JSLHR

Research Note

Comparison of Intelligibility Measures for Adults With Parkinson's Disease, Adults With Multiple Sclerosis, and Healthy Controls

Kaila L. Stipancic,[a] Kris Tjaden,[a] and Gregory Wilding[a]

Purpose: This study obtained judgments of sentence intelligibility using orthographic transcription for comparison with previously reported intelligibility judgments obtained using a visual analog scale (VAS) for individuals with Parkinson's disease and multiple sclerosis and healthy controls (K. Tjaden, J. E. Sussman, & G. E. Wilding, 2014).
Method: Speakers read Harvard sentences in habitual, clear, loud, and slow conditions. Sentence stimuli were equated for peak intensity and mixed with multitalker babble. A total of 50 listeners orthographically transcribed sentences. Procedures were identical to those for a VAS reported in Tjaden, Sussman, and Wilding (2014).
Results: The percent correct scores from transcription were significantly higher in magnitude than the VAS scores. Multivariate linear modeling indicated that the pattern of findings for transcription and VAS was virtually the same with respect to differences among groups and speaking conditions. Correlation analyses further indicated a moderately strong, positive relationship between the two metrics. The majority of these correlations were significant. Last, intrajudge and interjudge listener reliability metrics for the two intelligibility tasks were comparable.
Conclusion: Results suggest that there may be instances when the less time-consuming VAS task may be a viable substitute for an orthographic transcription task when documenting intelligibility in mild dysarthria.

Intelligibility refers to the degree or the accuracy with which a listener recovers the acoustic signal or message produced by a speaker (Duffy, 2013). Intelligibility has also been described or defined as how effective one is in his or her communication (Cannito et al., 2012), the ease with which the acoustic speech signal is understood (Kim & Kuo, 2012), or the extent to which the acoustic signal is understood (Tjaden, Sussman, & Wilding, 2014). Intelligibility should be distinguished further from the perceptual construct of comprehensibility (Yorkston, Strand, & Kennedy, 1996). *Comprehensibility*, as defined by Yorkston, Beukelman, and Tice (1996), refers to how much of the acoustic speech signal a listener understands when gestures, orthographic cues, semantic cues, and other types of contextual information are available (for a discussion of differences among the perceptual constructs of intelligibility, comprehensibility, and comprehension, see a review in Hustad, 2008).

[a]University at Buffalo, NY
Correspondence to Kaila L. Stipancic: kailalau@buffalo.edu
Editor: Jody Kreiman
Associate Editor: Amy Neel
Received August 3, 2015
Revision received October 20, 2015
Accepted October 26, 2015
DOI: 10.1044/2015_JSLHR-S-15-0271

Intelligibility is a common effect of dysarthria. Therefore, quantifying intelligibility is necessary for determining the overall degree of communication impairment and for demonstrating the efficacy of dysarthria therapy techniques. In addition, by measuring intelligibility over time, treatment progress and disease progression can be quantified. In everyday conversation, speech is typically produced in utterances composed of multiple words rather than single words or phonemes. Therefore, sentence-level metrics of intelligibility are presumed to index the magnitude of an individual's communicative difficulty (Weismer, 2009). As discussed in the following section, transcription and scaling tasks have frequently been used to measure sentence intelligibility in dysarthria.

Methods for Measuring Intelligibility

Transcription

Transcription has been characterized as an objective intelligibility measure (Miller, 2013; Weismer, 2009) and involves the listener writing the speaker's message word for word. The word-for-word transcription is then compared with the target production, and the percentage of words correctly transcribed is calculated. Orthographic transcription

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

is time consuming for both the listeners who must write or type what they think they hear and the individuals who score the accuracy of transcription. Computerized scoring improves efficiency, but even then, responses must be checked for spelling and other errors. However, transcription is the gold standard for quantifying intelligibility in the dysarthria literature, and the Sentence Intelligibility Test (SIT; Yorkston, Beukelman, & Tice, 1996) is undoubtedly one of the most widely used, published clinical tools for quantifying intelligibility (Duffy, 2013; Yorkston, Beukelman, Strand, & Hakel, 2010). Transcription also has been considered to yield good reliability both within and among listeners (Miller, 2013). Thus, dysarthria studies using transcription do not consistently report listener reliability (see Hustad, 2006a, 2006b; Liss, Spitzer, Caviness, & Adler, 2002; McHenry, 2011), although reliability has been reported in a few studies (see Bunton, Kent, Kent, & Duffy, 2001; Tjaden, Kain, & Lam, 2014; Tjaden & Wilding, 2010).

Scaling Tasks

In comparison to transcription, scaling tasks have been characterized as more subjective measures because listeners are instructed to estimate how much of the speaker's message they understand or to judge the extent to which the message was understood (Hustad & Weismer, 2007; Miller, 2013). Overall, scaling tasks for quantifying intelligibility have been criticized in the dysarthria literature. Listener reliability for these tasks has been questioned and, in certain cases, has been found to be poorer than is ideal for research purposes (Miller, 2013; Schiavetti, 1992). However, some research suggests that reliability for transcription and a visual analog scale (VAS) may be comparable. Tjaden, Kain, and Lam (2014) reported intrajudge correlation coefficients ranging from .57 to .99 (M = .80, SD = .13) for a scaling task and correlation coefficients ranging from .58 to 1.00 (M = .80, SD = .13) for a transcription task. Although this study included only 40 listeners who judged sentences produced by two speakers with Parkinson's disease, the results suggest that it should not be assumed that listener reliability for transcription is superior to a scaling task. Despite concerns regarding listener reliability, scaling tasks provide some attractive benefits. These tasks are less time consuming and labor intensive than orthographic transcription (Miller, 2013). Scaling tasks may also be easily applied to longer connected speech tasks commonly obtained in clinical practice, such as paragraph reading.

Visual analog scaling is a type of scaling task that shows promise for measuring intelligibility (Kent & Kim, 2011; Van Nuffelen, De Bodt, Vanderwegen, Van de Heyning, & Wuyts, 2010). A VAS involves listeners choosing a point on a continuous line that does not contain any ticks or intervals to represent their judgment of a given speech sample (Kent & Kim, 2011). For example, Tjaden, Sussman, and Wilding (2014) recently used a computerized VAS task in which listeners judged intelligibility. Listeners were presented with a continuous 150-mm vertical-oriented scale on a computer monitor, with endpoints of the scale labeled "*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.

Comparison of Intelligibility Measures

Several dysarthria studies have used multiple tasks (i.e., direct magnitude estimation vs. transcription) to index intelligibility for different types of speech stimuli (e.g., Metz, Schiavetti, Samar, & Sitler, 1990; Sussman & Tjaden, 2012; Yunusova, Weismer, Kent, & Rusche, 2005). However, limited knowledge is available about how objective and subjective metrics of intelligibility compare for the same stimuli. In fact, Weismer, Barlow, Smith, and Caviness (2008) commented that the "proper work to identify the benefits and problems of different measures has yet to be done" (p. 284).

In one of the few studies that directly compared intelligibility metrics for the same stimuli, Hustad (2006b) found that, for four speakers with dysarthria, on average, transcription scores were higher versus scores obtained from listeners estimating the percentage of words understood. However, the magnitude of the difference between transcription scores and percent estimates varied across speakers. Thus, although a few studies have reported different types of intelligibility metrics for the same stimuli, to date, no large-scale study has compared orthographic transcription and a VAS in dysarthria.

Summary and Purpose

Orthographic transcription is the gold standard for measuring intelligibility, but it is labor intensive for the listener and the individual scoring the accuracy of responses. Less time-consuming methods for measuring intelligibility, such as subjective scaling tasks, are attractive. However, few studies have looked at how objective and subjective metrics of intelligibility compare. If transcription and scaling are found to yield equivalent levels of severity or outcomes as well as similar listener reliability, then there may be instances when the more efficient scaling task could be used.

Therefore, the purpose of the present study was to compare the objective intelligibility metric of orthographic transcription with the subjective intelligibility metric of a VAS. Toward this end, sentences read by speakers with multiple sclerosis (MS) and Parkinson's disease (PD) as well as healthy controls were orthographically transcribed for comparison to VAS judgments of sentence intelligibility reported in Tjaden, Sussman, and Wilding (2014). Sentences in the VAS study were produced in a speaker's typical manner of talking (e.g., habitual condition) as well as in speaking conditions used therapeutically to maximize intelligibility in dysarthria, including clear, loud, and slow. A primary focus of the VAS study was to determine the impact of these speaking conditions on intelligibility. Speaking condition effects were of secondary interest in the present study. That is, inclusion of all sentences and speaking conditions from the VAS study was desirable for the purpose of maximizing the size of the data corpus and to allow more straightforward comparison of transcription

results to those previously reported for a VAS. The following research questions were addressed:

1. Does orthographic transcription yield similar intelligibility differences among speaker groups and speaking conditions, as previously shown using a VAS? In particular, is transcription intelligibility for the PD group significantly reduced relative to the control group, and do the clear and loud conditions yield significantly improved transcription intelligibility relative to the habitual condition?
2. What is the strength of the relationship between the percent correct scores from orthographic transcription and the scale values from a VAS?
3. Are there significant differences in intralistener and interlistener reliability for orthographic transcription and a VAS?

Method

Speakers

The 78 speakers and sentence stimuli from Tjaden, Sussman, and Wilding (2014) were used. Control speakers (n = 32) included 10 men (25–70 years old, M = 56) and 22 women (27–77 years old, M = 57) who reported the absence of neurological disease. Speakers with PD (n = 16) included eight men (55–78 years, M = 67) and eight women (48–78 years, M = 69) who had a medical diagnosis of idiopathic PD. Speakers with MS (n = 30) included 10 men (29–60 years, M = 51) and 20 women (27–66 years, M = 50) who had a medical diagnosis of MS.

Clinical metrics of single-word intelligibility, sentence intelligibility, and scaled speech severity for the Grandfather Passage were reported in detail in Sussman and Tjaden (2012) and are summarized in Table 1 for the purpose of describing the participants. Word intelligibility was obtained using the single-word test of Kent, Weismer, Kent, and Rosenbek (1989). Sentence intelligibility was obtained using the SIT (Yorkston, Beukelman, & Tice, 1996). To obtain perceptual judgments of speech severity for the Grandfather Passage, listeners used a computerized VAS with scale endpoints of 0 (*no impairment*) to 1 (*severe impairment*). These clinical metrics demonstrate that 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 and scaled severity suggest mild dysarthria for many speakers with MS or PD (Yorkston et al., 2010).

Experimental Speech Stimuli and Speech Tasks

Speakers read 25 Harvard psychoacoustic sentences (Institute of Electrical and Electronics Engineers, 1969) in habitual, clear, loud, and slow conditions. For each speaker, a subset of 10 sentences produced in each condition was used for intelligibility testing. Judgments of intelligibility for each speaker were obtained for 40 sentences (i.e., 4 conditions × 10 sentences). Each sentence contained between seven and nine words, and five key words (e.g., nouns, verbs, adjectives, and adverbs). An in-depth description of recording procedures was presented in Tjaden, Sussman, and Wilding (2014).

As reported in the previous study and summarized in Tables 2 and 3, acoustic measures of sound pressure level and articulatory rate were obtained using TF32 (Milenkovic, 2005) to verify the presence of production differences between the speaking conditions. Table 2 indicates that all speaker groups increased mean sound pressure level (SPL) for the loud and clear conditions relative to the habitual condition. Descriptive statistics in Table 3 further indicate a reduced rate for the slow and clear conditions relative to the habitual condition.

Listeners

Listener characteristics for the 50 individuals who orthographically transcribed sentences were the same as in the VAS study (Tjaden, Sussman, & Wilding, 2014). All 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 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.

Table 1. Clinical metrics of intelligibility and speech severity for speaker groups.

Group	Mean % single-word intelligibility[a]	Mean % sentence intelligibility[b]	Mean scaled speech severity score[c]
Control	97 (.01)	94 (2.7)	0.18 (.08)
MS	96 (.03)	93 (4.5)	0.44 (.25)
PD	95 (.03)	85 (10)	0.46 (.21)

Note. Standard deviations are in parentheses. MS = multiple sclerosis; PD = Parkinson's disease.

[a]Single-word test of Kent, Weismer, Kent, and Rosenbek (1989). [b]Orthographic transcription of sentences from the Sentence Intelligibility Test (Yorkston, Beukelman, & Tice, 1996). [c]Score on a visual analog scale of overall speech impairment for a reading passage as judged by three speech pathologists (0 = *no impairment*, 1 = *severe impairment*).

Table 2. Mean sound pressure level in dB SPL as a function of group and condition.

Group	Habitual	Clear	Loud	Slow
Control	73 (2.7)	77 (4.5)	83 (4.0)	73 (4.0)
MS	72 (3.0)	75 (4.4)	80 (3.6)	72 (4.7)
PD	72 (3.2)	75 (4.0)	79 (4.0)	72 (4.6)

Note. Standard deviations are in parentheses. SPL = sound pressure level; MS = multiple sclerosis; PD = Parkinson's disease.

Stimuli Preparation and Perceptual Task

Transcription data in the present study were collected using the same methods that were used to collect the VAS data (Tjaden, Sussman, & Wilding, 2014). Sentences were mixed with multitalker babble with a signal-to-noise ratio of −3 dB to induce a more challenging listening environment and to reduce the likelihood of ceiling effects. Stimuli were presented to individual listeners at 75 dB SPL via headphones (MDR V300, Sony) in a double-walled audiometric booth using custom software. The task took between 2 and 3 hours with breaks and was self-paced.

Sentences for all speakers and conditions were first pooled and 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. Five listeners were assigned to judge each list. Each listener also judged a random selection of 10% of sentences twice to determine intrajudge reliability. After hearing a sentence once, listeners were instructed to type exactly what they heard. Listeners had no knowledge of speakers' neurological diagnoses or the speaking conditions. Custom software saved typed responses for later scoring.

A key word scoring paradigm was used (see also Hustad, 2006a). This key word scoring paradigm involved scoring the five key informational words, including nouns, verbs, adjectives, and adverbs, in each Harvard sentence for a correct or incorrect match with the target. 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. In addition, 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. For each sentence production, the five listeners' responses were pooled, and the number of key words correctly transcribed was tallied. The percent correct scores was tabulated for each speaker in each condition.

Table 3. Mean articulation rate (syllables per second) as a function of group and condition.

Group	Habitual	Clear	Loud	Slow
Control	3.7 (0.44)	2.3 (0.32)	3.2 (0.46)	1.9 (0.48)
MS	3.6 (0.60)	2.7 (0.63)	3.3 (0.69)	2.4 (0.60)
PD	4.1 (0.58)	3.3 (0.75)	4.0 (0.71)	2.9 (0.75)

Note. Standard deviations are in parentheses. MS = multiple sclerosis; PD = Parkinson's disease.

Scoring Reliability

Scoring reliability refers to the consistency or reliability of scoring the transcription responses and 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). 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 .98 to 1.00, with a mean of .99 (SD = .01). Interscorer reliability was determined by having a second scorer who was not involved in the initial scoring rescore 10% of the listener responses. Pearson product–moment correlation coefficients for the first and second scoring of listener responses ranged from .92 to 1.00, with a mean of 0.98 (SD = 0.03). Both intrascorer and interscorer reliabilities are comparable to those from Hustad (2006a, 2006b) and McHenry (2011), and both measures indicate high levels of reliability in the scoring of transcribed responses.

Data Analysis

Dependent measures were characterized using both descriptive and parametric statistics. Analyses are described separately for each of the three research questions.

Research Question 1: Comparing VAS and Transcription Intelligibility

Descriptive statistics (i.e., means, standard deviations) were computed for the percent correct scores for comparison with the descriptive statistics of the VAS from Tjaden, Sussman, and Wilding (2014). This examination served as a descriptive comparison of overall means for transcription versus scaling.

Transcription data were also analyzed using the same parametric statistics applied to the VAS data in Tjaden, Sussman, and Wilding (2014). QQ-plots were generated to evaluate the need for transformations. Inspection of these plots based on the scaled residuals indicated that no transformation of the outcome was needed. A multivariate linear model was fit to the data using SAS version 9.1.3 (SAS Institute, Inc., Cary, NC). The percent correct scores were fit as a function of group (control, MS, PD), condition (habitual, clear, loud and, slow), and a Group × 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.

Research Question 2: Strength of Relationship Between Transcription and VAS

Pearson product–moment correlation coefficients were used to examine the strength of the relationship between the percent correct scores and scale values from the VAS. Two correlation analyses were performed. First, a correlation analysis was computed for each of the 78 speakers for data pooled across conditions and sentences. Second, correlations were computed separately for each condition and group. Given four conditions and three groups, a total of 12 correlations were computed for this second analysis. Because the present investigation is the first large-scale study examining these two metrics of intelligibility, it was deemed important to examine not only group trends but also individual speaker trends.

Research Question 3: Listener Reliability Comparison

For the transcription data, the number of exact word matches was calculated for the 10% of sentences judged twice by each listener. Intralistener 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, 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. By comparison, Pearson product correlation coefficients for scale values from original and reliability trials were calculated to assess intralistener reliability for the VAS data (Tjaden, Sussman, & Wilding, 2014).

Following Neel (2009) and Tjaden, Sussman, and Wilding (2014), interlistener reliability was assessed using intraclass correlation coefficients (ICCs). ICCs were calculated separately for each of the 10 sentence sets because the listeners assigned to judge each of these lists heard different sentences. A two-way mixed-effects model was used to determine the overall consistency of ratings among listeners. Aggregate listener performance was of interest; therefore, average ICC metrics were considered the primary measure of agreement among listeners. ICCs for transcription were summarized using descriptive statistics and descriptively compared with the ICC scores for the VAS data (Tjaden, Sussman, & Wilding, 2014).

Results

Research Question 1: Pattern of Findings for Intelligibility

Descriptive Statistics

Results for transcription intelligibility as a function of group and condition are shown in Figure 1A in the form of means and standard deviations. Results for VAS intelligibility (Tjaden, Sussman, & Wilding, 2014) are shown in Figure 1B. Scores from the VAS could range from 0 (*understand everything*) to 1 (*cannot understand anything*). To allow these scaled values to be more easily compared with the percent correct scores in Figure 1A, the scale was reversed and values were multiplied by 100, so that scale values closer to 100 in Figure 1B represent greater intelligibility.

Figure 1A indicates that transcription intelligibility in each condition was always highest for the control group, followed by the MS group and the PD group. Figure 1B shows this same pattern for the VAS, as the control group was the most intelligible in each condition, followed by the MS group and the PD group. Examination of the two figures further suggests that the overall percent correct scores from transcription were of greater magnitude than the scores from the VAS task.

Using the guideline that changes in sentence intelligibility of approximately 5% are likely clinically meaningful in the context of an adverse perceptual environment such as multitalker babble (e.g., Tjaden, Sussman, & Wilding, 2014; Van Nuffelen et al., 2010), the pattern of transcription intelligibility in Figure 1A was similar for all speaker groups. That is, for each group, the clear and loud conditions did not differ but increased intelligibility relative to the habitual condition by at least 5%. In addition, the slow and habitual conditions did not differ. For all groups, VAS judgments in Figure 1B show that the clear and loud conditions also did not differ but increased intelligibility relative to the habitual condition. As for transcription intelligibility, the habitual and slow conditions did not differ with the VAS.

Parametric Statistics

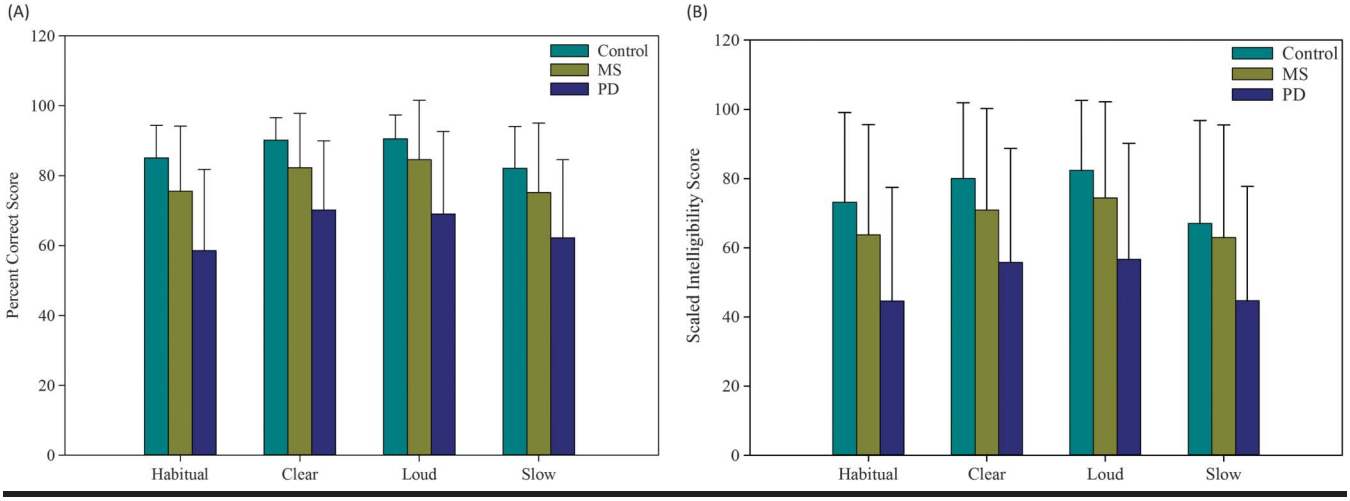
As previously noted in the Data Analysis section, a multivariate linear model was fit to the percent correct scores as a function of group (control, MS, PD), condition (habitual, clear, loud, slow), and a Group × Condition interaction. There were significant main effects of group, $F(2, 71) = 10.77$, $p < .0001$; and condition, $F(3, 71) = 35.75$, $p < .0001$. The Condition × Group interaction was not significant. Follow-up contrast tests indicated that the PD group had poorer intelligibility when compared with both the control ($p < .001$) and MS ($p = .015$) groups. Transcription intelligibility for the clear and loud conditions was significantly better than habitual ($p < .05$). To summarize, for all speaker groups, the clear and loud conditions significantly increased intelligibility relative to the habitual condition, but the clear and loud conditions did not differ. Transcription intelligibility for all groups also was not significantly different for the habitual and slow conditions. As elaborated in the Discussion, these results are virtually identical to those for the VAS (Tjaden, Sussman, & Wilding, 2014).

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

Correlation Analyses

For each of the 78 speakers, correlations for VAS scores and transcription scores were computed for all

Figure 1. Descriptive statistics for (Panel A) percent correct transcription scores and (Panel B) visual analog scale intelligibility scores. The colored bars represent mean intelligibility scores for each group in each condition, and vertical bars indicate *SD*. MS = multiple sclerosis; PD = Parkinson's disease.



sentences pooled across conditions. Across the 78 speakers, correlations ranged from .08 to .87, with an average of .57 (*SD* = .178). All correlations were significant ($p < .05$), with four exceptions (control female 14 [CSF14] $r = .227$, $p = .158$; MS female 7 [MSF07] $r = .165$, $p = .310$; MS female 12 [MSF12] $r = .124$, $p = .444$; MS female 16 [MSF16] $r = .083$, $p = .610$). When these nonsignificant correlations were excluded from the calculation of descriptive statistics, the mean correlation was .60 (*SD* = .151) for the remaining 74 speakers. Therefore, for the majority of speakers, there was a moderately strong relationship between the transcription task scores and the VAS task scores (Cohen, 1988).

We completed a second correlation analysis to examine the strength of the relationship between the transcription task scores and the VAS task scores on a per-condition and per-group basis. All 78 speakers were included in these computations. All correlations were significant ($p < .05$) and ranged from .83 to .99. On average, correlations were strongest for the PD group ($M = .96$, range = .94–.99), followed by the MS group ($M = .95$, range = .95–.97), and the control group ($M = .87$, range = .83–.89).

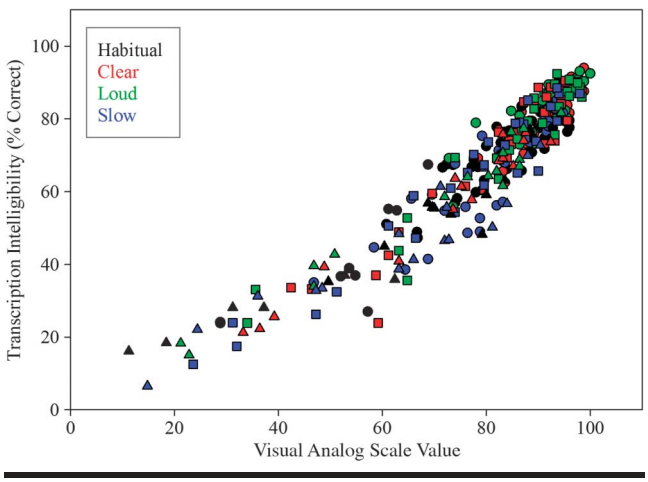
Figure 2 shows a scatter plot of the data with the percent correct transcription scores on the *x*-axis and the VAS intelligibility scores on the *y*-axis. Each symbol on the graph represents a single speaker in a given condition. Condition is designated by symbol color, and group is designated by symbol shape. A visual inspection of Figure 2 suggests a curvilinear relationship between the two intelligibility metrics when data for all groups and conditions are considered. To explore this possibility, we undertook a trend analysis for data pooled across all speakers, groups, and conditions. A linear regression function was statistically significant ($p < .001$) and accounted for nearly 90% of the variance in the relationship between the two intelligibility metrics (i.e., adjusted $R^2 = .89$). A quadratic regression function was also significant ($p < .001$) but only accounted for an additional 3% of the variance (i.e., adjusted $R^2 = .92$) in the relationship between the two intelligibility metrics.

Research Question 3: Listener Reliability

Intralistener Reliability

The intralistener reliability analysis examined the proportion of exact matches in transcription responses and yielded Pearson product–moment correlations from .32 to .88 across the 50 listeners, with a mean of .66 (*SD* = .13). All correlations were significant ($p < .05$). For comparison, in the VAS task, intralistener reliability

Figure 2. Percent correct transcription scores versus visual analog scale scores for each speaker in each condition (each point on the graph represents a single speaker in a given condition). Speaker group is designated by symbol shape, and condition is designated by symbol color. Control = circles, multiple sclerosis = squares, and Parkinson's disease = triangles.



correlation coefficients ranged from .60 to .88 across the 50 listeners, with a mean of .71 (SD = .07) (Tjaden, Sussman, & Wilding, 2014).

Interlistener Reliability

Average interlistener reliability ICCs for transcription intelligibility ranged from .78 to .86 (M = .81, SD = .02), and single-measure ICCs ranged from .33 to .54 (M = .45, SD = .06). All ICC measures, both single and aggregate, were significant ($p < .05$). By comparison, average ICCs across the 50 listeners for scaled intelligibility ranged from .85 to .91 (M = .87, SD = .02), and single-measure ICCs ranged from .54 to .68 (M = .59, SD = .04; Tjaden, Sussman, & Wilding, 2014).

Discussion

Research Question 1: Pattern of Findings for Intelligibility

The pattern of descriptive statistics for the two tasks, as well as the pattern of results for parametric statistics for the two tasks, was similar. For all groups, the clear and loud conditions, but not the slow condition, improved intelligibility relative to the habitual condition. In addition, the PD group was consistently judged to have the poorest intelligibility followed by the MS and control groups. This result held for both transcription and VAS intelligibility.

The results further suggest that raw scores were lower in magnitude for the VAS than for transcription. Hustad (2006b) also found that subjective intelligibility scores in the form of percent estimates were lower than scores derived from a transcription task for four speakers with dysarthria. The similar pattern and difference in magnitude of intelligibility scores for transcription and a VAS has implications for clinicians and researchers. To the extent that transcription and a VAS are both measuring the construct of intelligibility, clinicians and researchers may be able to choose the less labor-intensive VAS task with the knowledge that VAS scores can be expected to be lower than raw percent correct scores for transcription. In addition, because transcription and a VAS yield raw intelligibility scores of different magnitudes, when the purpose is to compare intelligibility findings either across time or across speakers, either transcription or a VAS should be used exclusively.

Research Question 2: Strength of Relationship Between Transcription and VAS

The correlation analyses indicated a moderately strong relationship between the percent correct scores derived from transcription and judgments of intelligibility from the VAS for each of the three speaker groups as well as the majority of individual speakers (Cohen, 1988). The implication is that although the magnitude of the scores may differ, the overall pattern of scores was broadly similar. One implication is that transcription and a VAS task are tapping into the same perceptual phenomenon. However, for four speakers, the two intelligibility metrics were not significantly correlated. This result may be due to the fact that these four speakers received intelligibility scores, both from transcription and VAS, on the higher end of intelligibility, leading to a very restricted range of intelligibility across sentences and conditions.

Research Question 3: Listener Reliability Comparison

Miller (2013) stated that because listeners' "internal yardsticks" differ on subjective intelligibility metrics such as the VAS, the end result is poor interrater reliability (p. 603). Both interlistener and intralistener reliabilities were slightly higher for the VAS task (Tjaden, Sussman, & Wilding, 2014) than for transcription. The present study would therefore appear to contradict Miller's (2013) statement because the VAS was found to be at least as reliable as transcription. Results further suggest that transcription should not be the preferred intelligibility metric solely on the basis of assumptions concerning reliability. Future studies are needed to statistically compare the reliability of the VAS and transcription.

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. However, because intelligibility of dysarthria in background noise has only begun to be investigated (Yorkston, Hakel, Beukelman, & Fager, 2007), drawing parallels to other listening environments should be done with caution. Speakers with MS and PD were also highly intelligible, as indicated by average intelligibility on the SIT in the vicinity of 90%. Thus, caution should be taken when extending the current results to other populations. Intelligibility results, and the difference between metrics of intelligibility, may differ more for less intelligible/more severe speakers.

Last, although listeners from Tjaden, Sussman, and Wilding (2014) were demographically similar to those in the current study and met the same inclusionary criteria as listeners who performed the transcription task, having the same listeners perform both transcription and a VAS may have yielded different results. Future studies may consider having the same listeners perform both the transcription and the scaling tasks as in Hustad (2006b; see also Tjaden, Kain, & Lam, 2014).

Clinical Implications

The present study both replicates and extends previous research. Hustad's (2006b) study included only four speaker participants, and the present study included 46 participants with a diagnosis of MS or PD, and although a much larger sample was used in the present study, the results were similar to those of Hustad (2006b), who also found that percent estimates underestimated transcription scores.

In the present study and in Hustad's (2006b) study, scores from an objective measure of intelligibility (i.e., transcription) and a subjective measure of intelligibility (i.e., VAS or percent estimates) were highly correlated, and listener reliability tended to be slightly higher in the VAS task than in the transcription task. These results support using a less time-consuming scaling measure as a substitute for orthographic transcription in at least some instances. However, because 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 and 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. Overall, the present results 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.

The primary purpose of the Tjaden, Sussman, and Wilding (2014) study was to compare the effects of reduced speech rate, increased vocal intensity, and clear speech on intelligibility in an attempt to inform therapy decisions. Although examining the effects of these conditions was not an aim of the present research, the transcription intelligibility results for the conditions are worth noting. Results showed that intelligibility improved for speakers with MS or PD in the clear and loud conditions relative to the habitual condition and that intelligibility was not improved in the slow condition relative to the habitual condition. The present findings further support the idea that both clear speech and increased vocal intensity have the potential to improve intelligibility in mild dysarthria and that a slowed speech rate 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 variables that contribute to intelligibility, such as listener error patterns and speech production characteristics, including severity and type of dysarthria, presence of background noise, listener experience, and type of stimuli. Furthermore, future research could examine whether transcription of SIT sentences (Yorkston, Beukelman, & Tice, 1996) and VAS judgments of the same sentences yield a similar pattern of results. This may have implications for future software development and advances in the widely used computerized SIT. Last, other perceptual metrics, such as perception of monopitch and naturalness as well speech comprehension are gaining traction (Anand & Stepp, 2015; Fontan, Tardieu, Gaillard, Woisard, & Ruiz, 2015). Their relationship to various intelligibility metrics or tasks is worth further examination.

Acknowledgments

This work was completed as part of the first author's master's thesis. Funding was provided by the Mark Diamond Research Fund of the Graduate Student Association at the University at Buffalo, The State University of New York, Buffalo, NY (PI: Kaila L. Stipancic) and National Institute on Deafness and Other Communication Disorders, Washington, DC, Grant R01DC004689 (PI: Kris Tjaden).

References

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 58,* 1134–1144.

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.

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

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

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., & Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research, 58,* 977–986.

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.

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

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.

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.

Kim, Y., & Kuo, C. (2012). Effect of level of presentation to listeners on scaled speech intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopedica, 64*(1), 26–33.

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. *The Journal of the Acoustical Society of America, 112,* 3022–3030.

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, WI: University of Wisconsin–Madison.

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

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.

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.

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,* 1208–1219.

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, 57,* 1191–1205.

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,* 155–168.

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.

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.

Weismer, G., Barlow, S., Smith, A., & Caviness, J. (2008). Driving critical initiatives in motor speech. *Journal of Medical Speech Language Pathology, 16*(4), 283–294.

Yorkston, K., Beukelman, D. R., & Tice, R. (1996). *Sentence intelligibility test* [Measurement instrument]. Lincoln, NE: Tice Technologies.

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.

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., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology, 5,* 55–66.

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.