Published in final edited form as:

Semin Speech Lang. 2019 November ; 40(5): 394–406. doi:10.1055/s-0039-1694997.

A Comparative Study of Auditory-Perceptual Speech Measures for the Early Detection of Mild Speech Impairments

Mili Kuruvilla-Dugdale, Ph.D.¹, Katie Threlkeld, M.S.¹, Mary Salazar, M.S.¹, Gwen Nolan, M.S.¹, Lindsey Heidrick, S.L.P.D.²

¹Department of Speech, Language and Hearing Sciences, University of Missouri, Columbia, Missouri;

²Department of Hearing and Speech, University of Kansas Medical Center, Kansas City, Kansas.

Abstract

Background and Purpose—The aim of the study was to compare transcription-based speech intelligibility and scaled speech severity for the detection of mild speech impairments, by studying these metrics across talkers with Parkinson’s disease (PD), age- and sex-matched older adults, and younger adults. An additional aim was to determine the impact of listener experience on these clinical measures.

Methods—Fifteen speakers from each experimental group were asked to read aloud 11 randomly generated sentences from the Speech Intelligibility Test at their typical speaking rate and loudness. Two groups of four listeners each, stratified as experienced or inexperienced listeners based on their clinical experience, judged the sentence samples. To estimate intelligibility, both listener groups were asked to orthographically transcribe exactly what they heard for each sentence. For severity estimates, the listener groups were asked to rate the sentences for overall quality based on voice, resonance, articulation, and prosody, using a visual analog scale.

Results—Transcription-based intelligibility and scaled severity scores of the PD group differed significantly from those of the older and younger adults. Between-age group differences in intelligibility and scaled severity were not observed. Listener experience had an impact on scaled speech severity, but not speech intelligibility. Between-group differences in speech severity were driven by the inexperienced group and not the experienced listener group.

Implications—Both transcription-based intelligibility and scaled severity estimates appear to be sensitive to relatively mild speech impairments in PD. Obtaining scaled severity is less labor intensive than transcription; therefore, visual analog scaling may be the preferred paradigm for clinical use. However, listener experience and training are important considerations for scaling techniques to be implemented clinically.

Keywords

Parkinson’s disease; dysarthria; speech intelligibility; speech severity; visual analog scale

For the assessment of dysarthria, the use of auditory-perceptual methods is the gold standard,¹ allowing clinicians to estimate metrics such as speech intelligibility, speaking rate, and speech severity for diagnostic, prognostic, and treatment monitoring purposes. A variety of auditory-perceptual methods are used for both clinical and research purposes including forced-choice methods,² orthographic transcription,³ and scaling techniques that involve the use of a visual analog scale⁴ (VAS) or direct magnitude estimation.⁵ Of these listening paradigms, orthographic transcription is the most widely used technique clinically to determine speech intelligibility because it provides an objective estimate of the overall effectiveness of the speech signal in conveying content.⁶ To estimate intelligibility, connected speech samples (i.e., standard reading passage or sentences) or word lists are transcribed word for word by listeners to obtain a percent score. Because intelligibility is a system-level measure, it is known to be relatively insensitive to subtle speech impairments presumably due to compensatory adjustments that are likely both within and across speech subsystems.⁷⁻⁹ Another limitation of transcription-based intelligibility estimates is that subsystem-specific underpinnings of the intelligibility loss are largely unknown.¹⁰ Additionally, orthographic transcription and the subsequent intelligibility calculation can be time consuming.¹¹

Scaling is less labor intensive and has been studied in comparison to traditional intelligibility transcription as a way to obtain more specific information about the speech deficit through subsystem-based severity estimates, which helps localize impairment and provide therapy targets. Specifically, using direct magnitude estimation, VAS, or an equal interval scale, listeners are asked to judge speech impairment severity based on certain salient perceptual features, for example, imprecise articulation¹² and hypernasality.¹³ Although scaled speech severity is assumed to have poorer listener reliability than transcription,^{14,15} it is known to provide unique information about subsystem contributions to speech loss if few, selective features are used to rate samples. Scaled severity may also be considered as a more global construct if ratings are based on several features spanning multiple subsystems. Either way, scaled severity estimates are likely to provide subsystem-specific impressions of speech loss that are otherwise unobtainable from traditional intelligibility transcription.⁴

Studies comparing transcription and scaling paradigms to assess dysarthric speech have produced conflicting findings, with studies that found no significant differences between transcription-based intelligibility estimates and scaled estimates of intelligibility,¹¹ as well as no differences between scaled estimates of intelligibility and scaled speech severity.¹² Contrastingly, other studies show that scaled severity estimates are sensitive to aspects of dysarthria not captured by transcription.^{4,12} In support of scaled speech severity, Sussman and Tjaden⁴ observed higher severity ratings for their Parkinson's disease (PD) and multiple sclerosis (MS) groups relative to controls despite comparable word and sentence intelligibility scores for the three groups. Similarly, Weismer et al¹² found scaled estimates of severity based on voice quality and prosody to differentiate PD from amyotrophic lateral sclerosis (ALS) even though both groups had similar word intelligibility scores. Existing comparative studies have several methodological differences including different scaling paradigms, perceptual dimensions, speech stimuli, and dysarthria severities resulting in little agreement on the clinical value of measures derived from different auditory-perceptual paradigms. Moreover, very few comparative studies have focused specifically on the early

detection of dysarthria.⁴ Therefore, the aim of the current study was to expand this line of research by comparing scaled speech severity to transcription-based intelligibility metrics, to determine the diagnostic capability of these measures for the early identification of mild speech impairments in PD and the elderly. Early identification is crucial as dysarthria can be the first and/or the most pronounced symptom of PD.¹⁰ Because speech decline in PD is often overlaid on aging-related subsystem changes that can independently affect speech, it is important to discern both healthy aging and disease processes.^{10,16} Use of less time consuming paradigms to obtain sensitive clinical measures for diagnosis and tracking will allow clinicians to provide speech interventions in a timely manner and will also allow them to better monitor the benefits of behavioral and/or medical treatments.^{17–19}

IMPACT OF LISTENER EXPERIENCE ON AUDITORY-PERCEPTUAL JUDGMENTS

Listener experience is known to have a significant impact on auditory-perceptual methods,²⁰ especially methods that involve subsystem-specific perceptual judgments. For example, Laczi et al¹³ found that naive listeners judged nasality more harshly than speech-language pathologists with over 10 years of clinical experience; contrastingly, Kreiman et al²¹ found that naive listeners judged voice disorders less harshly than experts. On the one hand, it has been argued that experience and training give listeners access to a wide range of perceptual anchors for various dimensions (e.g., voice quality) and allows them to implement flexible perceptual strategies compared with inexperienced listeners who have less exposure to perceptual patterns and tend to use inflexible strategies.²² On the other hand, it has been suggested that experience reduces the impact of deviant perceptual dimensions because of which they are considered less aberrant by experienced listeners.¹³ Not all studies to date, however, show experience-related differences in perceptual judgments.^{1,4,23} Sussman and Tjaden⁴ found that naive and experienced listener groups had comparable scaled estimates of severity for PD, MS, and control groups. Similar results were reported by Bunton and colleagues¹ who used an experienced and inexperienced listener group to perform multidimensional scaling of dysarthric speech samples. Of note, gross listener disagreements were observed for 17.1% of the trials in their study, but these disagreements were not attributed to any one experience group.

In terms of listener reliability, there are mixed reports across auditory-perceptual techniques. For instance, transcription is considered to yield better listener reliability than scaling by some researchers,^{14,24,25} but not others.¹¹ In one study, relatively inexperienced listeners had high intrarater reliability, but only satisfactory interrater reliability when asked to identify deviant perceptual dimensions for dysarthriatype classification.²⁶ On closer examination, interrater reliability was found to be higher for certain dimensions like imprecise consonants and fast rate than other dimensions. When comparing speech intelligibility and scaled speech severity, Stipancic and colleagues¹¹ found moderate intrarater reliability and good interrater reliability for both metrics obtained from inexperienced listeners. Marginally higher reliabilities were observed for scaled estimates of severity than transcription-based intelligibility in their study. Overall, the impact of listener experience appears to vary

depending on the clinical diagnosis (voice disorders, dysarthria), speech stimuli (sentence, paragraph), and listening paradigm (transcription, VAS) used.

STUDY AIMS AND HYPOTHESES

The primary goal of the study was to compare transcription-based speech intelligibility and scaled speech severity for the detection of mild speech impairments in PD and in the elderly, by studying these metrics across talkers with PD, age- and sex-matched older adults, and younger adults. Based on previous research,⁴ scaled speech severity was expected to capture subtle subsystem-specific impairments not inherent to orthographic transcription in talkers with PD relative to healthy controls. In addition, between-age group differences were predicted for scaled speech severity, but not speech intelligibility. An additional aim of the study was to examine the impact of listener experience on traditional sentence intelligibility and scaled speech severity estimates. Listeners with no clinical experience in motor speech disorders were expected to differ significantly in their speech severity ratings, but not in their intelligibility judgments, compared with those with clinical experience.

METHODS

Participants

SPEAKERS—Speakers included 15 healthy young adults (5 men, 10 women), 15 healthy older adults (9 men, 6 women), and 15 individuals with a medical diagnosis of PD (9 men, 6 women). Sex distribution between and within groups was not controlled because the aim of our study was to determine overall between-group differences rather than sex-specific differences for the three groups. All study participants met the following inclusionary criteria: (1) have no prior history of speech, language, cognitive, and hearing impairments per self-report; (2) be a monolingual, American English speaker; and (3) have no hearing aids or a prescription for hearing aids. For participants with PD, additional inclusionary criteria included a negative history of neurosurgical treatments such as deep brain stimulation and absence of comorbid neurological disorders such as ALS and/or dementia. Participants with PD who were taking anti-Parkinson's medications were scheduled for data collection approximately 1 to 2 hours after taking their medication (see Table 1 for demographic information).

All study participants completed a cognitive screening test, that is, either the Montreal Cognitive Assessment²⁷ (MoCA) or Mini-Mental Status Examination²⁸ (MMSE). The MMSE and MoCA are both cognitive screening tools that assess similar domains such as memory, language, attention, and orientation. The maximum score on the MoCA is 30, with a score of 25 or less indicating a possible cognitive impairment.²⁷ The maximum score on the MMSE is 30, with a score of 24 or less indicating a possible cognitive impairment.²⁸ Average cognitive scores for the experimental groups are listed in Table 1.

Although the hearing screening results were not used as an inclusionary criterion, all participants completed a pure tone hearing screening at 500, 1,000, 2,000, and 4,000 Hz. Majority of the PD and older control participants were able to detect pure tones at 45 dB HL at the tested frequencies in at least one ear. For all the younger adults and seven older adults,

no formal hearing screening was completed due to technical problems. However, these participants were able to follow instructions and conversation at normal loudness levels with no signs of hearing problems. In addition, none of these individuals self-reported difficulties with hearing.

LISTENERS—Two listener groups, that is, four females with no clinical experience and four females with clinical experience in motor speech disorders, participated in the current study. The inexperienced group comprised four undergraduate students in their junior year, majoring in communication science and disorders, without any clinical experience in motor speech disorders, but with one semester of research experience in a motor speech laboratory. The experienced group comprised speech-language pathologists with 3 to 5 years of daily clinical experience in motor speech disorders. All listeners passed a pure tone hearing screening at 25 dB HL at 500 Hz, 1 kHz, 2 kHz, and 4 kHz bilaterally. All listeners scored within normal limits on the MoCA (see Table 1) and had no history of speech, language, and/or cognitive impairments per self-report.

Data Collection and Analysis

ACQUISITION OF SPEECH SAMPLES—To compute sentence intelligibility and speech severity, speakers were asked to read aloud 11 sentences from the Speech Intelligibility Test³ (SIT) that varied in length from 5 to 15 words, at a typical speaking rate and loudness level. The sentences were randomly generated by the SIT software such that the test sentences varied across speakers to minimize listener's familiarity and were semantically unpredictable, to limit the contextual information available to listeners for intelligibility judgments. Audio recordings of the sentences produced by each participant were made with a high-quality condenser microphone (Shure, Model PG42, Niles, IL), placed 20 cm from each speaker's mouth, connected to a solid state recorder (Marantz, Model PMD670, Eindhoven, the Netherlands).

LISTENER JUDGMENTS OF SPEECH INTELLIGIBILITY AND SPEECH SEVERITY—The SIT sentences were used for both intelligibility and severity ratings so that these metrics could be compared directly for the same stimuli. Listeners were blind to the participant groups and listened to the speech samples in a randomized order via noise-cancelling headphones (Panasonic, RP-HC200-K, Osaka, Japan) in a quiet laboratory setting. Listeners were allowed to listen to each sentence no more than two times. For intelligibility, listeners were asked to transcribe exactly what they heard. The sentence transcription task took approximately 2.5 hours and was self-paced. Speech intelligibility was calculated as the percentage of words correctly transcribed out of the total number of words attempted. For intelligibility estimates, misspellings, homonyms, and capitalization errors were accepted as correct.

For speech severity ratings, we followed the method described by Sussman and Tjaden,⁴ where a computerized VAS was used to estimate speech severity based on overall quality related to voice quality, articulation, resonance, and prosody. Specifically, listeners were asked to rate each SIT sentence for voice quality (breathy, noisy, gurgly, high/low pitch, or OK); resonance (too nasal, not nasal enough, or OK); articulatory precision (some sounds

are slurred, most sounds are slurred, or OK); and speech rhythm (timing of speech is too slow/fast, or OK). After listening to each sample carefully for these perceptual features, listeners were asked to give a single severity score ranging from 0 to 100 with 0 indicating no impairment and 100 indicating severe impairment. Essentially, a single severity score was provided for each SIT sentence to represent an aggregate of the perceptions of voice quality, resonance, articulatory precision, and speech rhythm. Listeners were specifically instructed to rate each sentence based on overall speech quality and not intelligibility. Rating severity using VAS took approximately 1.5 hours, but was not self-paced in that the program automatically advanced to the next sample as soon as a score was provided until all the samples were rated.

Before commencing the severity rating task, listeners were provided with 48 training samples of healthy and disordered speech that represented a wide range of severities (i.e., typical to severe) of the deviant perceptual features pertinent to the current study such as imprecise articulation, low pitch, etc. Severity anchors were not provided, but listeners were informed that the training samples were representative of samples included in the experimental task. During training, listeners were also familiarized with the VAS rating procedure and asked to provide a single severity score that captures voice quality, resonance, articulatory precision, and speech rhythm for each sentence. None of the sentence samples from the experimental task were used for training, instead sentences collected as part of other studies were used. Training took approximately 45 to 60 minutes to complete and was paced exactly as described earlier for the experimental VAS task. Speech severity ratings were converted to percent scores by the same custom-written MATLAB script that was used for visual analog scaling.²⁹ The percent scores for both intelligibility and severity were averaged across experimental and listener experience groups.

LISTENER RELIABILITY—To estimate interrater reliability, intelligibility and severity scores from 100% of the samples were compared between the four listeners within each experience category. To determine intrarater reliability, 20% of the speech samples were randomly selected for rerating. However, because of work-related time constraints, particularly among clinicians in the experienced group, only one listener from each experience group completed intrarater reliability for both speech intelligibility and severity.

Statistical Analysis

To compare transcription-based speech intelligibility and scaled speech severity scores, a multivariate ANCOVA was conducted with experimental group (younger adults, older adults, PD) and listener experience (inexperienced, experienced) as the fixed factors and sex as the covariate. Statistical results for speech intelligibility and scaled speech severity are listed in Table 2 along with the effect size, partial eta-squared (η_p^2), where 0.01 may be interpreted as small, 0.06 as medium, and 0.14 as large.³⁰ If interactions were significant, pairwise comparisons were completed, and the family-wise Type I error rate was adjusted to 0.05 for multiple comparisons using the Bonferroni correction.

RESULTS

Experimental Group

For both speech intelligibility and speech severity, post hoc comparisons revealed that compared with older and younger adults, PD participants received significantly lower intelligibility scores ($p = 0.000$, $\eta_p^2 = 0.20$) and higher severity ratings ($p = 0.000$, $\eta_p^2 = 0.31$). Speech intelligibility and severity scores were not significantly different between older and younger adults (Fig. 1).

Listener Experience

For scaled severity, post hoc analysis revealed that severity ratings increased as listener experience decreased. Specifically, inexperienced listeners gave significantly higher severity ratings compared with the experienced group ($p = 0.000$, $\eta_p^2 = 0.17$). In contrast, for speech intelligibility, no significant differences were observed across the two listener groups (Fig. 1).

Experimental Group $\times$ Listener Experience

For the inexperienced group, there were significant differences in scaled severity scores between the PD group and older adults ($p = 0.000$, $\eta_p^2 = 0.34$) and the PD group and younger adults ($p = 0.04$, $\eta_p^2 = 0.34$). For the experienced group, there were no significant differences in scaled severity estimates across the three experimental groups. Scaled severity scores given to the PD group differed significantly between the inexperienced and experienced groups ($p = 0.002$, $\eta_p^2 = 0.27$), with PD participants receiving higher scores from inexperienced listeners relative to experienced listeners. There were no significant differences in speech severity scores given by inexperienced and experienced listeners for older and younger adults (Fig. 2).

Listener Reliability

Interrater reliability was calculated with intraclass correlation coefficient (ICC) estimates and their 95% confidence intervals, based on a single rating, absolute agreement, two-way mixed-effects model from 100% of the speech intelligibility and severity samples (Table 3). Intrarater reliability was calculated with ICC estimates and their 95% confidence intervals, based on a single rating, absolute agreement, two-way mixed-effects model from 20% of the speech intelligibility and severity samples (Table 3). The ICC coefficients may be interpreted as poor for values below 0.50; moderate for values between 0.50 and 0.75; good for values between 0.75 and 0.90; and excellent for values above 0.90.³¹

DISCUSSION

Diagnostic Capability of Transcription-Based Intelligibility and Scaled Speech Severity Metrics for Early Detection of Mild Speech Impairments

In contrast to our hypothesis about the diagnostic value of speech severity over intelligibility, our results suggest that both metrics are sensitive to mild speech decline in PD, but not in the elderly. In general, our finding on disease-related differences conflicts with that reported by Sussman and Tjaden⁴ who found speech severity, but not sentence or word intelligibility, to

be sensitive to speech decline in PD and MS. Interestingly, the average sentence intelligibility score for our PD group (i.e., 97%) is comparable to the score reported for their PD group (i.e., 96%), yet between-group differences in transcription-based intelligibility did not reach significance in their study. A closer examination of their group results stratified by sex, however, shows that sentence intelligibility scores were in fact significantly different between female speakers with PD and both male and female control speakers. For scaled speech severity, unlike intelligibility, they found significant differences between each participant group stratified by sex and each control group separated by sex, but females with PD received the highest severity rating. Taken together, our between-group difference in intelligibility is consistent with that observed between their PD female and control groups. Furthermore, our scaled severity results are consistent with those observed between their male and female PD and control groups. Sex-based differences are also likely in the current study and as suggested by Sussman and Tjaden,⁴ female PD speakers may drive the main effect for group. However, the current study does not have adequate power to investigate this and a much larger sample size with groups balanced for sex is necessary to adequately account for sex.

When considering the listening paradigms used in the current study, it is possible that both paradigms were tapping into similar global impressions of speech ability rather than subsystem-specific attributes as expected for speech severity. In that sense, it is reasonable that both speech intelligibility and speech severity yielded similar results. Future comparative studies that use scaled speech severity should consider fewer, salient perceptual features for their clinical population of interest, to obtain a better estimate of severity. Such studies are expected to produce different results from those of the current study, but could highlight the clinical value of scaled severity over speech intelligibility in obtaining subsystem-specific underpinnings of speech loss and in detecting preclinical speech impairments, for example, those associated with healthy aging.

Impact of Listener Experience on Auditory-Perceptual Judgments

As predicted, the inexperienced listeners differed significantly in their speech severity ratings, but not speech intelligibility judgments, compared with the experienced listeners. With regard to severity, our results are consistent with findings from Laczi et al¹³ who found that when judging nasality, naive listeners judged nasality more harshly than expert listeners. It should be noted, however, that Laczi and colleagues¹³ used a single perceptual measure (i.e., nasality) for severity ratings, which is in contrast to the current study where several perceptual features spanning multiple speech subsystems were used. Although a similar listening paradigm was used by Sussman and Tjaden⁴ for scaled speech severity, they found that experience did not affect severity ratings. Of note, they utilized samples from the Grandfather Passage³² for severity ratings instead of SIT sentences, which might be one reason for the divergent findings. In the current study, it is not surprising that scaled severity, but not intelligibility, was affected by listener experience because rating severity based on specific deviant perceptual features requires a certain amount of experience or exposure to motor speech disorders that is not required for intelligibility. Intelligibility is a more concrete construct based on how much the listener understands the spoken message by taking into account only the superficial aspects of speech.^{25,33}

Another finding from the current study related to listener experience was that the scaled severity results were driven by the inexperienced group and not the experienced group. This result was surprising considering the importance of training and/or experience to rate severity based on subsystem features. However, as suggested by Laczi et al,¹³ it is possible that experience reduces the impact that deviant perceptual characteristic has on the listener because of which dysarthric features are perceived as being less problematic by experienced listeners. Another possible explanation is that the type of experience has an influence on listener perception. In the current study, the experienced group had general experience in motor speech disorders, that is, not specific to PD or progressive dysarthrias, whereas the inexperienced group had at least one semester of research experience that was exclusively focused on PD and ALS speech, which may have affected their ratings. To reduce the variability in scaled estimates from listeners with varying levels of experience, it is important to provide specific training exemplars for the different perceptual features that are being assessed so listeners can replace their own unstable internal referents with stable external standards.²¹ Compared with prior studies that report at least 5 hours of training prior to study initiation, the training for the present study may have been limited.^{1,23} Future studies should critically consider both the quantity and specificity of training for obtaining dimensional scaling-based metrics that are reliable.

Listener Reliability

There was moderate interrater agreement, but good intrarater agreement for both listener experience groups and both metrics. Additionally, interrater reliability was slightly higher for speech severity than intelligibility, whereas intrarater reliability was higher for speech intelligibility compared with speech severity. In general, experienced listeners were noted to have higher reliabilities than inexperienced listeners. As such, listener reliabilities for the present study are comparable to values reported elsewhere.^{11,34,35} However, the reliability results from the present study should be interpreted with caution because listener agreement was examined only as a whole across experimental groups and not for each experimental group separately.

In sum, scaled estimates of speech severity shows promise for clinical use because they appear to be sensitive to early speech decline, are less time-consuming to estimate, and have comparable listener reliabilities to transcription-based intelligibility.

Limitations

Certain limitations need to be considered when interpreting the results from the present study such as the small sample size of the speaker and listener groups as well as the disproportionate sex representation. Because sex-related differences have been documented in prior perceptual studies,⁴ it is an important biological variable for disaggregating group data. Another limitation is that multitalker babble (i.e., competing background noise representative of everyday environmental noise encountered by listeners³⁶) was not used for intelligibility and severity judgments as a way to reduce ceiling effects, particularly for the older and younger adult groups. This speech-in-background noise rating procedure is typically followed to simulate communication issues in noisy environments when healthy speech is studied³⁷ and is being explored as an option for dysarthric speech.^{11,35} In terms of

speech stimuli, ratings based on longer, more complex utterances are likely to yield a more representative index of communication ability.^{6,38} In that sense, using a monologue, paragraph, or narrative sample may be better suited for assessing dysarthric speech. Future studies should consider comparing similar metrics across different listening paradigms, for example, comparing transcription-based intelligibility to scaled speech intelligibility. Lastly, specific training exemplars for the perceptual features of interest that represent a wide range of severities should be used so that listeners with varying levels of clinical experience can have comparable internal anchors to rate speech samples.

CONCLUSION

Although both transcription-based speech intelligibility and scaled speech severity metrics were sensitive to early speech changes in PD, obtaining a scaled estimate of severity is less labor intensive and, therefore, better suited for clinical use. However, as seen from the results of the current study, listener experience can significantly affect speech severity ratings. Therefore, when using a subsystem-based scaling paradigm clinically, it is important to provide targeted training so that listeners can have uniform, stable standards against which to base their ratings. With the added training component, it is possible that scaling techniques will be just as labor intensive as traditional transcription-based intelligibility paradigms.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health (R15 DC016383; PI: Kuruvilla-Dugdale). The authors would like to acknowledge the contributions of research assistants Kristina Adler, Kathryn Bauman, Jerrica Horn, Megan Kane, Joanna Kolker, Molly McDermott, Lyndsey Rector, Maggie Schneider, and Dylan Thompson for their help with data collection and analysis. We are particularly grateful to all those who participated in the study.

DISCLOSURES

M.K-D. has funding from the National Institute on Deafness and Other Communication Disorders as well as the Missouri Spinal Cord Injury/Disease Research Program. G.N. was awarded Phase II and Phase III funding from the Parkinson Voice Project as part of their National SPEAK OUT! and LOUD Crowd Grant Program.

K.T., M.S., and L.H. have no financial relationships to disclose. None of the authors have any nonfinancial relationships to disclose.

REFERENCES

1. Bunton K, Kent RD, Duffy JR, Rosenbek JC, Kent JF. Listener agreement for auditory-perceptual ratings of dysarthria. *J Speech Lang Hear Res* 2007;50(06):1481–1495 [PubMed: 18055769]
2. Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord* 1989;54(04):482–499 [PubMed: 2811329]
3. Yorkston K, Beukelman D, Hakel M, Dorsey M. Speech Intelligibility Test [Computer Software]. Lincoln, NE: Madonna Rehabilitation Hospitals; 2007
4. Sussman JE, Tjaden K. Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: intelligibility and beyond. *J Speech Lang Hear Res* 2012;55(04):1208–1219 [PubMed: 22232396]
5. Weismer G, Laures JS. Direct magnitude estimates of speech intelligibility in dysarthria: effects of a chosen standard. *J Speech Lang Hear Res* 2002;45(03):421–433 [PubMed: 12068996]
6. Weismer G Speech intelligibility In: Ball MJ, Perkins M, Müller N, eds. *The Handbook of Clinical Linguistics*. Oxford, UK: Blackwell Publishing; 2008:568–582

7. DePaul R, Brooks BR. Multiple orofacial indices in amyotrophic lateral sclerosis. *J Speech Hear Res* 1993;36(06):1158–1167 [PubMed: 8114482]
8. Green JR, Yunusova Y, Kuruville MS, et al. Bulbar and speech motor assessment in ALS: challenges and future directions. *Amyotroph Lateral Scler Frontotemporal Degener* 2013;14(7–8): 494–500 [PubMed: 23898888]
9. Yunusova Y, Green JR, Lindstrom MJ, Ball LJ, Pattee GL, Zinman L. Kinematics of disease progression in bulbar ALS. *J Commun Disord* 2010;43(01):6–20 [PubMed: 19683250]
10. Duffy JR. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. 3rd ed. St. Louis, MO: Elsevier Mosby; 2013
11. Stipancic KL, Tjaden K, Wilding G. Comparison of intelligibility measures for adults with Parkinson's disease, adults with Multiple Sclerosis, and healthy controls. *J Speech Lang Hear Res* 2016;59 (02):230–238 [PubMed: 26556727]
12. Weismer G, Jeng JY, Laures JS, Kent RD, Kent JF. Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatr Logop* 2001;53(01):1–18 [PubMed: 11125256]
13. Laczi E, Sussman JE, Stathopoulos ET, Huber J. Perceptual evaluation of hypernasality compared to HONC measures: the role of experience. *Cleft Palate Craniofac J* 2005;42(02):202–211 [PubMed: 15748113]
14. Miller N Measuring up to speech intelligibility. *Int J Lang Commun Disord* 2013;48(06):601–612 [PubMed: 24119170]
15. Schiavetti N Scaling procedures for the measurement of speech intelligibility In: Kent RD, ed. *Intelligibility in Speech Disorders*. Madison, WI: John Benjamins Publishing Company; 1992:11–34
16. Dromey C, Boyce K, Channell R. Effects of age and syntactic complexity on speech motor performance. *J Speech Lang Hear Res* 2014;57(06): 2142–2151 [PubMed: 25215529]
17. Bang YI, Min K, Sohn YH, Cho SR. Acoustic characteristics of vowel sounds in patients with Parkinson disease. *NeuroRehabilitation* 2013;32(03):649–654 [PubMed: 23648619]
18. Ho AK, Bradshaw JL, Iansek R. For better or worse: the effect of levodopa on speech in Parkinson's disease. *Mov Disord* 2008;23(04):574–580 [PubMed: 18163453]
19. Sapir S Multiple factors are involved in the dysarthria associated with Parkinson's disease: a review with implications for clinical practice and research. *J Speech Lang Hear Res* 2014;57(04): 1330–1343 [PubMed: 24686571]
20. Kent RD. Hearing and believing. *Am J Speech Lang Pathol* 1996;5(03):7–23
21. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res* 1993;36(01):21–40 [PubMed: 8450660]
22. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech Hear Res* 1990;33(01):103–115 [PubMed: 2314068]
23. Zyski BJ, Weisiger BE. Identification of dysarthria types based on perceptual analysis. *J Commun Disord* 1987;20(05):367–378 [PubMed: 3655012]
24. Hustad KC. A closer look at transcription intelligibility for speakers with dysarthria: evaluation of scoring paradigms and linguistic errors made by listeners. *Am J Speech Lang Pathol* 2006;15(03): 268–277 [PubMed: 16896176]
25. McHenry M An exploration of listener variability in intelligibility judgments. *Am J Speech Lang Pathol* 2011;20(02):119–123 [PubMed: 21317298]
26. Zeplin J, Kent RD. Reliability of auditory perceptual scaling of dysarthria In: Robin DA, Yorkston KM, Beukelman DR, eds. *Disorders of Motor Speech: Assessment, Treatment, and Clinical Characterization*. Baltimore, MD: Paul H. Brookes Publishing Company; 1996:145–154
27. Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005;53(04):695–699 [PubMed: 15817019]
28. Folstein MF, Folstein SE, McHugh PR. "Minimental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975;12(03):189–198 [PubMed: 1202204]

29. Stepp CE, Heaton JT, Rolland RG, Hillman RE. Neck and face surface electromyography for prosthetic voice control after total laryngectomy. *IEEE Trans Neural Syst Rehabil Eng* 2009;17(02): 146–155 [PubMed: 19304494]
30. Cohen J Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, MO: Lawrence Erlbaum Associates; 1988
31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15(02):155–163 [PubMed: 27330520]
32. Riper CV. Speech Correction. Englewood Cliffs, NJ: Prentice-Hall; 1963
33. Hustad KC. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *J Speech Lang Hear Res* 2008; 51(03):562–573 [PubMed: 18506035]
34. Tjaden K, Kain A, Lam J. Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with Parkinson's disease. *J Speech Lang Hear Res* 2014;57(04): 1191–1205 [PubMed: 24686409]
35. Tjaden K, Sussman JE, Wilding GE. Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *J Speech Lang Hear Res* 2014; 57(03):779–792 [PubMed: 24687042]
36. Plomp R Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am* 1978;63(2, S2):533–549 [PubMed: 670550]
37. Fontan L, Tardieu J, Gaillard P, Woisard V, Ruiz R. Relationship between speech intelligibility and speech comprehension in babble noise. *J Speech Lang Hear Res* 2015;58(03):977–986 [PubMed: 25809922]
38. Bunton K, Keintz CK. The use of a dual-task paradigm for assessing speech intelligibility in clients with Parkinson disease. *J Med Speech Lang Pathol* 2008;16(03):141–155 [PubMed: 21637738]

Learning Outcomes:

As a result of this activity, the reader will be able to (1) compare and contrast two listening paradigms that are used to assess dysarthria, namely, orthographic transcription and visual analog scaling; (2) evaluate the diagnostic capability of transcription-based speech intelligibility and scaled estimates of speech severity for the early detection of mild speech impairments in PD and in the elderly; (3) discuss how listener experience impacts speech intelligibility and speech severity estimates obtained through orthographic transcription and visual analog scaling, respectively.

Figure 1. Mean ($\pm$ SE) of transcription-based intelligibility and scaled speech severity scores for each experimental and listener experience group. PD, Parkinson's disease.

Figure 2.

Pairwise comparisons for the significant experimental group $\times$ listener experience interaction for scaled speech severity. PD, Parkinson's disease.

Table 1

Demographic and Screening Information for Speakers and Listeners

Participants	N	Sex (M, F)	Age Mean (SD)	MoCA/MMSE Mean (SD)
Speakers				
Younger adults	15	9 M, 6 F	24.21 (4.25)	28.29 (1.14)
Older adults	15	9 M, 6 F	66.93 (7.52)	26.75 (2.60)
Parkinson's disease	15	5 M, 10 F	68.07 (6.96)	27.79 (1.63)
Listeners				
Inexperienced	4	4 F	21.25 (0.05)	28.25 (1.26)
Experienced	4	4 F	28.00 (2.16)	29.5 (0.58)

Abbreviations: F, female; M, male; MMSE, Mini-Mental Examination; MoCA, Montreal Cognitive Assessment; SD, standard deviation.

Table 2
Multivariate ANCOVA for Transcription-based Sentence Intelligibility and Scaled Speech Severity

Source	Speech intelligibility			Speech severity		
	df	F	η_p^2	p	F	η_p^2 p
Experimental group ^a	2	10.35	0.20	0.000	18.63	0.31 0.000
Listener experience ^b	1	1.70	0.02	0.196	17.19	0.17 0.000
Experimental group × listener experience	2	0.83	0.02	0.441	5.08	0.11 0.008

Abbreviations: df, degrees of freedom; F, F-statistic; η_p^2 , partial eta-squared; p, significance level; PD, Parkinson's disease.

^a n = 15 for the younger adult group; n = 15 for the older group; n = 15 for the PD group.

^b n = 4 for inexperienced listeners; n = 4 for experienced listeners.

Interclass Correlation Coefficients for Interrater and Intrarater Reliability for Speech Intelligibility and Speech Severity

Table 3

Listener experience	Speech intelligibility		Speech severity	
	ICC (95% CI)	p-Value	ICC (95% CI)	p-Value
Interrater reliability ^a				
Inexperienced	0.509 (0.363–0.653)	0.000	0.642 (0.408–0.793)	0.000
Experienced	0.624 (0.492–0.745)	0.000	0.697 (0.528–0.817)	0.000
Intrarater reliability ^b				
Inexperienced	0.833 (0.452–0.951)	0.002	0.809 (0.653–0.899)	0.000
Experienced	0.910 (0.786–0.963)	0.000	0.825 (0.678–0.909)	0.000

Abbreviations: CI, confidence intervals; ICC, interclass correlation coefficients; *p*, significance level.

Note: ICC coefficients below 0.50: poor; 0.50–0.75: moderate; 0.75–0.90: good; above 0.90: excellent.

^a *n* = 4 inexperienced and 4 experienced listeners for speech intelligibility and speech severity.

^b *n* = 1 inexperienced and 1 experienced listener for speech intelligibility and speech severity.