Tipping the Scales: Indiscriminate Use of Interval Scales to Rate Diverse Dysarthric Features

[1]Kaila L. Stipancic, [2]Brooke-Mai Whelan, [3]Lauren Laur, [4]Yunxin Zhao, [5]Andrea Rohl, [6]Inyong Choi, and [6]Mili Kuruvilla-Dugdale

[1]Department of Communicative Sciences and Disorders, University at Buffalo

[2] Division of Speech Pathology, University of Queensland

[3]Department of Speech, Language and Hearing Sciences, University of Missouri

[4]Department of Electrical Engineering and Computer Science, University of Missouri

[5]Department of Neurosurgery, University of Iowa

[6]Department of Communication Sciences and Disorders, University of Iowa

Author Note

Running Title: Tipping the Scales

Correspondence: Mili Kuruvilla-Dugdale, mkuruvilladugdale@uiowa.edu

Keywords: Dysarthria, auditory-perceptual scaling, prothetic, metathetic dimensions

Abstract

Purpose: Error related to incorrect use of rating scales is problematic in the assessment and treatment of dysarthria. The main purpose of this project was to determine scale fit for cardinal speech features of hypokinetic dysarthria. A secondary aim was to determine rater reliability for the two different scales explored.

Methods: Forty-three speakers with Parkinson's disease (PD) and 25 neurologically healthy control talkers were recorded reading sentences from the Speech Intelligibility Test. Twenty-two healthy female listeners used both an equal appearing interval (EAI) scale and a direct magnitude estimation (DME) scale to rate five perceptual speech features (i.e., overall speech severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony) from these recordings. Regression analyses were used to determine the linearity of the relationship between the means of the EAI and DME ratings. Inter- and intrarater reliability was calculated using intraclass correlation coefficients and Spearman's correlation coefficients, respectively, for both EAI and DME ratings.

Results: There was a linear relationship between EAI and DME means for monotony, indicating it is a metathetic dimension. Curvilinear relationships were observed between the EAI and DME means for the other four features, indicating prothetic dimensions. Intra- and inter-rater reliability values were similar for EAI and DME ratings.

Discussion: Overall, results of this work suggest that DME is the best fit for scaling several hypokinetic dysarthria features, and not the conventionally used EAI scale. Prothetic dimensions best scaled by DME include overall speech severity, articulatory imprecision, reduced loudness, and short rushes of speech. Monotony was the only feature found to be a metathetic dimension and would be best scaled using EAI or DME. Findings call for rethinking the widespread use of EAI scales for rating perceptual features as part of the assessment and treatment of motor speech disorders.

Introduction

Auditory-perceptual judgements are the cornerstone of motor speech diagnostics, and are pivotal to tracking speech and voice changes over time (Kent, 1996). Perceptual methods and the speech characteristics they measure vary widely, with a preference toward rating scales for assessing system-level features (e.g., speech intelligibility) over word identification procedures and over subsystem dimensions (e.g., consonant imprecision). For motor speech disorders, subsystem-specific assessment was first introduced through the Mayo Clinic Rating System, which recommended the use of a 5-point scale, to identify clusters of deviant speech features salient to the six dysarthria subtypes (Darley et al., 1969a, 1969b). Over time, the utility and validity of the Mayo Clinic system have been questioned because of its time-intensive nature, limited reliability, and the indiscriminate use of an interval scale applied to all speech features (Ziegler et al., 2017). Notwithstanding the limitations of scale fit (i.e., construct validity) and reliability, a simplified feature-based system in conjunction with system-level measures provides a comprehensive account of the motor speech impairment and delineates the subsystem impairments contributing to global speech impairment severity and intelligibility loss.

In terms of validity, not all perceptual scales are considered equal. The three primary scaling methods include: i) interval scaling, which employs a definitive set of categories or numbers to assign to stimuli (e.g., equal appearing interval [EAI] and visual analog scaling [VAS]); ii) confusion scaling, which requires determination of difference thresholds (e.g., paired comparison ratings); and iii) magnitude or ratio scaling, which requires the assignment of numerical values proportional to the perceived ratio of the target features in a reference stimulus (e.g., direct magnitude estimation [DME]) (Stevens, 1975; Stevens & Galanter, 1957). Although EAI and VAS are amongst the most widely applied methods to evaluate speech and vocal characteristics (Kreiman et al., 1993), research indicates that EAI scaling is not appropriate for some perceptual dimensions, given evidence of listener bias toward subdividing the lower end of the scale into smaller intervals (Stevens, 1975). EAI scales also offer a limited set of categories to capture the listener's perception and listeners tend to use an equal amount of all the intervals when completing ratings (Zraick & Liss, 2000). For VAS, the level of measurement is unclear. Some researchers have reported that VAS provides only ordinal data (Kersten et al., 2012), while others suggest that VAS behaves like EAI and provides interval data (e.g., Reips & Funke, 2008). Yet other groups deem some types of VASs as ratio or magnitude scales because sensation ratios were in quantitative agreement with VAS ratings of sensory intensity (e.g., Price et al., 1983).

Scale validity (i.e., construct validity or how well a scale measures the dysarthric speech feature it is assigned to evaluate) is largely contingent upon the continuum class of a feature (Whitehill et al., 2002), i.e., whether a continuum is prothetic or metathetic. Simply put, a metathetically scaled dimension is a substitutive characteristic, or one that may be altered in terms of quality (e.g., pitch), as opposed to quantity. In contrast, prothetically scaled dimensions refer to additive characteristics that can be measured in relation to incremental changes in quantity or magnitude (e.g., loudness). In terms of the neurobiological difference, prothetic continua involve a quantitative receptor mechanism whereby increasing numbers of sensory receptors respond as the stimulus intensity increases. This differs from the metathetic continuum, which requires a substitutive or qualitative receptor mechanism whereby a different population of receptors is activated as the stimulus intensity increases (Ryan, 1971). A prothetic continuum is also distinct in that the magnitude of the psychological response grows as an exponent of the physical stimulus (e.g., loudness). Contrastively, for metathetic dimensions, the psychological response maintains a constant or uniform distance with the magnitude of the physical stimulus (e.g., pitch). A third difference between these continua is evident when plotting the mean scores of interval (e.g., EAI) and ratio (e.g., DME) scales. A linear relationship suggests a metathetic dimension whereby raters indicate interval equivalence on the measurement scales, and a non-linear relationship points to a prothetic dimension (Stevens, 1975). Although the distinction between prothetic and metathetic can be made based on the neural responses associated with each continuum (i.e., substitutive versus additive excitation) and/or the relationship between the physical stimulus and perceptual magnitude (e.g., dB values and loudness perception), it is most easily defined using interval and ratio scales, which is the focus of the current study. EAI scaling was chosen as one of the scaling methods (e.g., as opposed to VAS, for example), as was DME, to follow previous studies in this area.

The level of measurement (i.e., ordinal, interval, ratio) from an EAI scale can vary based on the type of continuum, the number of scale intervals (e.g., 5-, 7-, or 9-point), and whether descriptors (e.g., mild, moderate, severe) are used alongside scale values. In general, when listeners' sensitivity to differences is not constant over the scale, the equal interval nature of the data is questioned, and it is recommended that ratings from such a scale be treated as ordinal (Patel et al., 2008). Only one study has compared three methods for obtaining perceptual judgments based on the premise that equal perceptual distances between the methods suggests equivalent levels of measurement (Patel et al., 2010). The authors found that breathiness ratings obtained from a 7-point scale and a matching task (determined to provide ratio-level measurement) were best fit using a linear function for majority of the talkers, which suggested that the rating scale also has ratio-level measurement properties. This finding conflicts with the common notion that only interval level data is possible with an EAI scale, or that data from an EAI scale should be considered ordinal if listeners cannot apply each interval number as perceptually equidistant from its neighboring intervals (i.e., the perceptual difference between samples rated 4 and 3 should be the same as those rated 4 and 5). Similarly, numerical values combined with descriptors at scale intervals are assumed to provide ordinal-level measurement, where rank order, rather than the magnitude of difference between intervals, is measured. When a scale has fewer, less distinguishable intervals, rank ordering is more likely (Siegel, 1956; Stevens, 1969). To provide further clarification, Snijders and Bosker (2012) stated that variables measured using scales with five to ten intervals and with normal distributions can be considered as continuous rather than categorical variables obtained with an ordinal scale. The current study followed previous studies aimed at establishing psychophysical continua for voice, fluency, and resonance features, by using a 5-point interval scale with descriptors, and as done in these studies, we acknowledge that it cannot be assumed the intervals on an EAI scale represent equally distanced perceptual points.

Employing Stevens' (1975) approach, several perceptual speech and voice features have been mapped to metathetic and prothetic scaling dimensions. For example, non-linear relationships and thus, prothetic attribution have been reported for the following perceptual features: overall severity in dysphonic (Eadie & Doyle, 2002b) and tracheoesophageal speech (Eadie & Doyle, 2002a), and hypernasality in speakers with repaired cleft palate (Whitehill et al., 2002; Zraick & Liss, 2000). Prothetic dimensions are best scaled via DME methods; EAI is a poor fit for prothetic dimensions because listeners cannot keep the intervals perceptually equal as they assign stimuli to the various intervals (Stevens, 1971). Conversely, breathiness in dysphonic (Yiu & Ng, 2004) and normal speakers (Sewall et al., 1999), naturalness in stuttered speech (Metz et al., 1990) and tracheoesophageal speech (Eadie & Doyle, 2002a), and pleasantness in dysphonic speakers (Eadie & Doyle, 2002b), have been identified as metathetic. For these dimensions, either DME or EAI scaling is appropriate because listeners' sensitivity to differences is constant over the EAI scale for metathetic continua (Stevens, 1971). Voice disorders have received the most attention in this area and there is a significant paucity of studies on scale fit for measuring dimensions of dysarthria.

Error related to incorrect scale use is particularly problematic in the assessment of dysarthrias that progress quickly. As an example, when tracking dysarthria progression, if a change in score from 2 to 3 is more significant than a change from 4 to 5 because listeners tend to divide the lower end of the scale into smaller intervals, it can lead to underestimation of functional loss in the most active phase of the disease when speech changes are accelerated. Moreover, global, system-level measures of dysarthria severity, such as intelligibility and overall speech severity, are very popular in both research and clinical settings and are thought to serve as meaningful indicators of speech loss over time. In speakers with dysphonia, overall severity has been found to be a prothetic dimension best evaluated with DME; however, scale validity for overall severity has not yet been established in speakers with dysarthria. From a treatment standpoint, subsystem features, such as reduced loudness or short rushes of speech often observed in patients with Parkinson's disease (PD), are the targets of therapy aimed at improving speech production. If these are prothetic dimensions, using an EAI scale to quantify intervention outcomes is inappropriate because the change in scores pre- and post-intervention will be difficult to interpret because a change on the lower end is weighted differently than a change on the upper end of the scale. Lastly, relationships between perceptual and acoustic or perceptual and physiologic dimensions can be misinterpreted because invalid scaling options may have been employed. For example, one study showed that the perception of breathiness rated with an EAI scale was not supported by acoustic data (i.e., non-significant noise to harmonic ratio) across both early and late state PD (Holmes et al., 2000) – scale misfit may be one reason, among others, for this discrepancy.

Reliability of raters has also been an area of concern in the perceptual evaluation of dysarthric speech. A study by Bunton et al. (2007) examined both intra- and inter-rater reliability of EAI for the 38 perceptual dimensions in the Mayo Clinic Rating System across inexperienced and experienced listeners. In their study, speech samples from 47 speakers with dysarthria were rated by 20 listeners (i.e., 10 inexperienced listeners and 10 clinicians) on the 38 perceptual features using a 7-point EAI scale. Percentage of exact score agreements across the features ranged from 32.78% (for pitch level) to 100% (for grunt at the end of expiration). The authors also found no difference in agreement between inexperienced and experienced listeners. The wide range of agreement suggests that although listeners are highly reliable for some features, listeners may have more difficulty (or are more variable) with other perceptual features. Even rating scales that have reliability that is upwards of 50-70% agreement may be insufficient for clinical use, although there is not an agreed-upon standard for this. The authors also noted that more research is needed to determine how to improve reliability of these ratings. Eadie and Doyle (2002b) reported good reliability (i.e., ICCs ranging from 0.692-0.984) for *both* DME and EAI judgments of voice pleasantness and overall severity. These ratings were completed by 12 speech-language pathology (SLP) graduate students who listened to 24 speakers with dysphonia and 24 control speakers. Although the authors did not comment on a comparison of these reliability statistics in the discussion, it appears that reliability was grossly similar for the DME ratings and the EAI ratings for the features rated in their study. From these previous studies, it is clear that the reliability of ratings must be considered when attempting to characterize the psychometric properties of rating scales.

Current Study Aims

The first aim of the study was to determine scale fit for the cardinal speech features of hypokinetic dysarthria, namely consonant imprecision, reduced loudness, short rushes of speech, monotony, and overall severity with non-expert listeners as the judges. The focus was on PD because unlike more rapidly progressing dysarthria types (e.g., spastic-flaccid dysarthria due to amyotrophic lateral sclerosis) where there is overreliance on system-level dimensions, both system and subsystem-specific features are frequently used to assess and manage hypokinetic dysarthria. The salient features were selected to cover the speech subsystems predominantly involved in PD, i.e., reduced loudness (phonatory/respiratory subsystems), consonant imprecision (articulatory subsystem), short rushes of speech (articulatory/prosodic subsystems), and monotony (phonatory/respiratory/prosodic subsystems), in addition to overall speech impairment severity (several speech subsystems contribute to this dimension).

Based on the existing voice literature, we expected overall speech impairment severity to be prothetic. Because there are no existing studies on the remaining dysarthria features, we could not generate hypotheses for articulatory imprecision, reduced loudness, short rushes of speech, and monotony. Although the psychophysical scaling literature shows that both loudness and duration are prothetic continua (Stevens & Galanter, 1957), it is unclear if derivatives of these features that are considered pathological like reduced loudness and short rushes of speech are also prothetic in nature. Additionally, as mentioned earlier, the continuum class of various features can only be determined in these specific ways: (i) the relationship between scaled interval and ratio scores, (ii) the physical stimulus-perceptual response relationship, and (iii) the neural mechanisms that subserve each continuum. Therefore, it is challenging to hypothesize which dysarthria dimensions are prothetic versus metathetic and which features would be better suited to EAI versus DME measurement solely from perceptual impressions or knowledge about the nature of each feature.

Only inexperienced listeners were included in this study because they are more representative of the general population and tend to weight perceptual characteristics differently than highly trained experts. Further, untrained listeners play a role in evaluating dysarthric speech, both in clinical and research settings, when familiarity bias is likely to impede speech evaluation, and when the real-world significance of communication limitations needs to be determined (Lehner, 2021). It is not uncommon for nonexperts to complete fine-grained auditory analysis, for example, master's students in communication sciences and disorders, who are largely untrained when they enter research and clinical settings yet perform these types of ratings. Additionally, from a research perspective, inexperienced listeners are commonly employed to provide ratings of dysarthric speech. Examining how inexperienced listeners use rating scales is a necessary endeavor to support future research efforts that will continue to recruit nonexpert listeners. In psychophysics, most of the seminal studies that established the principles of optimal scaling used nonexpert listeners; more recent studies have also followed previous work in this area by using similar listeners.

The second aim of the study was to determine inter- and intrarater reliability for each scale and feature. Based on prior voice findings, for prothetic features, EAI scaling was expected to show lower inter- and intrarater reliability compared to DME. For metathetic features, the inter- and intra-rater reliability was expected to be similar for EAI and DME.

Methods

The study was conducted and received approval from the Institutional Review Board at the University of Missouri. A consent script was read to all the listeners, and written informed consent was received from the speakers. Both listeners and speakers were compensated for participating in the study.

Participants

Speakers

Forty-three individuals with Parkinson's disease (18 females, 25 males) and 25 neurotypical controls (11 females, 14 males) provided speech samples for the study. The mean age of the PD group was 68.14 years (*SD* = 7.65), and of the control group was 70.32 years (*SD* = 8.43). All participants met the following criteria: (i) negative history of speech, language, and hearing disorders (except those related to the diagnosis for the PD group); (ii) absence of a co-existing neurological diagnosis for the participants with PD, and any neurological diagnosis for the controls; (iii) negative history of head and neck surgery; (iv) no hearing aids or a prescription for hearing aids; and (v) monolingual, native speakers of American English. The participants with PD displayed a range of dysarthria severities with the majority showing mild dysarthria (*n* = 14), followed by moderate dysarthria (*n* = 13), severe dysarthria (*n* = 4), and lastly, profound dysarthria (n = 3), as determined by the clinical impressions of two experienced SLPs. Nine PD participants displayed typical speech.

Listeners

Twenty non-expert neurologically healthy females rated the speech samples provided by the PD and control groups using EAI and DME scales. The mean age of the listeners was 23.91 years (*SD* = 4.39). All listeners passed a bilateral hearing screening at 25 dB HL at 500Hz, 1 kHz, 2 kHz, and 4 kHz. They also met all the same inclusionary criteria as the speakers and had minimal exposure to communication disorders (i.e., they had not worked with clinical populations or received any formal instruction in communication disorders).

Experimental Task

Speakers

The speakers were asked to read aloud sentences from the Speech Intelligibility Test (SIT; Yorkston et al., 2007) at their typical rate and loudness. A print version of the SIT sentences was presented to each speaker. Font size was checked, and readability was confirmed before commencing the recording. The 11 sentences presented to each speaker were randomly generated by the SIT software such that majority (if not all) of the sentences differed across speakers. The speakers read each sentence aloud in their typical rate and loudness after the experimenter announced the sentence number. The speech samples were recorded in a quiet laboratory setting as part of other studies conducted in the last author's lab. A condenser microphone (Shure, Model PG42, Niles, IL) placed 20cm away from the mouth was used to record audio at a sampling rate of 22kHz and the samples were stored on a digital recorder (Marantz, Model PDM670, Eindhoven, Netherlands). Of the 11 SIT sentences, only one sentence per speaker was included in the listening task. The length of the selected sentence varied from 12 to 15 words across the speakers. For each speaker, we selected the sentence with the highest number of hypokinetic speech features represented (i.e., reduced loudness, consonant imprecision, short rushes of speech, and monotony). To determine which features were present in each sample, two trained research assistants, who had completed the graduate Motor Speech Disorders course, used the Dysarthria Rating Scale (Darley et al., 1969 a & b) to make independent judgments. They arrived at a consensus for features with divergent ratings, and the consensus ratings were used for sample selection.

Listeners

The listeners performed the perceptual ratings over two sessions that were about one week apart and lasted approximately an hour each. Either a 5-point EAI scale or modulus DME scale was used in each session. Across listeners, the order of the scales and speech samples was randomized. Listeners rated 82 samples (68 samples plus 14 re-rated for calculation of intra-rater reliability) in each session. Among the 68 samples, none of the sentences were repeated. Before commencing each task, the experimenter (last author) provided definitions for each of the five features to be rated, followed by instructions on appropriate scale use. For example, for overall severity, listeners were asked to rate the samples based on their general impression of speech impairment severity and not on understandability. For reduced loudness, they were instructed to rate the extent that the voice was insufficiently loud. For consonant imprecision, listeners were asked to determine if some or most sounds were produced crisply and sharply, and for short rushes they were asked to listen for rapid bursts of speech separated by pauses. For monotony, listeners were told to rate the extent that the sample sounded flat in terms of pitch, loudness, or duration. Regarding scale use, listeners were instructed to use the full range of the scale. Listeners were strongly encouraged to rate the first three features after listening to the sample once and rate the next two features after listening to the sample a second time to avoid confusing the features to be rated. The features were rated in the following order across all samples for both EAI and DME: overall speech impairment severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony. Listeners were only allowed to listen to each sample twice. The 5-point EAI scale had the following intervals: 1=typical, 2=mild, 3=moderate, 4=severe, and 5=profound. A 5-point scale was used to be consistent with the procedures of the Mayo Clinic Rating System from where the features were taken. The severity descriptors used at the intervals of the scale also follow the Mayo Clinic Rating System.

For DME, the listeners were instructed to first listen to the moduli (or references) for the first three features. Each modulus represented a moderate level of severity and was given a score of 100; each of the features had a different modulus. After listening to the moduli, the listeners played the sample and provided a comparative score between the modulus and the sample. Specifically, listeners were told that if they perceived the sample to be twice as severe as the modulus, a score of 200 would be appropriate. Similarly, if they perceived the sample to be half as severe as the modulus, a score of 50 would be appropriate. The lower limit was fixed at 1 to carry out geometric mean calculations; no upper limit was specified. After entering their scores for the first three features, listeners proceeded to listen to the last two moduli before listening to the sample a second time and scoring the last two features. Listeners were required to relisten to the moduli after every 6^{th} sample to maintain referential value of the modulus and to minimize shifting of the listeners' internal standards when performing the ratings (Eadie & Doyle, 2002; Kreiman et al., 1993).

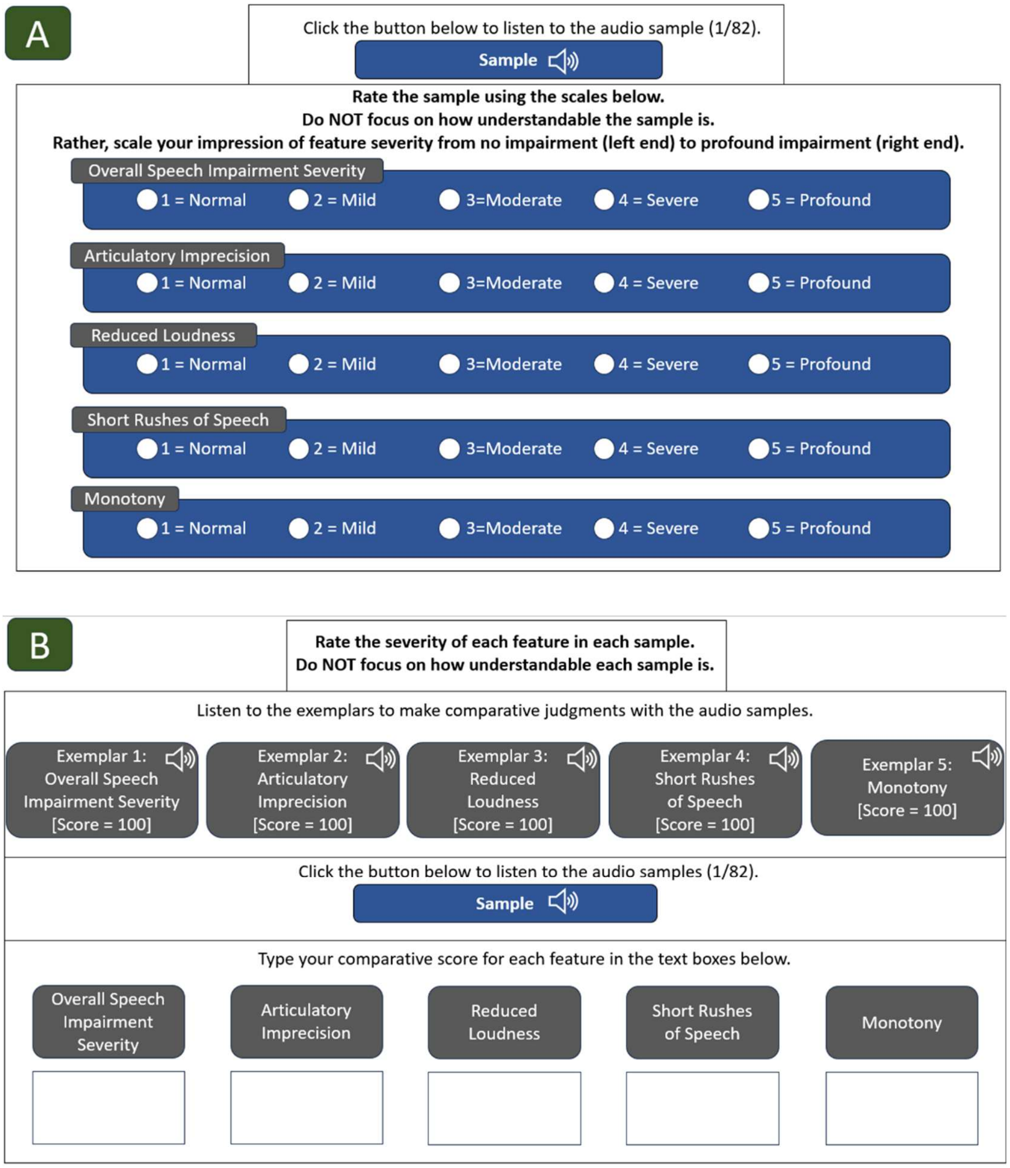
The DME moduli were initially selected by an experienced SLP who listened to dysarthria samples from the *Audio Seminar Series* (Darley et al., 1975) and chose samples for each feature representing mild, moderate, severe, and profound severity. Once these samples were identified for each feature, the last author rated them independently for severity, as well as the feature(s) represented in each sample. Consensus was sought when the two experts disagreed about feature representation or severity level; both experts confirmed that the modulus for each feature represented moderate severity.

Data Acquisition

All audio samples were presented through headphones (Panasonic, RP-HC200-K and RP-DJS150) from a desktop computer (Dell, OptiPlex 7010) in a quiet laboratory setting. The sound files recorded on different dates had different intensity levels. Therefore, the intensity levels were normalized across all the sound files using the Yoon script (Yoon, 2022) in Praat (Boersma & Weenink, 2021). When the script is run, the sound files are scanned to identify the maximum and minimum intensity in dB. The sound files are then normalized to 65 dB. Because two different computers and headsets were used, the sound pressure level (SPL) from each headphone was calibrated with a half-inch precision condenser microphone (PCB Piezotronics, 377C13) and a sound level meter (Larson Davis, SoundTrack LxT®), by connecting the headphone to an ear simulator (Larson Davis, AEC201). A one-minute speech sample devoid of long pauses or silences was played through the headphone connected to the artificial ear, and the SPL, via the sound card interface, was adjusted until the sound level meter showed 75dB SPL. The volume level corresponding to 75 dB SPL was noted for each computer-headset pair and checked before each session to ensure that all stimuli were played at the same loudness level across listener participants.

Five EAI scales clearly demarcated for each feature were available to the listeners via a custom-built MATLAB GUI. Listeners were instructed to use the correct scale to judge each dimension from 1 (i.e., no impairment) to 5 (i.e., profound impairment). A second custom-built MATLAB GUI was used for the DME ratings. The moduli were embedded at the top of the screen and five textboxes were available at the bottom of the screen to enter the comparative scores for the features. Renderings of the GUIs provided to listener participants are displayed in Figure 1. For both scales, listeners could only advance to the next sample after all five scores were entered. Moreover, for DME, listeners could not proceed if they missed a modulus or attempted to listen to a modulus more than once or when the modulus was inactive.

Figure 1. Renderings of the MATLAB GUIs used by listeners for (A) Equal Appearing Interval (EAI) Scales; (B) Direct Magnitude Estimation (DME).



Data Analysis

EAI and DME scores were generated automatically as part of the MATLAB output file. DME and EAI ratings were compared using Stevens' (1975) method. Geometric rather than arithmetic means are required for DME because the relationship between scores is multiplicative or exponential. Contrastively, using arithmetic means is appropriate for EAI because the relationship between scores is additive. To compute the arithmetic means from EAI scores, the scores were averaged across all 20 listeners for each feature and speaker. For DME, the geometric mean was calculated across all the listeners for each feature and speaker using the formula $\sqrt[n]{X_1 X_2 \ldots X_n}$ (where X = values of DME ratings, n = number of speakers). The EAI arithmetic means were plotted against the DME geometric means for all samples (excluding the samples repeated for intrarater reliability). Following Stevens' (1975) interpretation, a linear relationship between the two sets of means indicated a metathetic continuum, whereas a downward bowed curvilinear function indicated a prothetic continuum.

Statistical Analysis

To determine scale fit for each of the five features the following steps were taken in SPSS version 28 (IBM SPSS, Armonk, NY):

Linear Regression

A linear regression was performed for each feature to determine the linearity of the relationship between the means of the EAI and DME ratings. The geometric mean was placed in block 1, the computed square of the geometric mean in block 2, and the computed cube of the geometric mean in block 3. The arithmetic mean was the dependent variable in each model. The model summary, specifically the adjusted R^2 values and significant F change were checked for each feature to determine fit. Significance was set at $p < .05$. Cook's distance was examined to identify outliers; for all features, data were removed from one participant with PD due to Cook's distance being greater than 1.0. For the same reason, data from a second PD participant was also removed for all features except monotony. For reduced loudness, visual inspection of the predicted values versus residual plots revealed heteroscedasticity; therefore, the arithmetic means of the EAI scores were log transformed. Linear regressions were performed again after removing the outliers and correcting heteroscedasticity. Visual inspection of the residuals showed no signs of violating the normality assumption. Additionally, the Shapiro Wilk test was performed and showed normal distribution of residuals for overall severity ($W(61) = 0.97$, $p > 0.05$), articulatory imprecision ($W(61) = 0.96$, $p > 0.05$), reduced loudness ($W(61) = 0.98$, $p > 0.05$), short rushes ($W(61) = 0.97$, $p > 0.05$), and monotony ($W(61) = 0.99$, $p > 0.05$).

Collinearity Diagnostics

This step helped determine whether collinearity affected the fit of the linear, quadratic, and cubic models. When the variance inflation factor was above 10, the condition index was inspected for values greater than 30, and variance proportions greater than 0.90 among two of the predictors. For all five features, only the cubic model met these three criteria; therefore, the cubic model was excluded for all features.

Linear regression assumes interval or ratio-level data, and taking an average of scaled scores is regarded as requiring an interval scale. Because none of the final models were found to violate regression assumptions of normality, collinearity, homoscedasticity, and outliers, the regression results are considered valid.

Rater Reliability

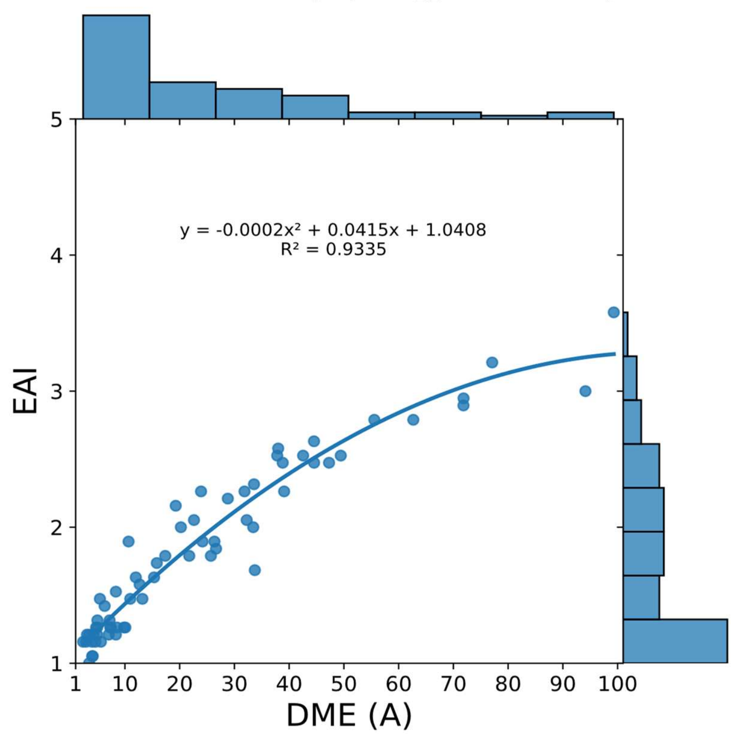
Reliability between raters (i.e., interrater reliability) was estimated using intraclass correlation coefficients (ICCs) for each scaling method. ICCs and their 95% confidence intervals (CIs) were calculated using SPSS statistical package version 28 (SPSS Inc., Chicago, IL) based on average-measures consistency, 2-way mixed-effects model with 20 listeners across 68 samples. Reliability within raters (i.e., intrarater reliability) was judged using Spearman’s correlation coefficients.

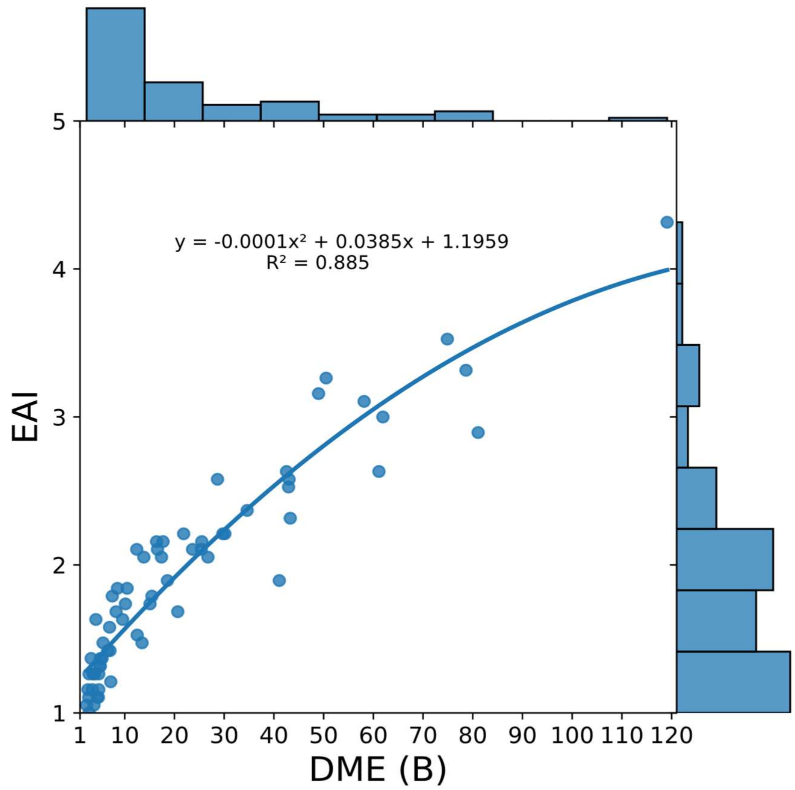
Results

EAI arithmetic means plotted as a function of the DME geometric means revealed a statistically significant result for a second-order polynomial curve of best fit for overall severity ($F(2, 64) = 432.47$, $p < .001$, $R^2_{\text{Adjusted}} = .93$), articulatory imprecision ($F(2, 65) = 241.61$, $p < .001$, $R^2_{\text{Adjusted}} = .88$), reduced loudness ($F(2, 65) = 85.98$, $p < .001$, $R^2_{\text{Adjusted}} = .72$), and short rushes of speech ($F(2, 65) = 197.50$, $p < .001$, $R^2_{\text{Adjusted}} = .86$). This indicates that the curvilinear model accounted for a statistically significant amount of the variance above that observed with a simple linear model. As shown in Figures 2 a-d, visual inspection of the model revealed a downward bowing towards the end of the curve, a finding that agrees with other data for prothetic continua reported in the literature (Eadie & Doyle, 2002 a & b; Zraick and Liss, 2000).

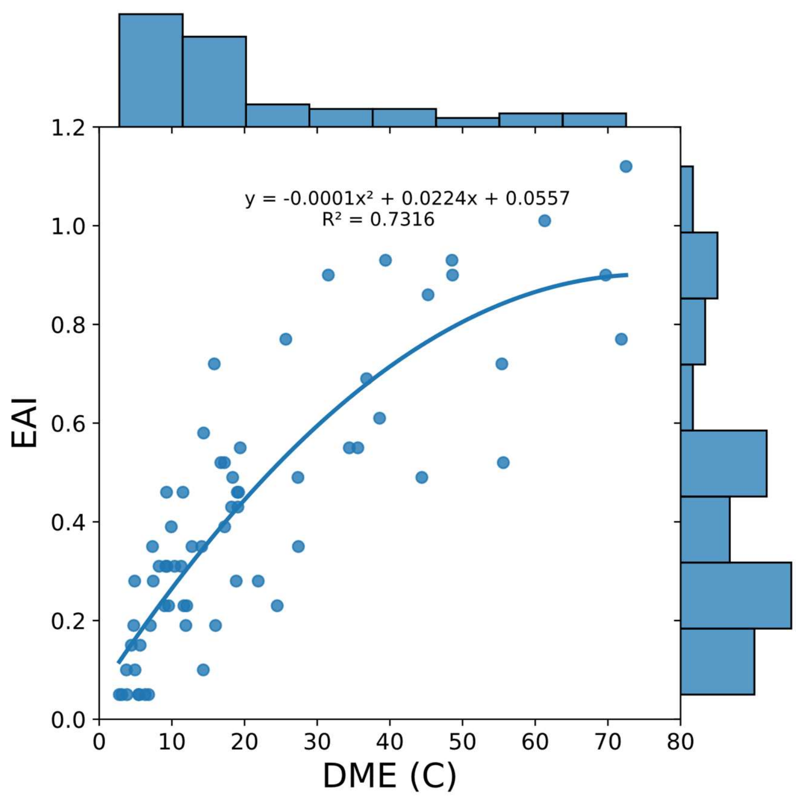
In contrast, a linear model accounted for a statistically significant amount of variance for monotony ($F(1, 66) = 376.26$, $p < .001$, $R^2_{\text{Adjusted}} = .85$). The quadratic and cubic curvilinear models revealed no significant improvement in the variance over the linear model. Visual inspection of the data also revealed a good approximation to the raw data by the linear regression, with no apparent downward bowing (see Figure 2 e).

Figure 2. Arithmetic means of equal appearing interval scale (EAI) scores plotted against the geometric means of the direct magnitude estimation (DME) scores for (A) overall severity, (B) articulatory imprecision, (C) reduced loudness, (D) short rushes of speech, and (E) monotony. The arithmetic means for reduced loudness were log transformed. The histograms based on the means of the EAI and DME scores are displayed opposite each respective scale axis.

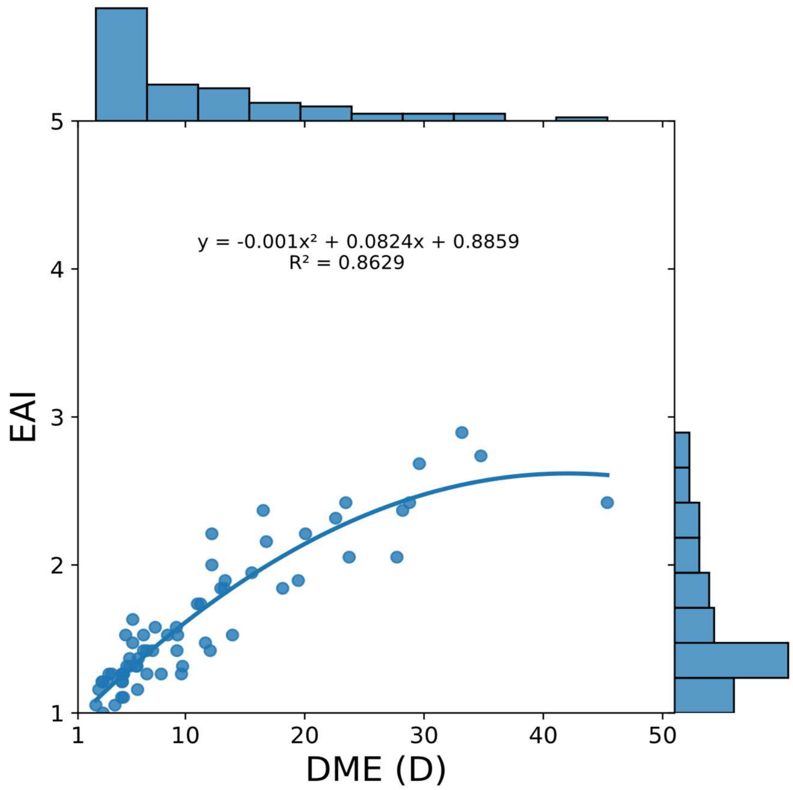


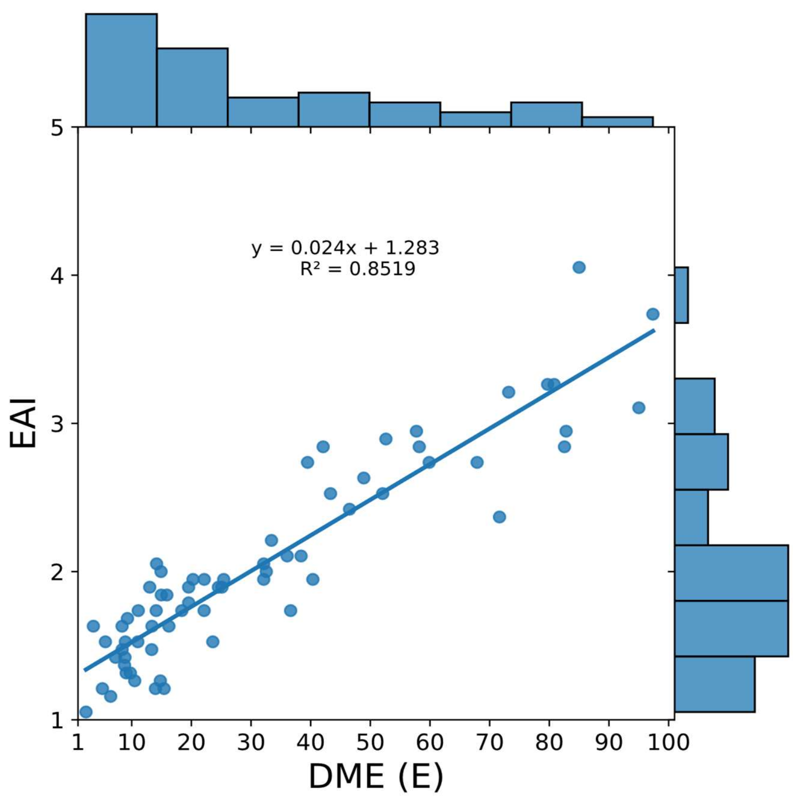

y = -0.0001x² + 0.0385x + 1.1959
R² = 0.885
EAI
DME (B)
1
2
3
4
5
1 10 20 30 40 50 60 70 80 90 100 110 120

437


y = -0.0001x² + 0.0224x + 0.0557
R² = 0.7316
EAI
DME (C)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0 10 20 30 40 50 60 70 80

438


y = -0.001x² + 0.0824x + 0.8859
R² = 0.8629
EAI
DME (D)
1
2
3
4
5
1
10
20
30
40
50


y = 0.024x + 1.283
R² = 0.8519
EAI
DME (E)
1
2
3
4
5
1
10
20
30
40
50
60
70
80
90
100

Rater Reliability

Table 1 displays the interrater and intrarater reliability results. Interrater reliability coefficients were highly acceptable (i.e., 0.89 or greater, indicating good or excellent reliability) for all features and both scaling methods (Koo & Li, 2016). Intrarater reliability was moderate (0.64-0.69) to strong (0.71-0.79) for both methods (Dancey & Reidy, 2004).

Table 1. Correlation coefficients for inter- and intrarater reliability of equal appearing interval (EAI) and direct magnitude estimation (DME) ratings for all features.

Dimensions	Inter-Rater Reliability Intraclass Correlation Coefficients (95% Confidence Intervals)		Intra-Rater Reliability Spearman's Correlation Coefficient (95% Confidence Intervals)	
	EAI	DME	EAI	DME
Overall severity	0.96 (.949-.975)	0.93 (.895-.948)	0.75 (.695-.803)	0.76 (.697-.804)
Articulatory imprecision	0.94 (.929-.975)	0.93 (.898-.949)	0.79 (.734-.829)	0.69 (.621-.751)
Reduced loudness	0.91 (.872-.937)	0.89 (.846-.924)	0.65 (.566-.712)	0.65 (.568-.713)
Short rushes of speech	0.93 (.895-.948)	0.92 (.884-.942)	0.71 (.638-.763)	0.72 (.651-.772)
Monotony	0.92 (.895-.948)	0.90 (.860-.930)	0.65 (.574-.718)	0.64 (.560-.707)

Discussion

The current research note sought to determine scale fit for several speech features whose measurements are critical to the evaluation of hypokinetic dysarthria associated with PD: overall speech severity, articulatory imprecision, reduced loudness, short rushes of speech, and monotony. Specifically, we were interested in whether the commonly used EAI scale, or the less popular DME scale would yield better validity and reliability. Schiavetti et al. (1981) referred to determining whether a continuum is prothetic or metathetic as an aspect of construct validity. As a reminder, three ways to determine continua class are by examining the (i) relationship between interval and ratio scales, (ii) relationship between physical and perceptual magnitude, and (iii) neural activation patterns. In this study, we compared the interval-ratio scale relationship, and our findings indicate that DME scaling has greater construct validity than EAI scaling for most of the features examined, except monotony. Our first hypothesis that overall speech impairment severity would be a prothetic continua was supported by these results.

Our hypotheses for reliability were not entirely supported because reliability was comparable for both EAI and DME ratings for all features. We hypothesized that for prothetic continua, like overall speech impairment, articulatory imprecision, reduced loudness, and short rushes of speech, reliability would be lower for EAI than for DME. The reliability hypothesis for metathetic continua was supported by the data for monotony ratings – the hypothesis was that for metathetic continua, rater reliability would be similar for the two rating tasks. Herein, we discuss these findings along with implications for clinical settings and research endeavors, as well as directions for future research in this area.

Early work describing methods for evaluating scale construction highlighted the importance of considering both reliability and validity of a scale to determine its usefulness, even though these concepts are complex and challenging to fully evaluate (Dawis, 1987). Therefore, as a brief, initial discussion point about the current study, inter-rater reliability was excellent for both scaling methods across all five features (ICCs ranged from 0.89 to 0.96). The reliability statistics in the current project are even slightly higher than those reported in recent work from our group for VAS ratings of speakers with PD (Stipancic et al., 2023). This was a somewhat unexpected finding, as rater reliability in scaling tasks has been questioned in the speech perception literature (Miller, 2013; Schiavetti, 1992; Stipancic et al., 2016). One difference among studies is whether single measures or average measures ICC is reported. The former is based on a single measurement from a single observer, and the latter is based on the average measurements of more than one observer. Average measures ICC tends to have higher values because it accounts for the reduction in measurement error achieved by averaging multiple measurements (Trevethan, 2017). Another unexpected finding was the slightly lower intrarater reliability compared to interrater reliability across all the features, given that previous literature has generally demonstrated the opposite (i.e., that listeners are more reliable *within* themselves than *across* other listeners). Because the two types of scaling procedures used in the current project have similar levels of reliability, scale choice should rely on other considerations, such as those discussed in the following sections.

Construct Validity is Better for DME than EAI for Most Dysarthric Features Explored

Non-linear relationships between EAI and DME ratings were observed for four out of five features (i.e., overall speech severity, articulatory imprecision, reduced loudness, and short rushes of speech), indicating that they are prothetic dimensions. This finding suggests that these four dimensions should not be scaled using EAI, which is what the Mayo Clinic Rating System recommends. These findings point to DME scaling having better construct validity than EAI scaling for the dimensions of overall speech severity, articulatory imprecision, reduced loudness, and short rushes of speech. The current results are similar to seminal findings from Schiavetti (1992) on the construct validity of scaling speech intelligibility. According to the author, EAI scaling is not appropriate to measure intelligibility because it cannot be partitioned linearly into equal intervals. Using an EAI scale for intelligibility results in an ordinal rather than interval level of measurement. Similarly, it follows that overall speech severity, articulatory imprecision, reduced loudness, and short rushes of speech, as rated in hypokinetic dysarthria, cannot be partitioned into equal intervals by listeners, making an EAI scale inappropriate to use for ratings.

Scaling Monotony Appears to Differ from the Other Dysarthric Features

In contrast to the other four dimensions, there was a linear relationship between EAI and DME ratings for monotony, indicating that it is a metathetic dimension. Interestingly, in recent work (Stipancic et al., 2023), our group identified ratings of monotony as having the poorest criterion validity and reliability compared to three other features of dysarthria (i.e., overall speech impairment, articulatory imprecision, and slow rate). Findings of the current study suggest that monotony can be rated using an EAI scale and, therefore, that listeners perceive monotony more linearly than the other dimensions. As a reminder, listeners were not asked to make a distinction between monopitch and monoloudness because it is difficult for non-expert listeners to distinguish these categories (Kim, 1994). In future studies, we propose that construct validity be determined for each sub-feature separately, regardless of the fit of the "root" feature. In this case, for example, ratings of monoloudness, monopitch, and monoduration should all have scale fit independently determined, outside of their contributions to monotony. The relative relationships of each of these features to monotony should also be considered in future work. Overall, taking into account the current study and our previous work (Stipancic et al., 2023), monotony appears to be a challenging dimension for non-expert listeners to rate, whether it be because it is a composite feature (i.e., made up of other features), or because its rating is limited due to psychophysical and neural restrictions in the ability to perceive monotony, or because listeners have a poor working definition of monotony. Improving the rating of monotony will be an interesting area of future research, given its importance as a feature that has long been considered central to the diagnosis of dysarthria subtypes like hypokinetic dysarthria (Darley et al., 1969 a & b).

Clinical and Research Implications

Results of this work suggest that DME is the best fit for scaling several hypokinetic dysarthria features, and not the conventionally used EAI scale. Error related to incorrect scale use has implications for both assessment and treatment tracking of individuals with dysarthria. This is especially true for the feature of overall severity which is ubiquitous in the field both in clinical practice and in research studies (Stipancic et al., 2021). Further, for individuals with PD participating in programs like the Lee Silverman Voice Treatment, multiple loudness ratings are typically carried out during assessment (e.g., baseline and stimulability check) and treatment (e.g., daily for four weeks). Similarly, features tied to speech rate, such as short rushes of speech, are likely to receive additional attention from a clinician during treatment sessions. If an EAI scale is used in these instances, listeners will be unable to place stimuli along the scale in an unbiased manner due to the propensity to divide the lower end of the continuum into finer segments than the upper end (Stevens, 1974; Whitehill et al., 2002). EAI scales, therefore, limit listeners from fully communicating their auditory perception when it comes to prothetic continua. This is thought to result in the ordinal partition of the interval scale, and if used in research studies, different descriptive and inferential statistics are recommended for ordinal versus interval variables to decrease the chance of erroneous conclusions from distorted effect sizes, inflated false alarm rates, etc. According to Stevens (1958), more statistical operations, particularly parametric statistics, are permissible only for higher measurement levels (i.e., ratio and interval) and not for ordinal and nominal levels. Based on this understanding, ordinal scaling disallows the use of metric models for interpreting results; however, this idea is debated (Torrin & Kruschke, 2018).

Several authors have noted that although psychometric properties of scales, such as reliability and validity, are crucial to determine, other considerations are equally important (Dawis, 1987). These other factors include administration concerns and usability of the scale in the setting(s) of interest. Even though the current results suggest that DME is more appropriate than EAI for rating most features considered here, DME is known to be cumbersome to use in both clinical and research settings – it takes longer to use, requires preparation to identify moduli, and poses challenges in conveying the meaningfulness of DME scores to clients. VAS is a very popular alternative to both approaches (as used in Stipancic et al., 2023), largely due to ease of implementation and use, but its scale properties are unknown, and we need to establish the validity associated with VAS. Particularly, as related to the current work, it will be important to determine whether VAS functions more like a DME or an EAI scale, and whether construct validity is comparable for VAS and DME since DME can be used with both prothetic and metathetic dimensions.

Final thoughts from Dawis in 1987 are still applicable to the current work: "In scale construction, as in much of human endeavor, there can be no single "best" method. One method may be best for one research purpose but not for another. Purpose, context, and limitations on the researcher must be considered. Trade-offs in advantages and disadvantages seem to be the rule…" (Dawis, 1987, p. 488). Therefore, it is important to consider the context in which a measure might be used. If, for example, as the current results suggest, DME has better validity than EAI for overall speech severity, but a practicing clinician is reluctant to use DME because of the inconvenience or the lack of a standard modulus, then using an easier-to-implement scale that might not have the best validity, may be appropriate. In contrast, a researcher who is interested in making very precise, valid measurements to evaluate treatment efficacy, and has the time and resources for a more complex procedure, should consider using DME. Overall, the current findings call for rethinking the widespread use of EAI scales for rating perceptual features as part of assessment and treatment of motor speech disorders.

Limitations and Future Directions

Only construct validity was examined in this project and future work could examine other types of validity along with construct validity to determine key scale properties. Robust analyses of validity *and* reliability (both interrater and intrarater) in the same study would be beneficial for determining psychometric properties of similar types of scales since validity and reliability are separate concepts and do not necessarily vary in tandem (i.e., a scale can be reliable, but not valid and vice versa). The limitations of reliability analyses in the current work may have contributed to the unexpected findings and could be addressed in future studies. Specifically, results revealed that intrarater reliability was lower than interrater reliability, when the opposite is typically true (i.e., reliability within a rater, is typically better than reliability across a group of raters). However, the small number of data points used in the reliability analyses, the difference in statistical analyses used for calculating intra vs. interrater reliability, and an inability to determine statistical or clinically meaningful differences between reliability estimates, may have contributed to this unexpected finding. More work examining scale fit, other types of validity, and reliability of the ratings used commonly in research and in clinic is critically needed. For example, only five of the 38 possible dysarthria features from the Mayo Clinic Rating System were included in the current study, and the psychophysical properties of the majority tied to other dysarthria types remain largely unexplored. This study used inexperienced listeners to rate the dysarthric speech features of interest and future work could examine the impact of listener characteristics on validity and reliability of scaling tasks. Given that continuum type is determined by neural activation patterns and how physical stimulus properties relate to perceptual magnitude, it remains to be determined if listener experience will alter how perceptual continua are classified. Lastly, the level of measurement represented by VAS needs to be established. Some voice researchers suggest that it behaves similar to EAI, without the fixed, pre-defined points along the scale (Wuyts, De Bodt, &Van de Heyning, 1999). Other groups suggest that VAS can function as a ratio scale like DME (Price et al., 1983). Since VAS has gained popularity for research and clinical use, this is a critical next step.

Conclusions

The validity of different scaling methods depends upon the continua class of the dimensions being rated. Four of the five cardinal features of hypokinetic dysarthria (i.e., overall severity, articulatory imprecision, reduced loudness, and short rushes of speech) were determined to be prothetic dimensions best scaled using DME and only monotony behaved as a metathetic dimension suited either to EAI or DME scaling. Given the unsuitability of EAI scales for prothetic dimensions, and the cumbersome nature of the optimal alternative, DME, it is recommended that researchers and clinicians consider the purpose and context of use while also weighing the advantages and disadvantages of each of these factors. To this end, inter- and intrarater reliability are comparable between the two scaling methods when rating features of hypokinetic dysarthria included in this study.

Acknowledgements

This research was supported by the National Institutes of Health (NIH) [(R15DC016383 and R21DC019952; PI: Kuruvilla-Dugdale)]. We are grateful to the research assistants and subjects who participated in the study. Special thanks to Dahlia Cukierkorn, Lexi Jacobsmeyer, Morgan Linneweh, Ella Meier, and Anna Mae Williams for helping with data collection and analysis. We also extend our appreciation to Minguang Song from the Department of Electrical Engineering and Computer Science at the University of Missouri for adapting the MATLAB GUI for VAS (Stepp et al., 2009) to the EAI and DME scaling tasks used in this study. Data visualization was handled by Dr. Francis Smith Jr., and Dr. Jacob Oleson served as the biostatistics consultant for the project.

Data Availability Statement

Data supporting the results reported in this manuscript are available for interested researchers on request from the authors.

References

Boersma, P., & Weenink, D. (2022): Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

Bunton, K., Kent, J., Duffy, J., Rosenbek, J., & Kent, R. (2008). Listener agreement for auditory perceptual ratings of dysarthria. *Journal of Speech, Language and Hearing Research, 50*(6), 1481-1495.

Chiu, Y.-F., Neel, A., & Loux, T. (2021). Exploring the acoustic perceptual relationship of speech in Parkinson's disease. *Journal of Speech, Language and Hearing Research*, *64*, 1560-1570.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, *12*, 462-496.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*(4), 481-489.

DeMaagd, G., & Philip, A. (2015). Parkinson's disease and its management. *Pharmacy & Therapeutics*, *40*(8), 504-510.

Duffy, J. R. (2013). *Defining, understanding, and categorizing motor speech disorders.* Elsevier Mosby.

Eadie, T. L., & Doyle, P. C. (2002a). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language and Hearing Research*, *45*(6), 1088-1096.

Eadie, T. L., & Doyle, P. C. (2002b). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America*, *112*. https://doi.org/https://doi.org/10.1121/1.1518983

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.

Ho, A. K., Bradshaw, J. L., Iansek, R., & Alfredson, R. (1999). Speech volume regulation in Parkinson's disease: Effects of implicit cues and explicit instructions. *Neuropsychologia*, *37*(13), 1453-1460.

Kempster, G. B., Kistler, D. J., & Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speeech and Hearing Research*, *34*, 534-543.

Kent, R. D. (1996). Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*(3), 7-23. https://doi.org/https://doi.org/10.1044/1058-0360.0503.07

Kim, H. (1994). *Monotony of speech production in Parkinson's disease: Acoustic characteristics and their perceptual relations* (Doctoral dissertation).

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *The Journal of the Acoustical Society of America*, *100*, 1787-1795.

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America*, *96*(3), 1291-1302.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. (1993a). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, *36*, 21-40.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993b). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, *36*, 21-40.

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and co-occurence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *Journal of Speech and Hearing Disorders*, *43*(1).

Ma, A., Lau, K. K., & Thyagarajan, D. (2020). Voice changes in Parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, *72*, 1-7. https://doi.org/https://doi.org/10.1016/j.jocn.2019.12.029

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychosocial dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders*, *55*, 516-525.

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

Patel, S., Shrivastav, R., & Eddins, D. A. (2010). Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice, 24*(2), 168-177.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45–56. https://doi.org/10.1016/0304-3959(83)90126-4

Ryan, M. L. (1971). *Investigation of the psychophysical relationship of kinesthetic extent of arm movement* (Doctoral dissertation, University of British Columbia).

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.

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

Schiavetti, N., Sacco, P. R., Metz, D. E., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering serverity. *Journal of Speech and Hearing Research, 26*, 568-573.

Sewall, A., Weglarski, A., Metz, D. E., Schiavetti, N., & Whitehead, R. L. (1999). A methodolohgical control study of scaled vocal breathiness measurements. *Contemporary Issues in Communication Sciences and Disorders*, *26*, 168-172.

Siegel, S. (1956). A method for obtaining an ordered metric scale. *Psychometrika, 21*(2), 207-216.

Snijders, B. & Bosker, J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd Ed. Sage, London.

Southwood, M. H., & Weismer, G. (1993). Listener judegements of the bizarreness, acceptibility, naturalness, and normalcy of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology*, *1*, 151-161.

Stepp, C. E., Heaton, J. T., Rolland, R. G., & Hillman, R. E. (2009). Neck and face surface electromyography for Prosthetic voice control after total laryngectomy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *17*(2), 146– 155. https://doi.org/10.1109/TNSRE.2009.2017805

Stevens, S. S. (1958). Measurement and man. *Science, 127*(3295), 383-389.

Stevens, S. S. (1969). Sensory scales of taste intensity. *Perception & Psychophysics, 6*, 302-308.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* Wiley.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*, 377-411.

Stipancic, K. L., Golzy, M., Zhao, Y., Pinkerton, L., Rohl, A., & Kuruvilla-Dugdale, M. (2023). Improving Perceptual Speech Ratings: The Effects of Auditory Training on Judgments of Dysarthric Speech. *Journal of Speech, Language, and Hearing Research*, *66*(11), 4236–4258. https://doi.org/10.1044/2023_JSLHR-23-00322.

Stipancic, K. L., Palmer, K. M., Rowe, H. P., Yunusova, Y., & Green, J. R. (2021). "You say severe, I say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research, 64*, 4718-4735.

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.

Trevethan, R. (2017). Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127-143.

Whitehill, T. L., Lee, A. S. Y., & Chun, J. C. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language and Hearing Research*, *45*(1), 80-88.

Wolfe, V. I., & Ratusnik, D. L. (1988). Acoustic and perceptual measurements of roughness influencing judgements of pitch. *Journal of Speech and Hearing Disorders*, *53*, 15-22.

Yiu, E. M.-L., & Ng, C.-Y. (2004). Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clinical Linguistics and Phonetics*, *18*(3), 211-229. https://doi.org/https://doi.org/10.1080/0269920042000193599

Yorkston, K. M., Beukelman, D., & Hakel, M. (2007). *Speech Intelligibility Test (SIT) for Windows [Computer software]*. Madonna Rehabilitation Hospital.

Zraik, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language and Hearing Research*, *43*, 979-988. https://doi.org/1092-4388/00/4304-0979

Zyski, B. J., & Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders*, *20*, 367-378.