80	92.40	68.80
CSF21	92.40	94.00	86.00	100.00	92.00
CSF22	94.40	95.20	92.00	98.00	96.40
CSM01	74.00	83.20	82.00	73.60	58.40
CSM02	83.60	87.20	78.00	86.40	83.60
CSM03	79.20	83.60	64.00	89.20	76.40
CSM04	62.80	93.60	60.00	89.60	46.80
CSM05	86.80	82.40	75.20	88.80	88.00
CSM06	66.67	78.40	51.20	78.00	74.00
CSM07	74.00	87.60	72.80	82.80	77.60
CSM08	72.40	86.80	55.60	84.40	83.20
CSM09	86.80	86.80	90.80	90.80	85.60

CSM10	92.00	98.80	90.00	94.40	93.60
MSF01	86.80	95.20	88.80	96.80	93.60
MSF02	57.20	59.20	62.80	64.80	23.60
MSF03	96.00	94.40	74.29	98.40	79.60
MSF04	61.20	86.00	57.20	89.20	79.60
MSF05	82.80	91.20	56.80	86.80	76.40
MSF06	52.00	58.80	55.60		47.20
MSF07	88.80	94.80	76.00	93.60	87.20
MSF08	84.40	92.40	89.20	94.00	86.40
MSF09	28.80	42.40	23.20	34.00	32.00
MSF10	88.80	87.20	80.80	93.60	92.40
MSF11	92.40	91.20	68.00	93.20	86.00
MSF12	92.80	98.40	82.86	98.00	89.60
MSF13	79.60	84.40	79.20	93.60	80.40
MSF14	90.40	91.60	84.80	95.60	85.60
MSF15	68.80	95.60	67.60	89.20	66.00
MSF16	94.40	90.00	85.60	97.20	98.00
MSF17	78.00	76.00	68.40	82.00	61.20
MSF18	77.60	93.20	66.00	90.80	79.60
MSF19	88.80	82.40	80.80	96.00	83.60
MSF20	60.80	69.60	54.00	75.20	73.20
MSM01	28.80	46.40	44.40	35.60	31.20
MSM02	77.20	84.00	79.60	88.00	77.60
MSM03	88.40	95.20	91.20	96.00	94.00
MSM04	53.60	63.20	44.40	63.20	51.20
MSM05	54.80	60.80	72.00	64.80	66.40
MSM06	94.80	96.80	97.60	93.20	93.60
MSM07	66.80	89.20	64.40	86.00	90.00
MSM08	94.40	94.40	88.80	97.60	88.00
MSM09	82.80	84.80	81.20	84.40	87.60
MSM10	74.40	79.60	60.40	82.40	74.00
PDF01	49.60	48.80	28.40	50.80	48.40
PDF02	69.60	77.20	70.80	83.20	66.00
PDF03	11.20	36.40	6.00	22.80	14.80
PDF04	79.20	83.20	22.40	87.20	63.20

PDF05	73.20	73.60	82.80	83.20	88.00
PDF06	80.00	92.40	58.80	86.40	72.00
PDF07	31.20	63.20	45.60	46.80	36.00
PDF08	88.40	87.20	80.40	94.40	90.40
PDM01	62.40	75.20	56.00	80.40	81.20
PDM02	83.60	74.00	71.20	72.00	84.00
PDM03	70.00	85.20	60.40	76.40	72.40
PDM04	68.80	87.60	52.80	84.80	71.20
PDM05	18.40	33.20	11.60	21.20	24.40
PDM06	52.80	82.40	51.60	82.00	72.80
PDM07	37.20	39.20	31.20	46.80	47.20
PDM08	60.40	83.20	48.40	86.40	62.20
Minimum	11.20	33.20	6.00	21.20	14.80
Maximum	96.00	98.80	97.60	100.00	98.00
Mean	75.96	83.03	68.22	83.83	75.34
SD	19.21	15.51	20.32	17.33	18.93

Appendix B: Correlations per Speaker

Speaker	Correlation (r)	P Value
CSF01	0.500	0.000
CSF02	0.685	0.000
CSF03	0.738	0.000
CSF04	0.742	0.000
CSF05	0.764	0.000
CSF06	0.387	0.005
CSF07	0.722	0.000
CSF08	0.775	0.000
CSF09	0.413	0.003
CSF10	0.561	0.000
CSF11	0.710	0.000
CSF12	0.737	0.000
CSF13	0.404	0.004
CSF14	0.387	0.006
CSF15	0.557	0.000
CSF16	0.559	0.000
CSF17	0.494	0.000
CSF18	0.355	0.110
CSF19	0.696	0.000
CSF20	0.813	0.000
CSF21	0.720	0.000
CSF22	0.518	0.000
CSM01	0.862	0.000
CSM02	0.556	0.000
CSM03	0.514	0.000
CSM04	0.818	0.000
CSM05	0.843	0.000
CSM06	0.741	0.000
CSM07	0.697	0.000
CSM08	0.586	0.000
CSM09	0.563	0.000
CSM10	0.580	0.000

MSF01	0.523	0.000
MSF02	0.551	0.000
MSF03	0.698	0.000
MSF04	0.707	0.000
MSF05	0.802	0.000
MSF06	0.787	0.000
MSF07	0.518	0.000
MSF08	0.366	0.009
MSF09	0.384	0.006
MSF10	0.472	0.001
MSF11	0.592	0.000
MSF12	0.311	0.028
MSF13	0.778	0.000
MSF14	0.355	0.011
MSF15	0.618	0.000
MSF16	0.422	0.002
MSF17	0.801	0.000
MSF18	0.799	0.000
MSF19	0.586	0.000
MSF20	0.568	0.000
MSM01	0.752	0.000
MSM02	0.802	0.000
MSM03	0.575	0.000
MSM04	0.760	0.000
MSM05	0.678	0.000
MSM06	0.465	0.001
MSM07	0.701	0.000
MSM08	0.257	0.072
MSM09	0.398	0.004
MSM10	0.667	0.000
PDF01	0.614	0.000
PDF02	0.730	0.000
PDF03	0.641	0.000
PDF04	0.856	0.000

Appendix C: Intralistener Reliability per Listener

Listener	Proportion of Exact Matches	Tjaden et al. (2014) Listener	Reliability (Pearson Product Coefficient)
L01	0.53	**L01**	0.80
L02	0.62	**L02**	0.70
L03	0.88	**L03**	0.74
L04	0.83	**L04**	0.60
L05	0.63	**L05**	0.74
L06	0.62	**L06**	0.75
L07	0.77	**L07**	0.63
L08	0.76	**L08**	0.82
L09	0.73	**L09**	0.80
L10	0.53	**L10**	0.87
L11	0.32	**L11**	0.72
L12	0.61	**L12**	0.68
L13	0.42	**L13**	0.67
L14	0.42	**L14**	0.67
L15	0.56	**L15**	0.78
L16	0.87	**L16**	0.65
L17	0.67	**L17**	0.67
L18	0.68	**L18**	0.88
L19	0.67	**L19**	0.71
L20	0.75	**L20**	0.70
L21	0.66	**L21**	0.79
L22	0.85	**L22**	0.80
L23	0.72	**L23**	0.72
L24	0.87	**L24**	0.61
L25	0.72	**L25**	0.68
L26	0.78	**L26**	0.63
L27	0.53	**L27**	0.73
L28	0.84	**L28**	0.67
L29	0.65	**L29**	0.68
L30	0.84	**L30**	0.68

L31	0.50	**L31**	0.70
L32	0.76	**L32**	0.75
L33	0.61	**L33**	0.83
L34	0.69	**L34**	0.65
L35	0.54	**L35**	0.70
L36	0.77	**L36**	0.73
L37	0.62	**L37**	0.65
L38	0.63	**L38**	0.60
L39	0.66	**L39**	0.68
L40	0.74	**L40**	0.64
L41	0.63	**L41**	0.64
L42	0.54	**L42**	0.83
L43	0.68	**L43**	0.77
L44	0.69	**L44**	0.74
L45	0.56	**L45**	0.85
L46	0.70	**L46**	0.72
L47	0.69	**L47**	0.60
L48	0.32	**L48**	0.73
L49	0.45	**L49**	0.62
L50	0.67	**L50**	0.70
Minimum	0.32	**Minimum**	0.60
Maximum	0.88	**Maximum**	0.88
Mean	0.66	**Mean**	0.71
SD	0.13	**SD**	0.07

Appendix D: Interlistener Reliability per List

List	ICC Single	ICC Average	Tjaden et al. (2014) List	ICC Single	ICC Average
A	0.45	0.80	**A**	0.59	0.88
B	0.48	0.82	**B**	0.68	0.91
C	0.48	0.82	**C**	0.57	0.87
D	0.33	0.80	**D**	0.58	0.87
E	0.54	0.86	**E**	0.57	0.87
F	0.42	0.78	**F**	0.55	0.86
G	0.49	0.83	**G**	0.64	0.90
H	0.44	0.79	**H**	0.56	0.86
I	0.41	0.78	**I**	0.60	0.88
J	0.45	0.80	**J**	0.54	0.85
Minimum	0.33	0.78	**Minimum**	0.54	0.85
Maximum	0.54	0.86	**Maximum**	0.68	0.91
Mean	0.45	0.81	**Mean**	0.59	0.87
SD	0.06	0.02	**SD**	0.04	0.02

References

Benedict, R. H. B., Bruce, J. M., Dwyer, M. G., Abdelrahman, N., Hussein, S., Weinstock-Guttman, B., . . . Zivadinov, R. (2006). Neocortical atrophy, third ventricular width, and cognitive dysfunction in Multiple Sclerosis. *Archives of Neurology, 63*, 1301-1306.

Benedict, R. H. B., Weinstock-Guttmann, B., Fishman, I., Sharma, J., Tjoa, C. W., & Bakshi, R. (2004). Prediction of neuropsychological impairment in Multiple Sclerosis: Comparison of conventional magnetic resonance imaging measured of atrophy and lesion burden. *Archives of Neurology, 61*, 226-230.

Beukelman, D. R., Fager, S., Ullman, C., Hanson, E., & Logemann, J. (2002). The impact of speech supplementation and clear speech on the intelligibility and speaking rate of people with traumatic brain injury. *Journal of Medical Speech-Language Pathology, 10*, 237-242.

Bunton, K. (2006). Fundamental frequency as a perceptual cue for vowel identification in speakers with Parkinson's disease. *Folia Phoniatricia et Logopaedica, 58*, 323-339.

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics, 15*(3), 181-193.

Cannito, M. P., Suiter, D. M., Beverly, D., Chorna, L., Wolf, T., & Pfeiffer, R. (2012). Sentence intelligibility before and after voice treatment in speakers with idiopathic Parkinson's disease. *Journal of Voice, 26*(2), 214-219.

Choe, Y., Liss, J. M., Azuma, T., & Mathy, P. (2012). Evidence of cue use and performance differences in deciphering dysarthric speech. *Journal of the Acoustical Society of America, 131*(2), 112-118.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Dagenais, P. A., Brown, G. R., & Moore, R. E. (2006). Speech rate effects upon intelligibility and acceptability of dysarthric speech. *Clinical Linguistics & Phonetics, 20*(2/3), 141-148.

Dagenais, P. A., Garcia, J. M., & Watts, C. R. (1998). Acceptability and intelligibility of mildly dysarthric speech by different listeners. In M. P. Cannito, K. M. Yorkston, & D. R. Beukelman, (Eds.), *Neuromotor speech disorders: Nature, assessment, and management* (pp. 229-239). Baltimore, MD: Paul H. Brookes Publishing Co.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). St. Louis, MO: Elsevier Mosby.

Duffy, J. R. (2007). History, current practice, and future trends and goals. In G. Weismer (Ed.), *Motor speech disorders* (pp. 7-56). San Diego, CA: Plural Publishing, Inc.

El Sharkawi, A., Ramig, L., Logemann, J. A., Pauloski, B. R., Rademaker, A. W., Smith, C. H., . . . Werner, C. (2002). Swallowing and voice effects of Lee Silverman Voice Treatment (LSVT™): A pilot study. *Journal of Neurology, Neurosurgery & Psychiatry, 72*, 31-36.

Feenaughty, L., Tjaden, K., & Sussman, J. (2014). Relationship between acoustic measures and judgments of intelligibility in Parkinson's disease: A within-speaker approach. *Clinical Linguistics & Phonetics*, 1-22.

Ferguson, S. H. (2012). Talker differences in clear and conversational speech: Vowel intelligibility for older adults with hearing loss. *Journal of Speech, Language, and Hearing Research, 55*, 779-790.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 112*(1), 259-271.

Ferguson, S. H., & Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 135*(6), 3570-3584.

Hanson, E. K., Beukelman, D. R., Fager, S., & Ullman, C. (2004). Listener attitudes toward speech supplementation strategies used by speakers with dysarthria. *Journal of Medical Speech-Language Pathology, 12*, 161-166.

Hustad, K. C. (2006a). A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology, 15*, 268-277.

Hustad, K. C. (2006b). Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatricia et Logopaedica, 58*, 217-228.

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51*, 562-573.

Hustad, K. C., & Weismer, G. (2007). A continuum of interventions for individuals with dysarthria: Compensatory and rehabilitative treatment approaches. In G. Weismer (Ed.), *Motor speech disorders* (pp. 261-303). San Diego, CA: Plural Publishing, Inc.

Huttunen, K., & Sorri, M. (2004). Methodological aspects of assessing speech intelligibility among children with impaired hearing. *Acta Oto-laryngologica, 124*, 490-494.

Kain, A., Amano-Kusumoto, A., & Hosom, J. P. (2008). Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *Journal of the Acoustical Society of America, 124*, 2308-2319.

Kent, R. D., & Kim, Y. (2011). The assessment of intelligibility in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 21-37). San Diego, CA: Plural Publishing Inc.

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*, 482-499.

Kuo, C., Tjaden, K., & Sussman, J. E. (2014). Acoustic and perceptual correlates of faster-than-habitual speech produced by speakers with Parkinson's disease and multiple sclerosis. *Journal of Communication Disorders, 52*, 156-169.

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research, 55*, 1807-1821.

Liss, J. M. (2007). The role of speech perception in motor speech disorders. In G. Weismer (Ed.), *Motor speech disorders* (pp. 187-219). San Diego, CA: Plural Publishing, Inc.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *Journal of the Acoustic Society of America, 112*(6), 3022-3030.

Logan, K. J., Roberts, R. R., Pretto, A. P., & Morey, M. J. (2002). Speaking slowly: Effects of four self-guided training approaches on adults' speech rate and naturalness. *American Journal of Speech-Language Pathology, 11*(2).

Logemann, J. A., Fisher, H. B., Boshes, B., & Blosnky, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *Journal of Speech and Hearing Disorders, 43*, 47-57.

Mackenzie, C. (2011). Cognition and its assessment in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 141-156). San Diego, CA: Plural Publishing Inc.

Maniwa, K., Jongman, A., & Wade, T. (2008). Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *The Journal of the Acoustical Society of America, 123*, 1113-1125.

McHenry, M. A. (2003). The effect of pacing strategies on the variability of speech movement sequences in dysarthria. *Journal of Speech, Language, and Hearing Research, 46*, 702-710.

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Journal of Speech-Language Pathology, 20*, 119-123.

Metz, D. E., Schiavetti, N., Samar, V. J., & Sitler, R. W. (1990). Acoustic dimensions of hearing-impaired speakers' intelligibility: Segmental and suprasegmental characteristics. *Journal of Speech and Hearing Research, 33*, 476-487.

Milenkovic, P. (2005). TF32 [Computer program]. Madison: University of Wisconsin-Madison.

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

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *Journal of the Acoustical Society of America, 128*, 435-443.

Molloy, D. (1999). *Standardized Mini-Mental State Examination*. Troy, NY: New Grange Press.

Molloy, D. W., Standish, T. I. M., & Lewis, D. L. (2005). Screening for mild cognitive impairment: Comparing the SSMSE and the ABCS. *Canadian Journal of Psychiatry, 51*(1), 52-58.

Neel, A. T. (2009). Effects of loud and amplified speech on sentence and word intelligibility in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 52*, 1021-1033.

Picheny, M. A., Durlach, N. L., & Braida, L. D. (1985). Speaking clearly for the hard of hearing: I. Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 28*, 96-103.

Ramig, L. O., Bonati, C., Lemke, J. H., & Horrii, Y. (1994). Voice treatment for patients with Parkinson's disease: Development of an approach and preliminary efficacy data. *Journal of Medical Speech-Language Pathology, 4*(2), 191-210.

Ramig, L. O., Countryman, S., Thompson, L. L., & Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson's disease. *Journal of Speech and Hearing Research, 38*, 1232-1251.

Samar, V. J., & Metz, D. E. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing impaired population. *Journal of Speech and Hearing Research, 31*, 207-316.

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

Schum, D. J. (1996). Intelligibility of clear and conversational speech of young and elderly talkers. *Journal of the American Academy of Audiology, 7*, 212-218.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3*, 236-264.

Spitzer, S. M., Liss, J. M., & Mattys, S. L. (2007). Acoustic cues to lexical segmentation: A study of resynthesized speech. *Journal of the Acoustical Society of America, 122*(6), 3678-3687.

Sussman, J., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and Multiple Sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208-1219.

The Institute of Electrical and Electronics Engineers. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*, 225-246.

Tjaden, K., Kain, A., & Lam, J. (2014). Hybridizing conversational and clear speech to investigate the source of increased intelligibility in Parkinson's disease. *Journal of Speech, Language, and Hearing Research.*

Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel acoustics in Parkinson's disease and Multiple Sclerosis: Comparison of clear, loud and slow speaking conditions. *Journal of Speech, Language, and Hearing Research, 56*, 1485-1502.

Tjaden, K., Richards, E., Kuo, C., Wilding, G., & Sussman, J. (2013). Acoustic and perceptual consequences of clear and loud speech. *Folia Phoniatrica et Logopaedica, 65,* 214-220.

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014). Impact of clear, loud and slow speech on scaled intelligibility and speech severity in Parkinson's disease and Multiple Sclerosis. *Journal of Speech, Language, and Hearing Research, 57*, 779-792.

Tjaden, K., & Wilding, G. (2010). Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics & Phonetics, 25*(2), 155-168.

Tjaden, K., & Wilding, G. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research, 47*, 766-783.

Tsao, Y. C., Weismer, G., & Iqbal, K. (2006). Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech, Language, and Hearing Research, 49*, 1156-1164.

Turner, G. S., & Weismer, G. (1993). Characteristics of speaking rate in the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Speech and Hearing Research, 36*, 1134-1144.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing: IV. Further studies of the role of speaking rate. *Journal of Speech and Hearing Research, 39*, 494-509.

Van Nuffelen, G., De Bodt, M., Vanderwegen, J., Van de Heyning, P., & Wuyts, F. (2010). Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatricia et Logopaedica, 62*, 110-119.

Van Nuffelen, G., De Bodt, M., Wuyts, F., & Van de Heyning, P. (2009). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatricia et Logopaedica, 61*, 69-75.

Weismer, G. (2009). Speech intelligibility. In M. J. Ball, M. R. Perkins, N. Muller, & S. Howard (Eds), *The handbook of clinical linguistics* (pp. 568-582). Oxford, UK: Blackwell Publishing Ltd.

Whitehall, T., Ciocca, V., & Yiu, E. (2004). Perceptual and acoustic predictors of intelligibility and acceptability in Cantonese speakers with dysarthria. *Journal of Medical Speech-Language Pathology, 12*, 229-233.

Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Hakel, M. (2010). *Management of Motor Speech Disorders in Children and Adults* (3rd ed.). Austin, TX: PRO-ED, Inc.

Yorkston, K., Beukelman, D. R., & Tice, R. (1996). *Sentence Intelligibility Test*. Lincoln, NE: Tice Technologies.

Yorkston, K., Beukelman, D., & Traynor, C. (1984). *Assessment of Intelligibility of Dysarthric Speech*. Austin, TA: PRO-ED, Inc.

Yorkston, K. M., Hakel, M., Beukelman, D. R., & Fager, S. (2007). Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review. *Journal of Medical Speech-Language Pathology, 15*(2), xi-xxxvi.

Yorkston, K. M., Hammen, V., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders, 55*, 550-560.

Yunusova, Y., Weismer, G., Kent, R. D., & Rusche, N. M. (2005). Breath-group intelligibility in dysarthria: Characteristics and underlying correlates. *Journal of Speech, Language, and Hearing Research, 48*, 1294-1310.