ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

Preliminary Psychometric Review of Neurologists' Speech Ratings on the Unified Parkinson's Disease Rating Scale

Kimberly L. Dahl,[a] Magdalen A. Balz,[a,b] Kara M. Smith,[c] and Cara E. Stepp[a,d,e]

[a] Department of Speech, Language, and Hearing Sciences, Boston University, MA [b] Department of Gerontology, University of Massachusetts, Boston [c] Department of Neurology, University of Massachusetts Chan Medical School, Boston [d] Department of Biomedical Engineering, Boston University, MA [e] Department of Otolaryngology—Head and Neck Surgery, Boston University School of Medicine, MA

ARTICLE INFO

Article History:
Received April 18, 2025
Revision received July 23, 2025
Accepted August 27, 2025

Editor-in-Chief: Rita R. Patel
Editor: Kelly Richardson

https://doi.org/10.1044/2025_AJSLP-25-00168

ABSTRACT

Purpose: The purpose of this study was to conduct a preliminary psychometric evaluation of the speech item of the Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III.
Method: Twenty neurologists specializing in movement disorders evaluated the speech—unscripted and sentence reading—of people with Parkinson's disease (PD) on the MDS-UPDRS Part III. The concurrent validity of the speech item was evaluated by comparing the ratings to a previously collected estimate of the speakers' intelligibility during sentence reading. The reproducibility of the ratings was evaluated in terms of reliability (intra- and interrater) and agreement.
Results: MDS-UPDRS Part III speech ratings based on sentence reading were moderately and negatively correlated with previously collected estimates of intelligibility. No such relationship was found for MDS-UPDRS Part III speech ratings based on unscripted speech. Individual correlations for each rater ranged widely, with some showing weak, negligible, or positive relationships with intelligibility. Intrarater reliability was high, but interrater reliability was fair for both speech tasks. Movement disorders specialists agreed on MDS-UPDRS speech ratings, on average, less than 70% of the time.
Conclusions: A critical communication outcome—intelligibility during unscripted speech—is not captured by MDS-UPDRS Part III speech ratings, suggesting weak concurrent validity of this widely used metric of speech function. The reproducibility of MDS-UPDRS Part III speech ratings across providers is fair. These speech ratings may be insufficient to assess the speech function of people with PD and identify those in need of speech services.

The speech symptoms of people with Parkinson's disease (PD) are heterogeneous, and all subsystems of speech—phonatory, articulatory, and prosodic—may be affected (Ho et al., 1999; Logemann et al., 1978). Pharmacological and surgical management of PD do not meaningfully alleviate speech symptoms and may even cause further deficits (Brabenec et al., 2017). Evidence-based behavioral speech treatments, however, are available (Behrman et al., 2020; Bryans et al., 2021; Ramig et al., 2001). Access to those treatments may depend on a neurologist's referral to a speech-language pathologist (SLP) based on the neurologist's determination that the patient's speech is sufficiently impaired to warrant intervention.

For people with PD, the standard clinical evaluation of speech by neurologists—in particular, movement disorders specialists—follows the Movement Disorder Society's revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III (Goetz et al., 2008; Ramaker et al., 2002). The 18-item MDS-UPDRS Part III motor examination includes a single, global rating of speech impairment. The speech item in the original UPDRS provided no specific instructions to the examiner and offered

Correspondence to Kimberly L. Dahl: dahl@bu.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

vague descriptions of how speech may be impaired at each level of the 5-point scale. This item showed poor to adequate reliability and was targeted for revision due to the scale ambiguity and unclear instructions (Goetz et al., 2007; Martínez-Martín et al., 1994; Richards et al., 1994). In the revised MDS-UPDRS, the examiner is instructed to assess the patient's conversational speech and assign a speech impairment score defined by deficits in "modulation, diction, or volume" and the degree to which these deficits affect intelligibility (Goetz et al., 2008). The validity and reliability of this revised speech item has not been fully evaluated.

The concurrent validity of MDS-UPDRS Part III speech ratings could be established by comparing them to a criterion measure. Such a measure need not be identical to the MDS-UPDRS Part III speech assessment but should be a relevant, reliable, and valid measure of speech function. Because speech symptoms of PD are variable, the criterion should capture impairment in any speech subsystem. Reduced intelligibility is a common concern among people with PD, with a notable impact on communicative function and well-being (Borrie et al., 2022; McAuliffe et al., 2017; Miller et al., 2007). Intelligibility may deteriorate due to impairment in any speech subsystem and with any combination of symptoms. It is a critical speech outcome closely tied to overall speech function and thus a compelling criterion measure. Furthermore, prior research provides useful guides for reliably measuring intelligibility (Abur et al., 2019; Dahl et al., 2024; Hirsch et al., 2022; Stipancic et al., 2016), thus ensuring its suitability as a criterion measure.

A prior study evaluated whether MDS-UPDRS Part III speech ratings relate to the intelligibility of speakers with PD and found a moderate negative correlation (Spencer et al., 2022). A crucial limitation of this work is that a single SLP with substantial expertise evaluating dysarthric speech completed the ratings. The findings of this study thus cannot be generalized to typical clinical use of the MDS-UPDRS, that is, by movement disorders specialists without specific, comprehensive training in speech assessment.

The clinical utility of a measure also depends on its reproducibility. That is, its repeated use by the same rater and across raters should yield similar results. Reproducibility can be assessed as reliability or agreement—two related but distinct psychometric features that are important for discriminative and evaluative purposes, respectively (de Vet et al., 2006). As noted above, speech ratings on the original UPDRS showed poor to adequate reliability (Martínez-Martín et al., 1994; Richards et al., 1994), but neither reliability nor agreement has been assessed for the revised MDS-UPDRS Part III speech item.

The purpose of this study was therefore to conduct a preliminary psychometric evaluation of the speech item of the MDS-UPDR Part III as rated by movement disorders specialists, focusing specifically on the item's concurrent validity and reproducibility. Neurologists who specialize in movement disorders evaluated speech samples of people with PD using the MDS-UPDRS. These speech ratings were compared to intelligibility scores to assess concurrent validity. Intelligibility scores were based on a robust estimate of intelligibility for each speaker calculated from transcriptions of sentence reading for a previous study (Dahl et al., 2024). The speech ratings were also evaluated for their reliability—both intrarater and interrater—and agreement. These analyses will help determine the psychometric properties of the MDS-UPDRS Part III speech item, with the broader goal of determining whether these ratings can be reasonably used for important purposes such as monitoring the progression of speech impairment, assessing the effects of treatment on speech function, identifying those in need of speech services, and collecting high-quality data for research.

Method

Raters

Raters were 20 movement disorders specialists (12 cisgender women, eight cisgender men; age: $M = 41.5$ years; $SD = 9.3$ years; range: 30–66 years), all of whom used the MDS-UPDRS regularly in their clinical work. They had an average of 10.8 years of experience evaluating people with PD ($SD = 9.2$ years; range: 1–38 years). They were recruited through the authors' professional networks using convenience and snowball sampling and through outreach to movement disorders clinics in Massachusetts and Rhode Island. All raters provided informed consent per the institutional review board of Boston University (#2625).

Speakers

Speakers were 20 people with PD (10 females, 10 males[1]), originally selected from a database of speech recordings at Boston University for a prior study on speech intelligibility (Dahl et al., 2024). Speakers were, on average, 66.6 years old ($SD = 5.6$ years; range: 54–77 years) and had been diagnosed with PD 1–16 years before their speech was recorded ($M = 7.2$ years; $SD = 3.8$ years). The speaker sample covered a wide range of motor symptom and dysarthria severity, as described in

[1]Only sex is reported, as gender information was not available for all speakers.

the original study (Dahl et al., 2024). All speakers provided informed consent per the institutional review board of Boston University (#2625) or the University of Washington (#36181).

Speech Samples

Two types of speech samples were collected from each speaker—(a) unscripted speech in response to a common conversational prompt (e.g., weather, travel, weekend plans) and (b) a set of 11 sentences randomly generated by the Speech Intelligibility Test (SIT; Yorkston et al., 1996). The speakers were recorded while wearing a microphone positioned 7 cm from the corner of the mouth at a 45° angle from the midline of the body. Recordings were collected in either a sound-treated booth with an omnidirectional condenser earset microphone (model MX153; Shure) and QuadMic II pre-amplifier or in a quiet room with a headset microphone (models WH20, WH20XLR, SM35XLR; Shure) and handheld digital recorder (model LS-10; Olympus). All recordings were sampled at 44.1 kHz with 16-bit resolution. Recordings were normalized to the peak amplitude across the entire study sample with a custom MATLAB script (Version 2022a; MathWorks). This amplitude-normalization eliminated wide fluctuations across recordings that could undermine the audibility of some samples when played at a set listening volume, but it maintained relative differences in amplitude between speakers and speech tasks.

Intelligibility Estimates

Intelligibility can be estimated through various approaches, and there is no consensus on a specific protocol. We drew on prior work to select intelligibility estimates that maximized objectivity and reliability—important features of a criterion measure. Specifically, we noted that (a) more listeners and more utterances improve intelligibility assessment (Abur et al., 2019; Dahl et al., 2024; Xue et al., 2023), (b) expert and inexperienced raters are closely aligned in their assessment of intelligibility (Abur et al., 2019; Dahl et al., 2024; Hirsch et al., 2022), and (c) there is some evidence that intelligibility of unscripted and read speech may be comparable (Tjaden & Wilding, 2011; cf. a case study by Kempler & Lancker, 2002).

With these considerations in mind, we selected intelligibility estimates that represented the most robust and reliable assessment of the intelligibility of each speaker in our sample. These estimates were drawn from a previous study on the accuracy of different approaches to assessing intelligibility (Dahl et al., 2024). Specifically, intelligibility was estimated as the percentage of words in 11 Speech Intelligibility Test (SIT) sentences, mixed with multitalker babble,[2] that were correctly transcribed by 10 listeners enrolled in the prior study. Those listeners were five SLPs with expertise in assessing dysarthric speech and five inexperienced listeners who were not SLPs. The intelligibility scores of speakers in our sample ranged from 1% to 84%, and these percentage-correct intelligibility scores served as the criterion measure for the current study.

Rating Procedure

Remote study sessions were conducted with the 20 movement disorders specialists via Gorilla Experiment Builder (gorilla.sc), an online behavioral experiment platform. These raters used their own computer or tablet and headphones to join an experimenter by videoconference. They did so from a location that was, by the rater's estimation, at least as quiet as the setting in which they would typically evaluate people with PD. The experimenter gave the rater instructions and then remained in the videoconference call throughout the session to answer questions and provide technical support.

Raters first played a brief recording of two people not included in the study to set the volume to a comfortable level, which they maintained throughout the session. Raters were then instructed to listen to each speech sample and rate the speech on the 0–4 scale of the MDS-UPDRS Part III speech item. The definition of each level of the scale was provided on the screen with each trial. Raters were allowed to play each sample as many times as they liked to mimic the flexibility of a clinical evaluation.

Raters evaluated all 20 speakers in two blocks of 20 speech samples each—one of unscripted speech and one of SIT sentences.[3] Ratings of unscripted speech were collected to mimic the typical process of evaluating speech for clinical purposes with the MDS-UPDRS. Ratings of SIT sentences, though not the typical task for the MDS-UPDRS Part III speech item, were collected to allow a more direct comparison with the SIT-based intelligibility outcome.

The order of these blocks of ratings was counterbalanced across raters. The order of speech samples within each block was randomized for each rater. One practice sample from a speaker not included in the study was presented at the start of each block to familiarize the rater with the task. Four randomly selected samples were

[2]Multitalker babble was added at a signal-to-noise ratio of 0 dB to minimize ceiling effects in the original study.

[3]Each speaker's SIT sentences were concatenated into a single recording for the rating procedure.

repeated at the end of each block to measure intrarater reliability. Raters thus evaluated a total of 50 speech samples, which took an average of 27.5 min (SD = 9.9 min; range: 12–47 min).

Statistical Analysis

Statistical analysis was conducted in RStudio (Version 2024.04.2; R Core Team, 2022). Statistical significance, where applicable, was set a priori at $\alpha = .05$ and corrected for multiple comparisons with the Holm–Bonferroni method (Holm, 1979). All statistics were calculated separately for each stimulus type (i.e., unscripted speech and SIT sentences).

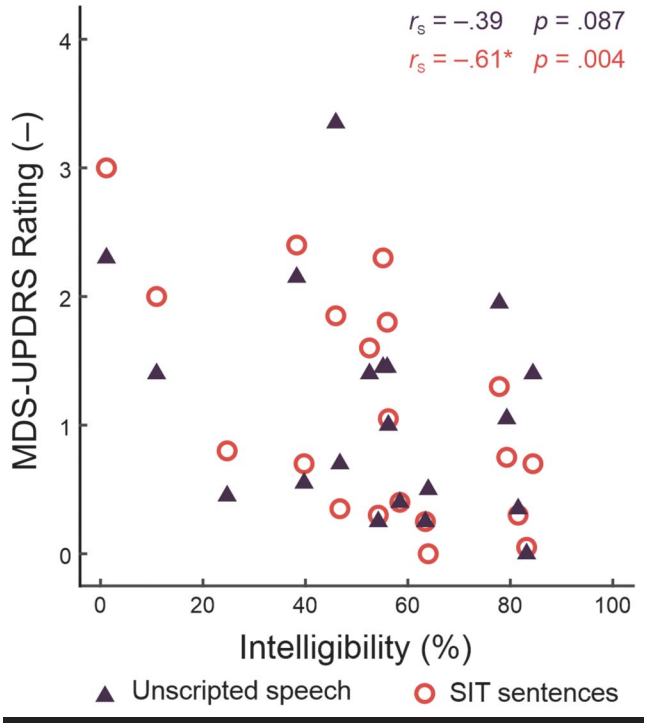
We evaluated the concurrent validity of the MDS-UPDRS Part III speech item by testing for a relationship between mean speech rating and intelligibility score via Spearman's correlation coefficient (R *psych* package; Revelle, 2023). We also evaluated the relationships between speech rating and intelligibility score for each movement disorder specialist via Spearman's correlation coefficients. These individual coefficients served to describe the data, not test any hypotheses, and so were uncorrected for multiple comparisons.

We assessed the reproducibility of MDS-UPDRS speech ratings in terms of both reliability and agreement (de Vet et al., 2006). We calculated intrarater reliability as Cohen's weighted kappa and interrater as Fleiss's kappa (R *irr* package; Gamer et al., 2019), both of which are weighted to account for the importance of distance between ratings on the MDS-UPDRS scale. We calculated agreement for each speaker as the number of ratings by movement disorders specialists that were equal to the mode for that speaker divided by the total number of ratings (i.e., percent agreement). Finally, to provide additional context for these results, we tested for a relationship between percent agreement and intelligibility score as Spearman's correlation coefficients.

Results

MDS-UPDRS Part III speech ratings were descriptively similar for both speech tasks; they covered the full range of the scale (0–4) but were skewed toward slight impairment (M = 1.1, SD = 1.0). At the group level, ratings by movement disorders specialists based on unscripted speech—as in the typical clinical use of the MDS-UPDRS—were not significantly correlated with intelligibility scores (see Figure 1). Ratings based on SIT sentences were significantly, moderately correlated with intelligibility scores, even after correcting for multiple comparisons.

Figure 1. Relationship between mean speech rating on the Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Part III and the intelligibility of people with Parkinson's disease. Ratings were based on unscripted speech (triangles) and Speech Intelligibility Test (SIT) sentences (circles). *statistically significant.

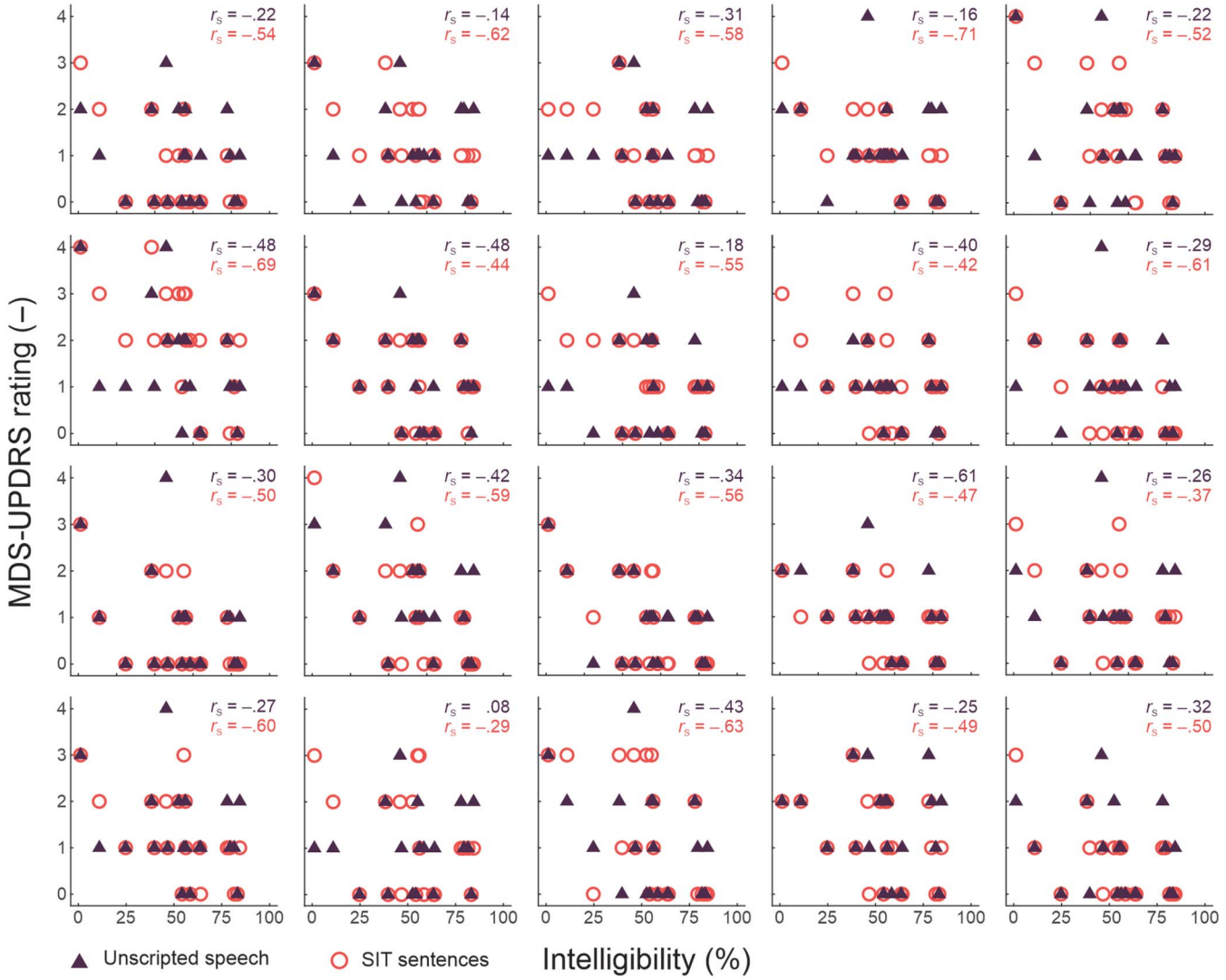


The relationship between MDS-UPDRS Part III speech ratings and intelligibility scores ranged widely for individual raters (see Figure 2). Correlations were negligible to moderate for ratings based on unscripted speech (M = −.30, SD = .15, range: .08 to −.61) and weak to strong for SIT-based ratings (M = −.53, SD = .10, range: −.29 to −.71).

Intrarater reliability of MDS-UPDRS speech ratings, based on the average weighted kappa, was strong for both unscripted speech (M = .76, SD = .25, range: .11–1) and SIT sentences (M = .75, SD = .23, range: .33–1). Most raters were highly reliable ($\kappa \geq .70$) when rating both unscripted speech (n = 14, seven with perfect agreement) and SIT sentences (n = 12, eight with perfect agreement). Two raters had poor intrarater reliability ($\kappa < .40$) with unscripted speech and one with SIT sentences.

Interrater reliability of MDS-UPDRS Part III speech ratings was fair for both unscripted speech (κ = .33) and SIT sentences (κ = .36). Raters agreed on the level of speech impairment for a given speaker, on average, 63% of the time (SD = 15%, range: 35%–100%) when rating unscripted speech and 69% (SD = 13%, range: 50%–100%) when rating SIT sentences (see Figure 3). After correcting for multiple comparisons, there was no

Figure 2. Relationship between the speech ratings of each movement disorders specialist and the intelligibility of people with Parkinson's disease. Ratings were based on unscripted speech (triangles) and SIT sentences (circles). In each panel, the upper Spearman's correlation coefficient is for unscripted speech and lower for Speech Intelligibility Test (SIT) sentences.



significant correlation between percent agreement and intelligibility for unscripted speech nor SIT sentences (see Figure 4).

Discussion

This study evaluated the concurrent validity and reproducibility of the speech item from the motor examination (Part III) of the MDS-UPDRS, offering a preliminary review of the psychometric properties of this global speech metric. The MDS-UPDRS is designed to comprehensively assess impairment and disability in people with PD. It does so through patient or caregiver reports and a motor examination. It is widely used and, overall, psychometrically sound (Goetz et al., 2008; Martinez-Martin et al., 2013; Ramaker et al., 2002). However, despite the high prevalence, heterogeneity, and substantial psychosocial impacts of speech impairment in people with PD (Ho et al., 1999; Logemann et al., 1978; McAuliffe et al., 2017; Swales et al., 2021), the MDS-UPDRS Part III motor exam includes only a single item dedicated to speech. Furthermore, the typical clinical examiner is an expert in neuromotor function without specific training in assessing speech deficits. Despite these limitations, this measure may factor into clinical decisions about the level of speech impairment of people with PD and their need for speech services.

Figure 3. Median and interquartile range (IQR) of percent agreement on Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating ScalePart III speech ratings, based on unscripted speech and Speech Intelligibility Test (SIT) sentences. Whiskers denote range of data points within 1.5 IQR. Circle markers show percent agreement for each speaker.

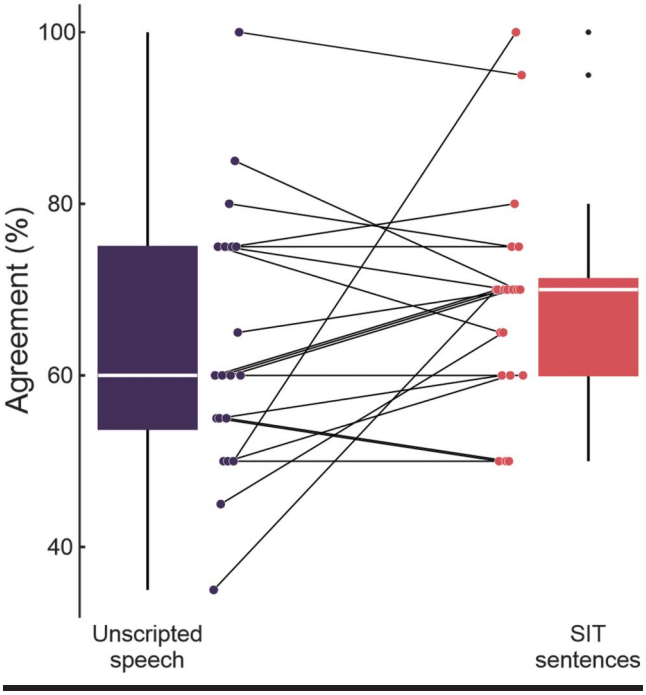
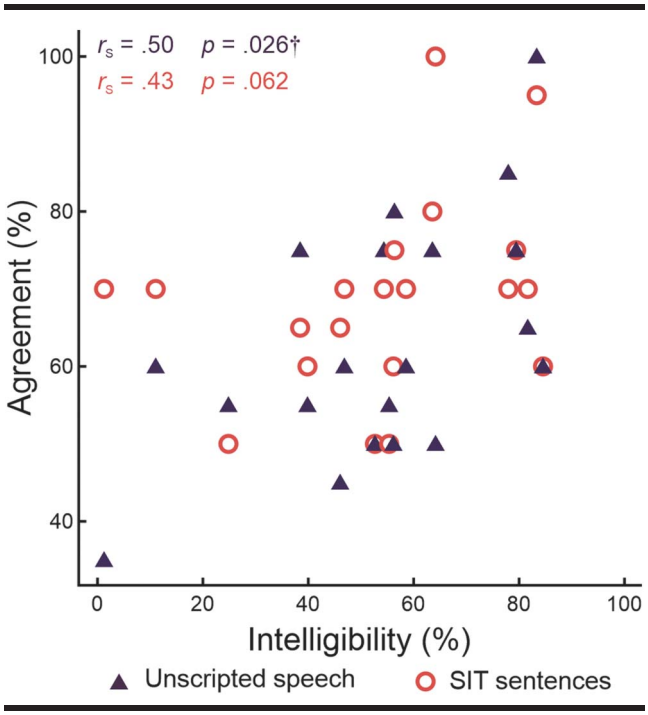


Figure 4. Relationship between the level of agreement by movement disorders specialists on Movement Disorder Society–sponsored revision of the Unified Parkinson's Disease Rating Scale Part III speech ratings and the intelligibility of people with Parkinson's disease. Ratings were based on unscripted speech (triangles) and Speech Intelligibility Test (SIT) sentences (circles). †not statistically significant after correcting for multiple comparisons.



Concurrent Validity

The clinical utility of MDS-UPDRS Part III speech ratings for identifying those in need of a speech referral may depend on whether it closely reflects important indicators of communicative function, such as intelligibility. Speech ratings collected for the present study were compared to this critical speech outcome to evaluate the concurrent validity of the MDS-UPDRS Part III speech item. Prior work found that these speech ratings were moderately correlated with the intelligibility of people with PD when the assessment was done by an expert SLP with extensive training in assessing dysarthric speech (Spencer et al., 2022). In our work, a similar relationship was found for movement disorders specialists' ratings of SIT sentences, but not for the more clinically relevant ratings of unscripted speech, that is, the speech task of the MDS-UPDRS.

The difference in concurrent validity according to speech task reveals a particular challenge of speech assessment. Prescribed speech such as SIT sentences offers greater control over potential confounds, such as linguistic content and phonemic coverage. But unscripted speech is more ecologically valid, offering a closer representation of how people speak in their day-to-day lives. The MDS-UPDRS prioritizes ecological validity by instructing the rater to base their assessment on speech elicited in conversation and thus sacrifices control over confounding variables. This approach is sound for the clinical purpose of approximating a person's speech function outside the clinic. However, the result is a lack of concurrent validity for MDS-UPDRS Part III speech ratings based on unscripted speech. It seems that revising the original UPDRS speech item to anchor each impairment score, in part, to overall intelligibility may not have successfully captured this important construct.

This test of concurrent validity, however, is based on ratings averaged over 20 raters. A person with PD is likely evaluated by a single movement disorders specialist at any given visit. So, we also assessed the concurrent validity of the MDS-UPDRS Part III speech item for each rater individually. For most raters (70%), the relationship between intelligibility and MDS-UPDRS ratings based on unscripted speech was weak at best, with one rater showing an unexpected, albeit negligible, positive relationship (r_s = .08). The degree to which the MDS-UPDRS Part III speech rating reflects intelligibility is thus dependent on the clinician rating the speech. These findings suggest the

need for (a) additional training in unscripted speech assessment for movement disorders specialists and (b) strengthening interprofessional practice between these specialists and SLPs, the latter of whom have undergone rigorous training in such assessments. Additional training and close collaboration with SLPs could enhance the validity of MDS-UPDRS Part III speech ratings, thus making these ratings more useful for monitoring symptom progression, determining treatment effects, making referrals, and conducting rigorous research.

Reproducibility

MDS-UPDRS Part III speech ratings for both speech tasks were highly reproducible at the level of the individual rater (i.e., intrarater reliability). Most raters were consistent in how they rated the speech of a given speaker and in several cases perfectly so. Speech ratings were considerably less reproducible, however, across raters—neither interrater reliability nor agreement were strong. This finding is consistent with earlier psychometric reviews of the original UPDRS speech item (Martínez-Martín et al., 1994; Richards et al., 1994). The revisions incorporated into the MDS-UPDRS—the version assessed here—do not seem to have improved the reproducibility of these speech ratings.

Speech task was also a factor in reproducibility. Movement disorders specialists agreed less when rating unscripted speech than SIT sentences. This was true of average agreement across speakers (63% unscripted vs. 69% SIT) as well as the lowest observed agreement for a given speaker (35% unscripted vs. 50% SIT). Thus, unscripted speech again offers the benefit of ecological validity but at the cost of reproducibility.

The level of disagreement we observed among raters about a person's speech impairment may represent a clinically meaningful difference. Except for a very intelligible speaker (always rated 0, or "normal" speech), speakers' unscripted speech samples were rated one to three levels apart on the five-level MDS-UPDRS scale. There is no published research on what represents a clinically meaningful change in the MDS-UPDRS Part III speech item. However, it is likely that clinical decision making would differ if a patient were rated as having "mild" versus "severe" speech impairment (i.e., two levels apart). Thus, the level of disagreement in our findings suggest that the suboptimal psychometric performance of this rating approach could have a practical impact on people with PD. Furthermore, low interrater reliability and agreement undermines the utility of these speech ratings for determining response to speech interventions if different clinicians assess a person with PD before and after treatment.

The reproducibility of other auditory-perceptual ratings of speech tends to be strongest at the extremes, such that raters agree more closely when rating those with mild and severe symptoms and less so for speakers with moderate speech impairment (Eadie & Kapsner-Smith, 2011). We found no compelling evidence that the same is true of MDS-UPDRS Part III speech ratings. Although there tended to be closer agreement for highly intelligible speakers, that was not the case for those who were least intelligible (see Figure 4). The reproducibility of MDS-UPDRS Part III speech ratings is thus not closely tied to the actual severity of speakers' speech symptoms.

Nonspeech Factors

PD affects all subsystems of speech, and it is specifically speech that is evaluated on the MDS-UPDRS Part III. However, PD also affects language and cognition. Although SLPs are explicitly trained to assess speech impairment separately from cognitive–linguistic function, movement disorders specialists—who have not received the same training in speech assessment—may be influenced by these other nonspeech factors. That is, movement disorders specialists may rate speakers who, for example, use simplified grammar or have word-finding difficulties as having greater speech impairment than speakers who do not show these cognitive–linguistic changes. The degree to which cognitive–linguistic factors influence speech ratings may vary by rater. These individual differences may have contributed to the wide-ranging relationships between MDS-UPDRS Part III speech ratings and intelligibility at the level of the rater (see Figure 2) and the high intrarater but low interrater reliability.

Importantly, people with PD report disruptions to their communication even when intelligibility is preserved (Miller et al., 2011). This disruption may be the result of cognitive–linguistic function, self-perceptions, or other nonspeech factors. Although these factors are outside of the scope of the MDS-UPDRS Part III speech item, they should certainly be considered in a comprehensive assessment of communicative function and in determining who would benefit from speech services.

Clinical Implications

Up to 90% of people with PD develop hypokinetic dysarthria (Duffy, 2019; Logemann et al., 1978), and speech symptoms can appear years before diagnosis (Fereshtehnejad et al., 2019). Even mild speech symptoms can disrupt communicative participation (Schalling et al., 2017). Yet less than 10% of people with PD are referred for speech services (Roberts et al., 2021). We do not know the reasons for this alarming gap between need and access. It is important,

however, for movement disorders specialists to consider this disparity along with our findings—weak concurrent validity and reproducibility of the MDS-UPDRS speech item—when making decisions about speech referrals for people with PD.

Our findings suggest that MDS-UPDRS Part III speech ratings are not closely related to intelligibility, an important indicator of communicative function. They also show that a speaker may be rated on the MDS-UPDRS scale quite differently from one provider to the next. Thus, other measures of communicative function, particularly self- and caregiver reports of changes in communication (Baylor et al., 2024; Mach et al., 2021; Miller et al., 2011), may offer better evidence of impairment and the need for referral. Furthermore, the use of the MDS-UPDRS in clinical trials and other research should be implemented with these limitations in mind, taking care that participants are evaluated by the same rater at each time point.

Limitations

In typical clinical practice, MDS-UPDRS Part III speech ratings are based on conversational speech evaluated face to face. The intelligibility scores that served as our criterion measure were not intended to be identical to the MDS-UPDRS Part III speech ratings but rather another important, reliable, and valid indicator of speech function. However, it is important to note some differences in how these intelligibility estimates were collected. First, they were estimated from audio samples that provided no visual cues, which may play a role in intelligibility. Second, the audio samples were amplitude-normalized in a way that maintained relative differences between speakers and speech tasks but eliminated inaudibility, which is a concern for those with severe dysarthria, as a factor in intelligibility. Finally, intelligibility estimates were based on SIT sentences. Sentence reading offered the most robust estimate of intelligibility (Dahl et al., 2024) and allowed for rigorous control of potential confounds. Evidence for differences in the intelligibility of unscripted and read speech is mixed (Bunton & Keintz, 2008; Kempler & Lancker, 2002; Tjaden & Wilding, 2011), and we accounted for the possible effect of speech task by conducting our analysis with MDS-UPDRS Part III ratings based on both unscripted speech and SIT sentences. Nevertheless, sentence reading may not capture the way a person speaks in day-to-day conversation, and thus using reading-based estimates of intelligibility is a potential limitation.

We also did not have information on the cognitive function of all participants, and a language analysis was outside of the scope of this project. However, cognitive–linguistic factors may have played a role here and are worth investigating in future research.

Lastly, this study offers a limited evaluation of the MDS-UPDRS Part III speech item, focusing on those psychometric properties likely to have a notable impact on the clinical utility of this tool. A more comprehensive review that evaluates other important metrics of validity and reliability and fully documents the psychometric strengths and weaknesses of the speech item is warranted.

Conclusions

Ratings of speech impairment by movement disorders specialists when following the standard protocol of the MDS-UPDRS Part III show no concurrent validity with the criterion measure of intelligibility during sentence reading. These speech ratings also show weak reproducibility in terms of both interrater reliability and agreement. The results of this preliminary psychometric review of the MDS-UPDRS Part III speech item suggest it has limited clinical utility for evaluating the communicative function of people with PD, with implications for its use in clinical decision making and research. These results further suggest that the MDS-UPDRS Part III speech ratings should not determine who is referred for speech services to address communication concerns related to PD.

Data Availability Statement

The data sets generated or analyzed during the current study are not publicly available due to commitments to protect participant confidentiality.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders Grants DC021080 (awarded to K.L.D); DC016656 (awarded to K.M.S.); and DC015570, DC013017, and DC020867 (awarded to C.E.S.), as well as the Council of Academic Programs in Communication Sciences and Disorders PhD Scholarship (awarded to K.L.D.).

References

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology, 28*(3), 1222–1232. https://doi.org/10.1044/2019_AJSLP-18-0275

Baylor, C., Cook, K. J., & McAuliffe, M. J. (2024). "Take us into account": Perspectives of family members of people with Parkinson's disease regarding speech-language pathology intervention. *American Journal of Speech-Language Pathology, 33*(2), 736–755. https://doi.org/10.1044/2023_AJSLP-23-00273

Behrman, A., Cody, J., Elandary, S., Flom, P., & Chitnis, S. (2020). The effect of SPEAK OUT! and The LOUD Crowd on dysarthria due to Parkinson's disease. *American Journal of Speech-Language Pathology, 29*(3), 1448–1465. https://doi.org/10.1044/2020_AJSLP-19-00024

Borrie, S. A., Wynn, C. J., Berisha, V., & Barrett, T. S. (2022). From speech acoustics to communicative participation in dysarthria: Toward a causal framework. *Journal of Speech, Language, and Hearing Research, 65*(2), 405–418. https://doi.org/10.1044/2021_JSLHR-21-00306

Brabenec, L., Mekyska, J., Galaz, Z., & Rektorova, I. (2017). Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission, 124*(3), 303–334. https://doi.org/10.1007/s00702-017-1676-0

Bryans, L. A., Palmer, A. D., Anderson, S., Schindler, J., & Graville, D. J. (2021). The impact of Lee Silverman Voice Treatment (LSVT LOUD®) on voice, communication, and participation: Findings from a prospective, longitudinal study. *Journal of Communication Disorders, 89,* Article 106031. https://doi.org/10.1016/j.jcomdis.2020.106031

Bunton, K., & Keintz, C. K. (2008). The use of a dual-task paradigm for assessing speech intelligibility in clients with Parkinson disease. *Journal of Medical Speech-Language Pathology, 16*(3), 141–155.

Dahl, K. L., Balz, M. A., Díaz Cádiz, M., & Stepp, C. E. (2024). How to efficiently measure the intelligibility of people with Parkinson's disease. *American Journal of Speech-Language Pathology, 34*(1), 70–84. https://doi.org/10.1044/2024_AJSLP-24-00080

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

Duffy, J. R. (2019). *Motor speech disorders: Substrates, differential diagnosis, and management* (4th ed.). Elsevier. https://shop.elsevier.com/books/motor-speech-disorders/duffy/978-0-323-53054-5

Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research, 54*(2), 430–447. https://doi.org/10.1044/1092-4388(2010/09-0205)

Fereshtehnejad, S.-M., Yao, C., Pelletier, A., Montplaisir, J. Y., Gagnon, J.-F., & Postuma, R. B. (2019). Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: A prospective study. *Brain, 142*(7), 2051–2067. https://doi.org/10.1093/brain/awz111

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (R package Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Goetz, C. G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stebbins, G. T., Stern, M. B., Tilley, B. C., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., Olanow, C. W., . . . LaPelle, N. (2007). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. *Movement Disorders, 22*(1), 41–47. https://doi.org/10.1002/mds.21198

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., . . . LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders, 23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Hirsch, M. E., Thompson, A., Kim, Y., & Lansford, K. L. (2022). The reliability and validity of speech-language pathologists' estimations of intelligibility in dysarthria. *Brain Sciences, 12*(8), Article 1011. https://doi.org/10.3390/brainsci12081011

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural Neurology, 11*(3), 131–137. https://doi.org/10.1155/1999/327643

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

Kempler, D., & Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, 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*(1), 47–57. https://doi.org/10.1044/jshd.4301.47

Mach, H., Baylor, C., Pompon, R. H., & Yorkston, K. (2021). Beyond the patient: A mixed-methods inquiry into family members' involvement in the treatment of Parkinson's disease to target third-party disability. *American Journal of Speech-Language Pathology, 30*(1), 169–185. https://doi.org/10.1044/2020_AJSLP-20-00168

Martínez-Martín, P., Gil-Nagel, A., Gracia, L. M., Gómez, J. B., Martínez-Sarriés, J., Bermejo, F., & The Cooperative Multicentric Group. (1994). Unified Parkinson's Disease Rating Scale characteristics and structure. *Movement Disorders, 9*(1), 76–83. https://doi.org/10.1002/mds.870090112

Martinez-Martin, P., Rodriguez-Blazquez, C., Alvarez-Sanchez, M., Arakaki, T., Bergareche-Yarza, A., Chade, A., Garretto, N., Gershanik, O., Kurtis, M. M., Martinez-Castrillo, J. C., Mendoza-Rodriguez, A., Moore, H. P., Rodriguez-Violante, M., Singer, C., Tilley, B. C., Huang, J., Stebbins, G. T., & Goetz, C. G. (2013). Expanded and independent validation of the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS). *Journal of Neurology, 260*(1), 228–236. https://doi.org/10.1007/s00415-012-6624-1

McAuliffe, M. J., Baylor, C. R., & Yorkston, K. M. (2017). Variables associated with communicative participation in Parkinson's disease and its relationship to measures of health-related quality-of-life. *International Journal of Speech-Language Pathology, 19*(4), 407–417. https://doi.org/10.1080/17549507.2016.1193900

Miller, N., Allcock, L., Jones, D., Noble, E., Hildreth, A. J., & Burn, D. J. (2007). Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry, 78*(11), 1188–1190. https://doi.org/10.1136/jnnp.2006.110171

Miller, N., Andrew, S., Noble, E., & Walshe, M. (2011). Changing perceptions of self as a communicator in Parkinson's disease: A longitudinal follow-up study. *Disability and Rehabilitation, 33*(3), 204–210. https://doi.org/10.3109/09638288.2010.508099

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramaker, C., Marinus, J., Stiggelbout, A. M., & van Hilten, B. J. (2002). Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders, 17*(5), 867–876. https://doi.org/10.1002/mds.10248

Ramig, L. O., Sapir, S., Countryman, S., Pawlas, A. A., O'Brien, C., Hoehn, M., & Thompson, L. L. (2001). Intensive voice treatment (LSVT®) for patients with Parkinson's disease: A 2 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry, 71*(4), 493–498. https://doi.org/10.1136/jnnp.71.4.493

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package Version 2.3.9) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Richards, M., Marder, K., Cote, L., & Mayeux, R. (1994). Interrater reliability of the Unified Parkinson's Disease Rating Scale motor examination. *Movement Disorders, 9*(1), 89–91. https://doi.org/10.1002/mds.870090114

Roberts, A. C., Rafferty, M. R., Wu, S. S., Miao, G., Cubillos, F., Simuni, T., Marras, C., Davis, T., Dahodwala, N., Neault, M., Ramirez-Zamora, A., Rafferty, M., Malaty, I., Parashos, S., Kraakevik, J., Simuni, T., Dahodwala, N., Jankovic, J., Simon, D., . . . Litvan, I. (2021). Patterns and predictors of referrals to allied health services for individuals with Parkinson's disease: A Parkinson's Foundation (PF) QII study. *Parkinsonism & Related Disorders, 83,* 115–122. https://doi.org/10.1016/j.parkreldis.2020.11.024

Schalling, E., Johansson, K., & Hartelius, L. (2017). Speech and communication changes reported by people with Parkinson's disease. *Folia Phoniatrica et Logopaedica, 69*(3), 131–141. https://doi.org/10.1159/000479927

Spencer, K. A., Brown, K. A., & Elder, G. (2022). Does the Unified Parkinson's Disease Rating Scale adequately estimate severity of dysarthria? *Journal of Neurology and Neuroscience, 13*(4), 1–4. https://www.itmedicalteam.pl/articles/does-the-unified-parkinsons-disease-rating-scale-adequately-estimate-severity-of-dysarthria.pdf [PDF]

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research, 59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Swales, M., Theodoros, D., Hill, A. J., & Russell, T. (2021). Communication and swallowing changes, everyday impacts and access to speech-language pathology services for people with Parkinson's disease: An Australian survey. *International Journal of Speech-Language Pathology, 23*(1), 70–82. https://doi.org/10.1080/17549507.2020.1739332

Tjaden, K., & Wilding, G. (2011). Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics & Phonetics, 25*(2), 155–168. https://doi.org/10.3109/02699206.2010.520185

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech: Test types, ratings and transcription measures. *Clinical Linguistics & Phonetics, 37*(1), 52–76. https://doi.org/10.1080/02699206.2021.2009918

Yorkston, K. M., Beukelman, D. R., & Tice, R. (1996). *Speech Intelligibility Test* [Computer software]. Tice Technologies.