ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

"You Say Severe, I Say Mild": Toward an Empirical Classification of Dysarthria Severity

Kaila L. Stipancic,[a] Kira M. Palmer,[b] Hannah P. Rowe,[b] Yana Yunusova,[c] James D. Berry,[d] and Jordan R. Green[b]

[a] Department of Communicative Disorders and Sciences, University at Buffalo, NY [b] Department of Communication Sciences and Disorders, MGH Institute of Health Professions, Boston, MA [c] Department of Speech-Language Pathology, University of Toronto, Ontario, Canada [d] Sean M. Healey and AMG Center for ALS, Massachusetts General Hospital, Boston

ARTICLE INFO

Article History:
Received April 2, 2021
Revision received July 7, 2021
Accepted August 12, 2021

Editor-in-Chief: Bharath Chandrasekaran
Editor: Kate Bunton

https://doi.org/10.1044/2021_JSLHR-21-00197

ABSTRACT

Purpose: The main purpose of this study was to create an empirical classification system for speech severity in patients with dysarthria secondary to amyotrophic lateral sclerosis (ALS) by exploring the reliability and validity of speech-language pathologists' (SLPs') ratings of dysarthric speech.
Method: Ten SLPs listened to speech samples from 52 speakers with ALS and 20 healthy control speakers. SLPs were asked to rate the speech severity of the speakers using five response options: normal, mild, moderate, severe, and profound. Four severity-surrogate measures were also calculated: SLPs transcribed the speech samples for the calculation of speech intelligibility and rated the effort it took to understand the speakers on a visual analog scale. In addition, speaking rate and intelligible speaking rate were calculated for each speaker. Intrarater and interrater reliability were calculated for each measure. We explored the validity of clinician-based severity ratings by comparing them to the severity-surrogate measures. Receiver operating characteristic (ROC) curves were conducted to create optimal cutoff points for defining dysarthria severity categories.
Results: Intrarater and interrater reliability for the clinician-based severity ratings were excellent and were comparable to reliability for the severity-surrogate measures explored. Clinician severity ratings were strongly associated with all severity-surrogate measures, suggesting strong construct validity. We also provided a range of values for each severity-surrogate measure within each severity category based on the cutoff points obtained from the ROC analyses.
Conclusions: Clinician severity ratings of dysarthric speech are reliable and valid. We discuss the underlying challenges that arise when selecting a stratification measure and offer recommendations for a classification scheme when stratifying patients and research participants into speech severity categories.

A primary goal of motor speech assessment is to grade the severity of speech impairment. Although severity ratings are a ubiquitous part of clinical practice and a critical design feature of many studies on atypical speech, current approaches to assessing speech severity are largely untested, and there is no accepted definition or classification scheme (Stipancic et al., 2018). Dysarthric speech is routinely described using adjectival labels by both clinicians (i.e., physicians, nurses, speech-language pathologists [SLPs], and/or other rehabilitation professionals) and researchers (King et al., 2012). King et al. (2012) found that the two most frequently used informal measures for quantifying speech intelligibility by SLPs were (a) estimating the percentage of a patient's speech that was understood and (b) using adjectival labels such as "normal," "mild," "moderate," "severe," or "profound." Most dysarthria research also relies on similar adjectival categories to stratify study participants into severity groups (e.g., see Connaghan & Patel, 2017; Hustad,

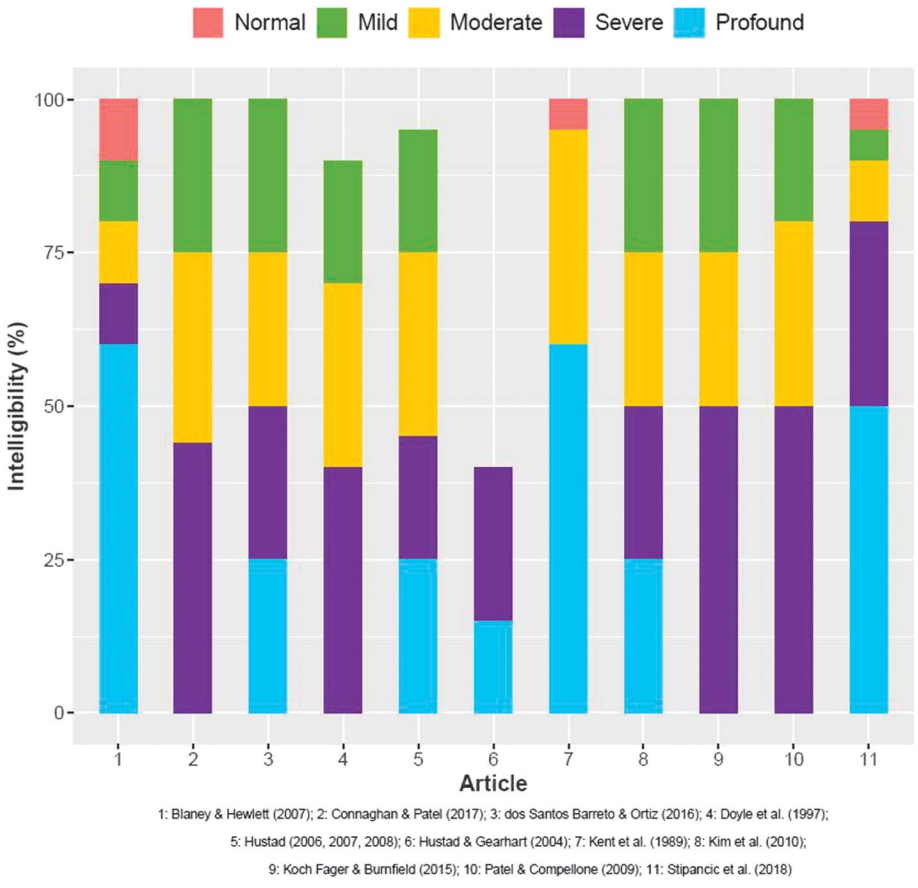
Correspondence to Jordan R. Green: jgreen2@mghihp.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

2006a, 2006b; Koch Fager & Burnfield, 2015; Stipancic et al., 2018). Research is needed to understand the psychometric qualities of these ratings (Miller, 2013; Streiner et al., 2015) and the factors that influence the perception of speech severity. This knowledge will have significant implications for improving clinical and methodological rigor (i.e., validity and reproducibility) in studies of persons with communication disorders.

In the dysarthria literature, speech intelligibility, or how understandable a speaker is to a listener (Duffy, 2013), is often used as a proxy for speech severity (Strand & Yorkston, 1994; Sussman & Tjaden, 2012). Figure 1 illustrates the variable cutoff points that investigators have used to define severity levels based on speech intelligibility scores. Inconsistencies were noted not only in the ranges of intelligibility assigned to each category but also in the methods used to quantify intelligibility across studies (see Figure 1). Furthermore, although some studies stratified speakers with dysarthria dichotomously (Hustad & Gearhart, 2004; Paja & Falk, 2012), many others used three, four, or five adjectival categories with differing criteria for category inclusion (see Blaney & Hewlett, 2007; dos Santos Barreto & Ortiz, 2015; Doyle et al., 1997; Hustad, 2006a, 2006b, 2007, 2008; Stipancic et al., 2018). Thus, for many studies, intelligibility cutoff points for severity group membership appear to have been chosen somewhat arbitrarily and based on available data.

To improve across-study reproducibility and minimize assumptions inherent to clinician-derived severity categories, some authors have used data-driven stratification approaches such as quartile or quintile criteria (i.e., Gordon-Brannan & Hodson, 2000), statistical cluster analyses (i.e., Rong, 2019), and machine classification (i.e., Wang et al., 2018). Other authors have reported their findings using several different stratification approaches (i.e., Stipancic et al., 2018). Although these methods may improve the reproducibility of severity groupings, they are also contingent on the data used for such an analysis (i.e., if two different

Figure 1. The distribution of severity categories in the literature. Percent intelligibility ranges for each clinical severity category across studies are presented. Method of intelligibility quantification differed across studies: 1: single-word multiple-choice; 2: single-word multiple-choice; 3: single-word transcription; 4: single-word multiple-choice; 5: narrative transcription; 6: sentence transcription (Speech Intelligibility Test [SIT]); 7: single-word multiple-choice; 8: single-word transcription; 9: sentence transcription (SIT); 10: single-word multiple-choice; 11: sentence transcription (SIT). Some studies did not use the exact adjectival labels used in this figure; we attempted to match the stratifications to these labels to allow for comparison across studies.



studies have different numbers of participants across the severity continuum, even data-driven approaches will provide different results). In addition, data-driven approaches disregard the role of the listener in characterizing speech, which is a key component for understanding speech impairments at an activity level (World Health Organization, 2002). To our knowledge, only two studies have attempted to examine the convergent validity of speech severity ratings in participants with impaired speech. Shriberg and Kwiatkowski (1982) investigated severity ratings of speech involvement in children with speech sound disorders. Of the several factors investigated (including percent consonants correct [PCC], suprasegmental ratings, and speech intelligibility derived from orthographic transcription), speech intelligibility and PCC contributed the most to the clinicians' ratings of speech severity, sharing 55% and 38% common variance with severity ratings, respectively. In another study, Schiavetti et al. (1983) investigated severity ratings of speakers who stutter. The authors examined categorical judgments of severity and their nonlinear relationship with scaled judgments of intelligibility. They found that the severity categories corresponded to different ranges of intelligibility; the "mild" category corresponded to 95%–100% intelligibility, or 5% of the total possible range of intelligibility, whereas the "moderate" category corresponded to 80%–95% intelligibility, or 15% of the total possible range of intelligibility. Although these two articles have provided important insights into the complex nature of speech severity ratings, their findings may not generalize to the idiosyncrasies and heterogeneity inherent to dysarthric speech. It should also be noted that authors in the voice literature have begun to examine how to stratify participants into groupings of voice severity (see, e.g., Y. W. Lee et al., 2020).

What Measure Should Be Used to Stratify Speech Severity?

Speech severity ratings have previously been found to be at least moderately correlated with intelligibility (Sussman & Tjaden, 2012; Tjaden et al., 2014; Tjaden & Wilding, 2004; Weismer et al., 2000), and there is evidence supporting the validity of stratifying speakers by intelligibility scores. For example, Hustad (2007) found that ratings of listeners' confidence in transcribing dysarthric speech (to derive intelligibility) worsened as dysarthria severity increased. Evidence, therefore, suggests that intelligibility likely accounts for a large portion of the information clinicians use to make severity judgments. However, use of intelligibility alone to define severity categories may be problematic given previous findings that many other parameters, such as speaking rate, prosody, and naturalness, can have a profound effect on perceived speech severity (Tjaden et al., 2014), such that intelligibility can be unaffected while severity is impaired. Indeed, work by Sussman and Tjaden (2012) suggested that, although the measures are related, transcription intelligibility is an imperfect proxy for overall speech severity.

Given the current literature base, it is unclear how adjectival ratings of severity compare to well-accepted metrics of speech severity. In addition, we lack information about whether speech intelligibility or another clinical severity-surrogate measure best aligns with clinician severity ratings. Although much less prevalent than studies on intelligibility as a proxy for severity, a number of recent studies used clinical judgment (Kuruvilla et al., 2012), scores on patient-reported outcomes (Allison et al., 2017; Yunusova et al., 2016), and more objective measures such as speaking rate (Shellikeri et al., 2016; Stipancic et al., 2018; Yunusova et al., 2012) to define speech severity categories. Two largely unexplored measures are intelligible speaking rate and subjective ratings of listener effort. Intelligible speaking rate is a measure of speech efficiency first described in 1981 (Yorkston & Beukelman, 1981) that has not been readopted until recently (Hustad et al., 2019; Wang et al., 2018). Since intelligible speaking rate combines two sensitive and responsive measures of speech severity (i.e., intelligibility and speaking rate; Barnett et al., 2019; Green et al., 2013; Rong, Yunusova, Wang, & Green, 2015; Rong et al., 2016), it may serve as a superior proxy for speech severity in developing a standardized categorization scheme. Furthermore, the more subjective measure of listener effort may tap into subconscious parameters that contribute to listeners' perceptions of severity. Listener effort also likely reflects a combination of parameters that contribute to perceived speech severity (Cote-Reschny & Hodge, 2010; Landa et al., 2014; Nagle & Eadie, 2012, 2018; Whitehill & Wong, 2006; Wilson et al., 2020). Indeed, intelligibility and speaking rate have previously been found to account for a significant amount of variance in listener effort (Nagle, 2015). Other factors, such as voice quality, have also been found to contribute to measures of listener effort (Whitehill & Wong, 2006). Thus, similar to intelligible speaking rate, listener effort has potential to be an informative measure that aligns with the clinician-derived severity categories; however, reliability on this subjective measure may be reduced relative to the reliability of more objective measures such as speaking rate.

In the current study, speech samples from patients with dysarthria secondary to amyotrophic lateral sclerosis (ALS) were investigated. ALS is a progressive motor neuron disease caused by the degeneration of both upper and lower motor neurons, resulting in the weakening, fasciculation, and atrophy of all the body's muscles (Turner et al., 2013). Motor neuron deterioration affects muscles that are important for speech, leading to flaccid and/or

spastic dysarthria and reduced speech intelligibility. Because this was the first investigation, to our knowledge, to explore an empirical stratification of dysarthria severity, we chose to constrain the heterogeneity of speakers by selecting individuals within one disordered population. Additionally, within this population, we included speakers across the continuum of severity.

The purpose of this study was to investigate the reliability (intrarater and interrater) and validity (construct) of clinician severity ratings of dysarthric speech. Additionally, we examined the efficacy of other commonly used clinical measures of speech (i.e., intelligibility, speaking rate, intelligible speaking rate, and listener effort) for classifying overall speech severity levels and, thus, performing as severity surrogates. We chose the four severity-surrogate measures as they have been widely used to characterize speech in individuals with ALS (Kalra et al., 2020), have been found to be sensitive to longitudinal changes in speech (Rong et al., 2016), and are relatively easy to implement across research and clinical settings. Within our experimental framework, a severity-surrogate measure was considered a valid proxy for overall severity if (a) it trended with the clinician-derived severity labels (i.e., intelligibility scores uniformly increase as adjectival severity levels increase) and (b) the adjacent severity groups were statistically different in regard to the severity-surrogate measure. Following these analyses, we used a data-driven approach to determine the optimal severity group stratification scheme and recommended a severity classification for both research and clinical purposes. To this end, our research questions were as follows:

1. What is the reliability of clinician-based adjectival ratings of speech severity?
2. What is the construct validity of clinician-based adjectival ratings of speech severity?
3. Which of the four severity surrogates explored is best suited to represent the adjectival labels of speech severity provided by the clinicians?

Method

This study was approved by the institutional review boards at Mass General Brigham, University of Nebraska, University of Toronto, and University of Texas at Dallas. Written informed consent was obtained from all participants prior to being enrolled in the study.

Participants

Speakers

Participants included 52 speakers with ALS and 20 neurologically healthy control speakers with recordings from the Speech Intelligibility Test (SIT; Yorkston et al., 2007) from a larger study of bulbar motor impairment in ALS (Green et al., 2013). The participants with ALS had been previously diagnosed with ALS by a neurologist following El Escorial criteria (Brooks et al., 2000). Site of ALS onset (i.e., bulbar vs. spinal) varied among participants (see Table 1). All participants spoke English as their primary language; had no history of other neurological disorders; and were required to have adequate hearing, vision, and literacy skills with which to hear, see, and read stimuli. The speech intelligibility of all speakers ranged from 0% to 100% (M = 63.95%, SD = 33.60) based on the SIT as transcribed by a trained research assistant (procedure described below). Demographic information for the speakers is provided in Table 1.

Listeners

Ten SLPs from the Boston area, all of whom indicated that their caseloads have included patients with dysarthria (etiologies of dysarthria included stroke, traumatic brain injury, ALS, primary progressive aphasia, other neurodegenerative diseases, cerebral palsy, and other pediatric neurodevelopment disorders), were recruited to participate in this study. The only inclusion criterion for listeners was that they were licensed SLPs who had some experience with assessing/treating patients with dysarthria. Because adjectival labels of dysarthria severity are often used by clinicians (King et al., 2012), we were particularly interested in how trained clinicians assign these labels to speakers with ALS and how the severity-surrogate measures contribute to clinicians' perceptions of speech severity. The selection of only clinicians who have experience with dysarthria has been implemented in previous studies (Gurevich & Scamihorn, 2017; King et al., 2012) and would, presumably, constrain some variability in the listener

Table 1. Demographic information of the speakers included in the study.

Variable	ALS participants	Control participants
Total *N* (number of men)	52 (30)	20 (9)
Mean age in years (*SD*)	58.73 (9.37)	63.21 (9.98)
Site of onset = *N*	Spinal = 25 Bulbar = 21 Mixed = 3 Unknown = 3	
Mean % intelligibility (*SD*) [range]	58.85 (32.24) [0–100]	99.86 (0.61) [97.27–100]
Mean speaking rate in WPM (SD) [range]	94.63 (46.11) [19.33–196.43]	176.00 (21.18) [137.2–216.26]

Note. ALS = amyotrophic lateral sclerosis; *SD* = standard deviation; WPM = words per minute.

group (i.e., instead of also including some SLPs with no experience with dysarthria). All 10 SLPs had received a master's degree in an SLP program, three had received a PhD, and two were in a PhD program at the time of data collection. Each clinician completed a demographic questionnaire prior to participating in study procedures. On average, clinicians had been practicing as SLPs for 6.6 years (range: 2–14 years, *SD* = 5.1), had been practicing with patients with dysarthria for a mean of 6.25 years (range: 0.5–14 years, *SD* = 5.2), and estimated the average percentage of individuals in their monthly caseloads who received services for dysarthria to be ~25% (range: 1%–70%, *SD* = 21.65%). Demographic information for each SLP is provided in Table 2.

Procedure

Speech Samples

During data collection sessions, the SIT (Yorkston et al., 2007) was administered to calculate speech intelligibility of the speakers. Each participant was presented with a unique set of 11 computer-generated sentences to read aloud, which was recorded using a head-mounted microphone. Using these recordings, a research assistant, who was unfamiliar with participants' diagnoses and speech severity, orthographically transcribed the sentences. These data were collected over 15 years in four different research labs for a large, longitudinal study. As such, different research assistants transcribed these SIT recordings as they were collected to provide an overall indication of speech severity. We refer here to a single research assistant completing these transcriptions because one research assistant transcribed each of the collected SIT recordings. Percent intelligibility was calculated and averaged across the sentences as the total number of words correctly transcribed, divided by the total number of words produced, multiplied by 100. This listening protocol has been used in several previous studies (Rong, Yunusova, Wang, & Green, 2015; Stipancic et al., 2018; Yunusova et al., 2010, 2011). The initial transcription performed by a trained research assistant was completed to provide baseline demographic information about the speakers to create the following listening task for completion by the SLPs. For this portion of the study, transcriptions were only completed by one judge; strong reliability of transcription on the SIT has been previously reported in a work from our group (Stipancic et al., 2018). The transcription by the research assistant was only used to ensure that our data set contained a diverse sample of speech severities, which was corroborated by the SLPs' ratings (see range of severity ratings in the Results section). Throughout the rest of the article, "speech sample" refers to a set of 11 SIT stimuli from a single speaker.

Listening Task Preparation

A visual representation of the organization of the listening task can be seen in Table 3. Using percent intelligibility calculated from the research assistant's transcriptions of the speech samples, speakers with ALS were broken into five lists based on quintiles of intelligibility (i.e., 0%–20%, 21%–40%, 41%–60%, 61%–80%, and 81%–100%). Each list contained four speakers with ALS per quintile of intelligibility. Overall, there were a total of 20 different speakers with ALS per listener (see Table 3). To ensure that there were 20 speakers in each list, a subset of speech samples was repeated in the data set. In addition, to ensure an even distribution of intelligibility across the speakers with ALS, a subset of speech samples from multiple data collection sessions for the same speaker was included in the data set (i.e., as a speaker progressed in the disease, their intelligibility declined, and thus, the same speaker at a different data collection session fell into a different quintile of intelligibility); however, the same speaker was never repeated within a list (i.e., a listener

Table 2. Demographic information of the listeners included in the study.

Listener number	Number of years practicing	Current work setting	Number of years with patients with dysarthria	Percent monthly caseload of patients with dysarthria	List assigned for listening task
1	12	Acute care	12	40	List 1
2	4	Acute care, rehabilitation, outpatient	4	25	List 1
3	2	Outpatient	2	70	List 2
4	14	Other (research), acute care	14	1	List 2
5	3	Acute care	3	10	List 3
6	2	Acute care, outpatient	2	10	List 3
7	14	Other (research)	14	—	List 4
8	2	Outpatient, acute care	0.5	40	List 4
9	4	Outpatient, rehabilitation	4	10	List 5
10	9	Outpatient	7	20	List 5

Note. The em dash indicates data not reported. This listener forgot to answer the question about the percentage of monthly caseload consisting of patients with dysarthria; however, the clinician did endorse that they had been working with patients with dysarthria for 14 years.

Table 3. Listening task design.

Samples in listening task		List 1	List 2	List 3	List 4	List 5	Total *N*
ALS speakers	Intelligibility: 0%–20%	4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
	Intelligibility: 21%–40%	4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
	Intelligibility: 41%–60%	4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
	Intelligibility: 61%–80%	4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
	Intelligibility: 81%–100%	4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
Control speakers		4 speakers	4 speakers	4 speakers	4 speakers	4 speakers	20 speakers
Reliability, speakers with ALS repeated		2 speakers	2 speakers	2 speakers	2 speakers	2 speakers	10 speakers
Total number of samples rated		26 samples	26 samples	26 samples	26 samples	26 samples	130 samples
Completed by SLP listeners		SLP Listeners 1 and 2	SLP Listeners 3 and 4	SLP Listeners 5 and 6	SLP Listeners 7 and 8	SLP Listeners 9 and 10	10 listeners

Note. We created five lists for the speech-language pathologists (SLPs) to judge. Each list included four speakers with amyotrophic lateral sclerosis (ALS) within each quintile of intelligibility and four healthy control speakers. For the purposes of speaker selection, intelligibility percentages were used to ensure an even range of intelligibility across all lists that were heard by the SLPs. Additionally, two randomly selected speakers with ALS were repeated in each list for reliability calculations. Therefore, each SLP judged a total of 26 samples. Only 52 speakers with ALS participated, but due to multiple longitudinal sessions from individual participants, we had a total sample of 100 Speech Intelligibility Tests.

never heard the same speaker with ALS twice). For speakers in which multiple longitudinal sessions were included in the data set, there was an average of 61 days between data collection sessions (range: 21–127 days, SD = 35). One speaker had five sessions included, three speakers had four sessions included, and three had two sessions included. In addition, four healthy control speakers were assigned to each listener, and 10% of the speakers with ALS (i.e., two speakers) were chosen at random and were repeated in order to calculate intrarater reliability. Therefore, each list contained a total of 26 speech samples, which each consisted of 11 sentences. Thus, listeners each heard 286 sentences in total. Each of the 10 SLPs listened to one of these lists, with two listeners per list to allow for the calculation of interrater reliability. SLPs were consecutively assigned to lists as they were recruited (i.e., the first two SLPs who were recruited were assigned to List 1, followed by the next two SLPs who were assigned to List 2, etc.). Order effects were partially controlled for by presenting the list of speakers in different orders for each of the two listeners assigned to each list (i.e., the first listener heard them in random order, and the second listener heard the latter half of the speakers first). It should be noted that several instances of repeated SIT stimuli (~10% of the sentences; the most times a single sentence was repeated for a single listener was three) were randomly assigned across listeners, creating the potential for bias if SLPs transcribed stimuli they heard previously; however, there was a large number of stimuli and speakers and also a random assignment of these variables across listeners.

Listening Task Procedure and Measures

The listening study was administered electronically via REDCap (Harris et al., 2009), an online survey platform. Before accessing the REDCap survey, SLP listeners used an online tool to individually control for appropriate listening volume (Miracle Ear, n.d.). Listeners were instructed to wear headphones and listen to brief scenarios mixed with background noise, after which they were given multiple-choice questions to answer pertaining to the scenario presented. Listeners were instructed to adjust the volume on their computer until they could achieve 100% accuracy on the questions provided. Listeners were subsequently instructed to refrain from adjusting the volume for the remainder of the experiment. All speech samples heard by the listeners were normalized for amplitude. As a brief overview, clinicians first heard all 11 SIT sentences from a single speaker and were asked to rate overall speech severity. Then, they were provided with each sentence of the SIT from the same speaker and asked to transcribe each to the best of their ability. Immediately following this, the clinician was asked to rate their listening effort for that single speaker. The clinician then listened to the next speaker's speech sample and followed the identical procedure. All 26 samples for each clinician were completed in the same manner. Details on each measure are provided below.

Clinician severity ratings. Following the volume-adjustment procedures described above, listeners were instructed to listen to all 11 SIT sentences from one speaker, after which they were asked to rate the speaker's speech severity ("Please indicate the severity of speech for this individual").

Listeners were given five response options: normal, mild, moderate, severe, and profound. In contrast to previous work (Sussman & Tjaden, 2012), no instructions were provided regarding which speech features listeners should focus on (i.e., intelligibility, imprecision). We made this decision because we wanted clinicians to use the features they would naturally use to inform their severity ratings.

Severity-Surrogate Measures

Speech Intelligibility

To measure speech intelligibility, the listeners were asked to provide an orthographic transcription of each speaker's sentences. They were given the following instructions: "Please write, word for word, what you hear. You can listen to each audio file twice." The listeners heard each SIT sentence and were given an opportunity to type their response following each sentence recording. These transcriptions were later translated into intelligibility percentages for each participant following the SIT protocol (Yorkston et al., 2007), consistent with procedures used in previous work (Allison et al., 2017; Stipancic et al., 2018; sentence intelligibility = number of words correctly transcribed/total number of words produced × 100).

Speaking Rate

Average speaking rate was calculated for each speaker from their recordings of the SIT sentences. The number of words produced divided by the total duration of each sentence supplied a rate for each sentence in words per minute (WPM). The rates from all 11 sentences were then averaged to create an overall speaking rate for each speaker (Stipancic et al., 2018).

Intelligible Speaking Rate

The number of intelligible words per minute (IWPM) was also calculated for each participant. Percent intelligibility from the listeners' transcriptions was divided by 100 and then multiplied by the speaking rate (in WPM) to derive an average intelligible speaking rate (in IWPM).

Listener Effort

To measure listener effort, the listeners were asked about their perceived effort of listening to each speaker. Answers were recorded using a visual analog scale (VAS), which consisted of a vertical line, without tick marks, with end points labeled "very" and "not at all" (Picou et al., 2017). Listeners were asked to use the VAS to answer the following question: "How effortful was it for you to understand? Remember, we are asking how hard you worked, not how well you did." The position of the slider on the VAS was automatically translated into a score from 0 (at the "not at all" end point) to 100 (at the "very" end point). This score was later subtracted from 100 for ease of interpretation and alignment with other measures used in this study. Therefore, scores closer to 100 indicated the least effort, and scores closer to 0 indicated the most effort.

Statistical Analysis

To prepare the data for statistical analysis, we dummy-coded the clinician severity ratings and averaged across the listeners who heard the same speakers—when severity ratings did not agree across listeners (for 33 speakers), we rounded up to the more severe rating. All discrepancies between SLPs differed by adjacent categories (i.e., one SLP rated a speaker as mild, but another SLP rated the same speaker as moderate). Thus, each speaker was assigned a single severity category for the remainder of the analyses. Furthermore, scores on each of the severity-surrogate measures were averaged across the listeners who rated the same speakers to derive single scores for intelligibility, speaking rate, intelligible speaking rate, and listener effort for each speaker. All statistical analyses were completed in R (R Development Core Team, 2013).

Reliability

Separate reliability statistics were calculated for the clinician severity ratings and for each of the severity-surrogate measures. The statistics used for intrarater and interrater reliability, which can be seen in Table 4, included intraclass correlation coefficients (ICCs; psych package in R [Revelle, 2018]), percent agreement, and weighted kappa. The statistics used for each measure differed based on the type of data (i.e., continuous vs. categorical) and the number of data points available. For example, intrarater reliability for clinician severity ratings included only four speakers' severity ratings (which resulted in only four scores), while intrarater reliability transcription included 11 sentences across two speakers (which resulted in 22 scores).

Intrarater reliability. Listener responses and measurements from the two repeated samples were compared to the same two samples heard earlier in the task. Because of the small number of repeated samples for the clinician severity ratings, percent agreement for the two repeated samples was calculated for each listener, and an average was calculated across all 10 listeners. A two-way random single measure for consistency/absolute agreement ICC (2, 1) was used to assess intrarater reliability for intelligibility. For speaking rate and intelligible speaking rate, 10% of the data (12 speech samples = 132 sentences) were measured by the same analyst twice, and ICCs (2, 1) were used to assess the relationship between the first and second measurements. Again, because of the small number of repeated samples for listener effort, percent agreement for the two repeated samples was calculated for each listener, and an average was calculated across all 10 listeners.

Table 4. Reliability of measures.

Type of reliability	Clinician severity ratings	Intelligibility	Speaking rate	Intelligible speaking rate	Listener effort
Intrarater	.94 (across all listeners) *% agreement*	.91 (.88–.92, $p < .001$) *ICC (2, 1)*	1.00 (1.00–1.00, $p < .001$) *ICC (2, 1)*	.99 (.98–1.00, $p < .001$) *ICC (2, 1)*	.97 (across all listeners) *% agreement*
Interrater	.91 (across all listeners) *Weighted kappa*	.94 (.92–.96, $p < .001$) *ICC (2, 1)*	.99 (.94–1.00, $p < .001$) *ICC (2, 1)*	.98 (.95–.99, $p < .001$) *ICC (2, 1)*	.93 (.90–.95, $p < .001$; across all listeners) *ICC (2, 1)*

Note. The method used for each reliability statistic appears in italics. All reliability statistics were significant at $p < .001$. Confidence intervals are included in the parentheses after the correlation statistics. ICC = intraclass correlation coefficient.

Interrater reliability. In order to determine interrater reliability, responses and measurements from the listeners who heard the same list of speech samples (i.e., Listeners 1 and 2; see Table 3) were compared. Because clinician severity ratings were categorical data, interrater reliability was calculated using weighted kappa. For intelligibility, reliability was calculated with an ICC (2, 1). For speaking rate and intelligible speaking rate, 10% of the data (12 speech samples = 132 sentences) were measured by a second analyst, and ICCs (2, 1) were used to assess the relationship between the first and second analysts' measurements. For listener effort, interrater reliability was also calculated with an ICC (2, 1).

Validity

Convergent validity. Convergent validity is a subtype of construct validity and is estimated by comparing the output of a new scale to that of a known scale that measures the same construct (Streiner et al., 2015). We assessed convergent validity of the clinician severity ratings with correlation coefficients to evaluate the association between the clinician severity ratings and the severity-surrogate measures (i.e., intelligibility, speaking rate, intelligible speaking rate, and listener effort). Spearman correlations were calculated between the clinician severity ratings and the severity-surrogate measures because severity ratings were categorical data, and Pearson product correlation coefficients were calculated between the severity-surrogate measures because all severity-surrogate measures were continuous data. One-way analysis of variance tests were used to determine whether there were between-groups differences in each of the four severity-surrogate metrics across the five severity categories. Average scores for intelligibility, speaking rate, intelligible speaking rate, and listener effort were fit as a function of severity category (normal, mild, moderate, severe, and profound) to evaluate statistical differences between the categories. Post hoc contrasts (Tukey's honestly significant difference) with Bonferroni corrections for multiple comparisons were conducted for statistically significant main effects. Post hoc contrasts were evaluated for significance at $p < .005$.

Known-groups validity. Construct validity by known groups was examined with Cohen's *d* effect sizes using the effsize package in R (Torchiano, 2020). We calculated effect sizes for each severity-surrogate measure between all combinations of the groups to see which severity-surrogate measures would demonstrate the largest group differences. In particular, groups adjacent to one another in ranking (i.e., mild and moderate, moderate and severe, etc.) were of greatest interest because, in theory, they should be the most difficult for clinicians to distinguish between.

Classification Scheme for Dysarthria Severity

Receiver operating characteristic (ROC) curves, using the ROCR package in R (Sing et al., 2005), were used to identify optimal cutoff points for the clinician-rated severity groups for each severity-surrogate measure. To do this, we made each cutoff point between categories a binary decision. Four ROC curves were created for each severity-surrogate measure. Thus, we calculated four optimal cutoff points that maximized sensitivity and specificity for defining the boundaries between (a) the normal and mild groups, (b) the mild and moderate groups, (c) the moderate and severe groups, and (d) the severe and profound groups. We also calculated the area under the curve (AUC) for each cutoff point along with sensitivity, specificity, and accuracy. We used these cutoff points to provide a range of values for each severity-surrogate measure within each clinician-rated severity category.

Results

Reliability

Results of the reliability analyses are displayed in Table 4.

Intrarater Reliability

Across all listeners, average percent agreement for the clinician severity ratings was .94. The ICC (2, 1) for transcription reliability was .91 ($p < .001$), that for speaking

rate was 1.00 ($p < .001$), and that for intelligible speaking rate was .99 ($p < .001$). Across all listeners, average percent agreement for listener effort was .97. These statistics indicate excellent intrarater reliability for all measures.

Interrater Reliability

The weighted kappa for the clinician severity ratings across all listeners was .91. The ICCs for transcription reliability ranged from .92 to .96 ($p < .001$) for all listener pairs, with an average ICC of .94. The ICCs for speaking rate and intelligible speaking rate ranged from .94 to 1.00 ($p < .001$) with average ICCs of .99 and .98, respectively. Across all listener pairs, the ICCs for listener effort ranged from .90 to .95 ($p < .001$) with an average ICC of .93. These statistics indicate excellent interrater reliability for all measures.

Validity

Convergent Validity

Results from the correlation analyses are presented in Table 5. All correlations between the clinician severity ratings and the severity-surrogate measures were significant ($p < .001$), as were correlations between all the severity-surrogate measures ($p < .001$).

Boxplots for each severity-surrogate measure across the clinician severity ratings are presented in Figure 2. On average, for intelligibility (see Figure 2A), the normal group had the highest intelligibility ($N = 23$, $M = 97.8\%$, $SD = 2.29$), followed by the mild group ($N = 12$, $M = 91.2\%$, $SD = 6.65$), the moderate group ($N = 16$, $M = 75.6\%$, $SD = 16.6$), the severe group ($N = 21$, $M = 55.6\%$, $SD = 18.2$), and the profound group ($N = 21$, $M = 22.4\%$, $SD = 18.6$). There was a significant main effect of clinician severity rating on intelligibility, $F(4, 88) = 89.19$, $p < .001$. Post hoc tests revealed all contrasts were significant at $p < .001$ except the normal–mild contrast ($p = .70$) and the mild–moderate contrast ($p = .04$), which were not significant.

For speaking rate (see Figure 2B), the normal group had the fastest speaking rate ($N = 23$, $M = 173$ WPM, $SD = 23.1$), followed by the mild group ($N = 12$, $M = 135$ WPM, $SD = 36.1$), the moderate group ($N = 16$, $M = 92.4$ WPM, $SD = 39.6$), the severe group ($N = 21$, $M = 70.91$ WPM, $SD = 35.2$), and the profound group ($N = 21$, $M = 64.8$ WPM, $SD = 24.0$). There was a significant main effect of clinician severity rating on speaking rate, $F(4, 88) = 45.74$, $p < .001$. Post hoc tests revealed that the mild–moderate contrast was significant at $p = .0047$. The normal–mild ($p = .009$), moderate–severe ($p = .24$), moderate–profound ($p = .07$), and severe–profound ($p = .97$) contrasts were not significant. All other nonadjacent contrasts were significant at $p < .001$.

For intelligible speaking rate (see Figure 2C), the normal group had the fastest intelligible speaking rate ($N = 23$, $M = 170$ IWPM, $SD = 24.2$), followed by the mild group ($N = 12$, $M = 124$ IWPM, $SD = 36.2$), the moderate group ($N = 16$, $M = 66.3$ IWPM, $SD = 29.7$), the severe group ($N = 21$, $M = 39.6$ IWPM, $SD = 23.9$), and the profound group ($N = 21$, $M = 14.0$ IWPM, $SD = 11.7$). There was a significant main effect of clinician severity rating on intelligible speaking rate, $F(4, 88) = 135.3$, $p < .001$. Post hoc tests revealed all contrasts were significant at $p < .001$ except the moderate–severe ($p = .015$) and severe–profound ($p = .011$) contrasts, which were not significant.

Lastly, for listener effort (see Figure 2D), listeners rated the least amount of effort on the VAS for the normal group ($N = 23$, $M = 96.98$, $SD = 2.81$), followed by the mild group ($N = 12$, $M = 70.7$, $SD = 17.5$), the moderate group ($N = 16$, $M = 35.9$, $SD = 17.6$), and the severe group ($N = 21$, $M = 20.3$, $SD = 12.8$), and the most effort for the profound group ($N = 21$, $M = 4.5$, $SD = 4.39$). There was a significant main effect of clinician severity rating on listener effort, $F(4, 88) = 220.4$, $p < .001$. Post hoc tests revealed all contrasts were significant at $p < .001$.

Known-Groups Validity

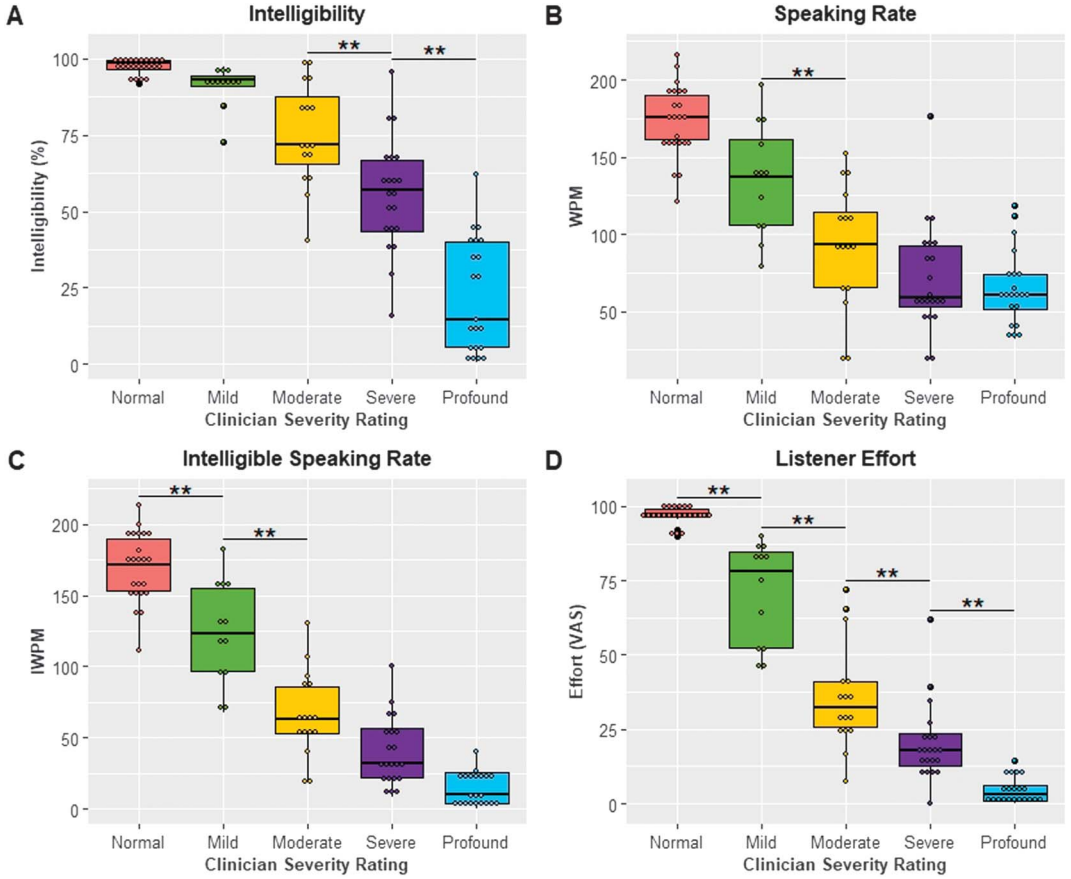
Effect sizes between all groups are displayed in Table 6, with the most-difficult-to-distinguish contrasts (i.e., adjacent severity categories) delineated with a superscript lowercase "a". All effect sizes were large (Cohen, 1988), with the exception of speaking rate between the severe and profound categories, which was considered small. For the normal–mild contrast, listener effort had the largest effect size ($d = 2.54$), followed by intelligible speaking rate

Table 5. Correlations of clinician severity ratings with severity-surrogate measures.

Variable	Intelligibility	Speaking rate	Intelligible speaking rate	Listener effort
Clinician severity rating (Spearman correlations)	−.89 ($p < .001$)	−.70 ($p < .001$)	−.87 ($p < .001$)	−.89 ($p < .001$)
Intelligibility (Pearson correlations)	—	.63 ($p < .001$)	.84 ($p < .001$)	.84 ($p < .001$)
Speaking rate (Pearson correlations)	—	—	.89 ($p < .001$)	.78 ($p < .001$)
Intelligible speaking rate (Pearson correlations)	—	—	—	.89 ($p < .001$)

Note. Spearman correlations were completed between the clinician severity ratings and the severity-surrogate measures. Pearson correlations were completed between the severity-surrogate measures. All correlations were significant at $p < .001$.

Figure 2. Severity-surrogate measures across clinician-rated severity categories. All nonadjacent comparisons were significant at $p < .005$ except the moderate–profound contrast in Panel B, which was not significant ($p = .07$). Only significant adjacent comparisons are marked in the figure with $^{**}p < .005$. IWPM = intelligible words per minute; VAS = visual analog scale; WPM = words per minute.



($d = 1.58$) and intelligibility ($d = 1.56$). Speaking rate had the smallest effect size for the normal–mild contrast ($d = 1.35$). For the mild–moderate contrast, listener effort had the largest effect size ($d = 1.98$), followed by intelligible speaking rate ($d = 1.77$), intelligibility ($d = 1.19$), and speaking rate ($d = 1.12$). For the moderate–severe contrast, intelligibility had the largest effect size ($d = 1.12$), followed by listener effort ($d = 1.04$), intelligible speaking rate ($d = 1.01$), and speaking rate ($d = 0.58$). For the severe–profound contrast, intelligibility had the largest effect size ($d = 1.80$), followed closely by listener effort ($d = 1.64$) and intelligible speaking rate ($d = 1.36$). Again, speaking rate had the smallest effect size for this contrast ($d = 0.20$).

Classification Scheme for Dysarthria Severity

The optimal cutoff points from the ROC curve analyses that maximize sensitivity and specificity, along with AUCs and specificity, sensitivity, and accuracy values for each severity-surrogate measure, are presented in Table 7. AUCs and specificity, sensitivity, and accuracy values for listener effort were the largest compared to all the other measures with all values being .81 or greater, indicating good diagnostic accuracy. Values were lowest for speaking rate, with a few exceptions (i.e., specificity was lower for intelligibility for the moderate–severe comparison and for intelligible speaking rate for the mild–moderate comparison), ranging from .48 to .83, indicating poorer diagnostic accuracy than the other severity surrogates. Values for intelligibility ranged from .69 to .95, and those for intelligible speaking rate ranged from .67 to .96, indicating adequate diagnostic accuracy. Table 8 presents the proposed classification scheme for each clinical severity measure based on the cutoff points obtained from the ROC analyses.

Discussion

Clinician Ratings of Dysarthria Severity Are Valid and Reliable

The current findings demonstrated that the adjectival labels of "normal," "mild," "moderate," "severe," and

Table 6. Cohen's *d* effect sizes with 95% confidence intervals in brackets, between each severity group for each severity-surrogate measure.

Measure	Mild	Moderate	Severe	Profound
Intelligibility				
Normal	1.56 [0.74, 2.37][a]	2.10 [1.28, 2.92]	3.32 [2.38, 4.26]	5.81 [4.42, 7.20]
Mild	—	1.19 [0.34, 2.04][a]	2.34 [1.40, 3.28]	4.44 [3.10, 5.78]
Moderate	—	—	1.12 [040, 1.84][a]	2.97 [2.00, 3.94]
Severe	—	—	—	1.80 [1.06, 2.54][a]
Speaking rate				
Normal	1.35 [0.55, 2.14][a]	2.62 [1.72, 3.51]	3.47 [2.51, 4.43]	4.61 [3.44, 5.77]
Mild	—	1.12 [0.28, 1.97][a]	1.81 [0.95, 2.68]	2.44 [1.48, 3.40]
Moderate	—	—	0.58 [–0.11, 1.27][a]	0.88 [0.17, 1.58]
Severe	—	—	—	0.20 [–0.42, 0.83][a]
Intelligible speaking rate				
Normal	1.59 [0.76, 2.40][a]	3.89 [2.78, 5.00]	5.40 [4.09, 6.71]	8.06 [6.22, 9.90]
Mild	—	1.77 [0.85, 2.70][a]	2.93 [1.88, 3.97]	4.68 [3.29, 6.06]
Moderate	—	—	1.01 [0.29, 1.72][a]	2.45 [1.56, 3.34]
Severe	—	—	—	1.36 [0.67, 2.05][a]
Listener effort				
Normal	2.54 [1.59, 3.49][a]	5.35 [3.96, 6.75]	8.45 [6.53, 10.37]	25.36 [19.87, 30.84]
Mild	—	1.98 [1.03, 2.94][a]	3.45 [2.31, 4.58]	6.02 [4.34, 7.71]
Moderate	—	—	1.04 [0.32, 1.76][a]	2.62 [1.71, 3.54]
Severe	—	—	—	1.64 [0.92, 2.37][a]

Note. Cells with the superscript lowercase "a" indicate severity categories next to each other (i.e., most-difficult-to-distinguish groups).

"profound" to describe dysarthria severity have strong construct validity. Convergent validity of clinician severity ratings was demonstrated by robust associations with other widely used measures of speech motor severity (i.e., speech intelligibility, speaking rate, intelligible speaking rate, and listener effort). Additionally, all severity-surrogate measures followed the expected direction across severity groups (i.e., average intelligibility was the highest for the normal group, followed by the mild, moderate, severe, and profound groups), providing further evidence of convergent validity. Known-groups validity was demonstrated by large effect sizes between all severity groups—even those that are most difficult to distinguish (i.e., normal–mild, mild–moderate)—on each of the four severity-surrogate measures.

Our study also found strong intrarater and interrater reliability for the clinician severity ratings, which was

Table 7. Optimal cutoff points and diagnostic power values for each severity-surrogate measure.

Clinician severity rating	Optimal cutoff point	AUC	Specificity	Sensitivity	Accuracy
Intelligibility					
Mild	94.11	.92	.75	.96	.88
Moderate	84.18	.77	.69	.92	.79
Severe	69.39	.81	.86	.69	.78
Profound	45.02	.90	.95	.71	.83
Speaking rate					
Mild	158.29	.81	.75	.83	.80
Moderate	117.25	.77	.75	.67	.71
Severe	86.97	.70	.71	.69	.70
Profound	60.51	.48	.52	.57	.54
Intelligible speaking rate					
Mild	135.50	.83	.67	.96	.86
Moderate	93.52	.91	.88	.83	.86
Severe	50.56	.76	.71	.81	.76
Profound	27.90	.85	.95	.67	.81
Listener effort					
Mild	90.5	1.00	1.00	.96	.97
Moderate	43.50	.93	.81	1.00	.89
Severe	24.25	.81	.81	.81	.81
Profound	11.75	.93	.95	.86	.90

Note. The "normal" severity group was not included in this table—this group did not need a cutoff point as it would consist of any values higher than the cutoffs for the "mild" severity group. AUC = area under the curve.

Table 8. Based on the optimal cut-points identified by the receiver operating characteristic curves, this table provides recommended speech severity groupings for both clinical and research purposes for each of the four severity-surrogate measures considered in this study.

Clinician severity rating	Intelligibility	Speaking rate	Intelligible speaking rate	Listener effort (VAS)
Normal	> 94%	> 158 WPM	> 136 IWPM	> 90
Mild	85%–94%	118–158 WPM	94–135 IWPM	44–90
Moderate	70%–84%	88–117 WPM	52–93 IWPM	25–43
Severe	45%–69%	61–87 WPM	28–51 IWPM	12–24
Profound	< 45%	< 60 WPM	< 28 IWPM	< 12

Note. VAS = visual analog scale; WPM = words per minute; IWPM = intelligible words per minute.

comparable to reliability for the four severity-surrogate measures. Reliability for the severity ratings, while similar to reliability of the other metrics reliant on listener judgment (i.e., transcription intelligibility and listener effort), was slightly lower than reliability for the metrics involving speaking rate (i.e., speaking rate and intelligible speaking rate). The reliability statistics reported here are consistent with those reported in previous studies (Bunton et al., 2001; Hustad, 2006a, 2006b; Keintz et al., 2007; Stipancic et al., 2018; Xue et al., 2020; Yorkston & Beukelman, 1978, 1981).

Taken together, our results suggest that clinician-based adjectival ratings of dysarthria severity, as one of the most commonly used measures of speech severity (King et al., 2012), are a valid and reliable method for indexing known clinical measures of speech severity in persons with ALS, at least for the five-category approach we implemented. However, our findings also support the efficacy of other methods, such as cutoff points of intelligibility, speaking rate, intelligible speaking rate, and listener effort, for providing alternative means of severity stratification in research or clinical settings and offer insights into contributors to clinician severity judgments. The cutoff points provided in this article were derived from severity ratings made by experienced clinicians and, thus, may differ if the raters/listeners are naive listeners.

A Classification Scheme for Dysarthria Severity

The importance of a universal language and standardized method for dysarthria severity stratification across research studies has long been acknowledged. Establishing an empirical stratification method and standard cutoff points is critical to the replicability of future studies. Furthermore, a standardized method will reduce the bias that is inherently introduced when arbitrarily selecting cutoff points for different measures. A thorough review of previous literature, however, revealed significant disparities in the classification of dysarthria severity and an overall lack of consistency across studies (see Figure 1). These inconsistencies and the findings from this study beg the question: What is the *best* measure to use to stratify research participants or clinical patients into severity categories? The current work answers this question by evaluating the severity-surrogate measures to find the one(s) that best align with the clinician severity ratings and, thus, may contribute the most to clinicians' perception of speech severity.

Although our findings validate the use of adjectival labels for dysarthria severity stratification, the current study also presents evidence in favor of four other measures: intelligibility, speaking rate, intelligible speaking rate, and listener effort. Although reliability was comparable for all of the measures explored, it was slightly lower for the more subjective measure of intelligibility derived from transcription. However, listener effort, which is also a highly subjective measure, had surprisingly strong reliability. Based on the high reliability and strong construct validity, listener effort was objectively superior to the other severity-surrogate measures analyzed in this study for severity stratification and most closely aligned with the clinician severity ratings. We propose that each of the severity-surrogate measures explored might provide distinct information (discussed in the following sections) about the participants' speech characteristics and at least partially contribute to clinicians' ratings of severity. We, therefore, provided cutoff points (see Table 8) for each of the severity-surrogate measures that can be used in research and clinical settings for classifying dysarthria severity in individuals with neurodegenerative disease similar to ALS. Again, it should be noted that the cutoff points provided may differ if the listeners employed are naive listeners, rather than the experienced clinicians used in the current study. The following sections will discuss our findings regarding each of these four measures and offer recommendations for when and how to employ these measures for severity stratification.

Intelligibility

Intelligibility was found to be statistically different between some of the clinician-derived severity groups, demonstrating convergent validity of the clinician severity rating. However, intelligibility did not yield a statistical difference between the normal and mild or between the mild and moderate severity groups. This finding was not surprising given previous longitudinal work that has explored

the rate of progression across measures of speech severity in individuals with ALS and found little effect of mild speech abnormalities on intelligibility (Rong, Yunusova, & Green, 2015). Indeed, intelligibility is known to lack sensitivity for early detection of neurodegenerative disease, as other aspects of speech, such as speaking rate (Rong, Yunusova, Wang, & Green, 2015), tend to decline earlier in the disease. A potential reason for this lack of sensitivity may be, in part, because speech abnormalities that occur in the early stages of disease progression are still within the normal range of variation (Cooke, 2006), resulting in unaffected intelligibility. Thus, despite the promising construct validity found for intelligibility and its frequent use as a proxy for speech severity, intelligibility may lack sensitivity to subtle speech changes. The strong relationship between clinician severity ratings and intelligibility is consistent with previous work in dysarthria that found VAS and direct magnitude estimation (DME) ratings of speech severity to be highly correlated with measures of transcription-derived intelligibility (Sussman & Tjaden, 2012; Tjaden et al., 2014). Interestingly, the derived cutoff points yielded similar severity stratifications to those used in previous studies that selected intelligibility cutoffs for severity groups based on clinical acumen (see Stipancic et al., 2018). This result further highlights that intelligibility plays a large role in how clinicians judge speech severity. However, we recommend that researchers avoid using intelligibility to distinguish between normal, mild, and moderate severity groups. In general, the absence of sensitivity in the early stages of disease progression suggests that intelligibility may not be the best measure for five-group stratification (i.e., "normal," "mild," "moderate," "severe," and "profound"). Intelligibility did, however, perform very well for discriminating between more severely impaired groups of participants, with the largest effect sizes between moderate and severe groups and between severe and profound groups. It should also be noted that the cutoff points provided here were calculated on intelligibility derived from orthographic transcription and may, therefore, differ from intelligibility calculated using other methods such as VAS, DME, or machine-based transcription. We anticipate that our findings would be similar for VAS ratings of intelligibility, as previous work has found transcription intelligibility and scaled ratings of intelligibility to be strongly related (Abur et al., 2019; Stipancic et al., 2016).

Speaking Rate

Speaking rate has been widely used to track bulbar disease progression in ALS (Green et al., 2013; Yorkston et al., 1993). Previous work has demonstrated the linear decline of speaking rate with symptom duration in ALS (Ball et al., 2002; Yorkston et al., 1993) and its association with other aspects of bulbar function, such as voice quality, velopharyngeal functioning, and tongue movement (Ball et al., 2001). Despite these characteristics and its use as a primary component of bulbar assessment in ALS, our study showed that speaking rate only statistically differentiated between the mild and moderate severity groups and did not yield statistical differences between groups at the more mild end of the continuum (i.e., normal and mild) and at the most severe end of the continuum (i.e., moderate, severe, and profound). In addition, the accuracy values from the ROC analyses were generally the smallest for speaking rate as compared to the other three clinical measures. The lack of statistical differences between groups is consistent with recent work illustrating plateaus in speaking rate as speech motor performance deteriorates (Barnett et al., 2019; Rong, Yunusova, & Green, 2015). This plateau in speaking rate suggests that, over time, speaking rate becomes less responsive to change while other features (e.g., intelligibility) continue to decline. To explain these findings, researchers have speculated that the use of a single measure, such as speaking rate, in isolation, may preclude insight into other factors that impact perceived severity (Stipancic et al., 2021). For example, Stipancic et al. (2018) found that neither intelligibility nor speaking rate was able to adequately distinguish between patient groups whose self-report of speech change on the ALS Functional Rating Scale–Revised (Cedarbaum et al., 1999) would constitute a meaningful change. The authors concluded that these measures may not be capturing a meaningful element of communication, which likely contributes to self-judgments of severity, such as the amount of effort patients must exert to produce intelligible speech (Stipancic et al., 2018). Additionally, speaking rate was the only severity-surrogate measure that was completely free of listener judgment, which is likely part of the reason it showed a limited association with clinician severity ratings. If speaking rate is the only option for severity stratification, we recommend that researchers and clinicians exercise caution when stratifying groups on either end of the severity continuum. Similar to the intelligibility findings, speaking rate may not perform well with five-group stratification. Therefore, researchers and clinicians may need to consider collapsing across groups, particularly those who exhibit severe speech impairments. Speaking rate may, instead, be more useful for coarse-grained or binary (i.e., unimpaired vs. impaired) stratifications of severity, consistent with methods used in prior work (Shellikeri et al., 2016; Stipancic et al., 2018).

Intelligible Speaking Rate

Intelligible speaking rate provided statistical differences between normal and mild speakers and between mild and moderate speakers, with trends toward statistical differences of the other adjacent contrasts. Intelligible speaking rate, as measured in IWPM, has traditionally been

considered an indicator of communication efficiency (Kent et al., 1989; Yorkston & Beukelman, 1981), which refers to the ability to quickly transmit a clear message. This measure, although attractive for combining two readily used measures of speech impairment (i.e., intelligibility and speaking rate), has been vastly understudied. In Hustad et al.'s (2019) work, the authors highlighted the utility of intelligible speaking rate as a measure that "simultaneously provides information about two areas of deficits that are common in dysarthria and interact with one another" (p. 811). Intelligible speaking rate performed similarly to intelligibility in that it was statistically different between two of the adjacent severity categories and had comparable effect sizes and diagnostic accuracy values between groups. We acknowledge that intelligible speaking rate may be an onerous metric to acquire given its reliance on calculating two other measures. However, for researchers who have the ability to obtain both intelligibility and speaking rate, the combination of these measures appears to outperform speaking rate in stratifying between severity categories. In light of these findings, we advocate for the continued exploration and use of this measure in research and clinical settings.

Listener Effort

Based on our analyses, listener effort yielded the strongest validity for severity stratification of the four severity-surrogate measures examined in this study. In contrast to the other three severity surrogates, listener effort provided clear distinctions between all five severity categories. Moreover, listener effort had surprisingly excellent reliability, particularly for a subjective measure. Aside from its strong, linear relationship with adjectival severity ratings, listener effort performed the best with distinguishing between the severity groups at the mild end of the continuum (i.e., normal–mild; see Table 6) and performed second best, after intelligibility, with distinguishing the severity groups at the more severe end of the continuum (i.e., moderate–severe–profound; see Table 6), indicating its sensitivity to smaller changes in severity that were not detected by speaking rate or intelligible speaking rate. Listener effort also yielded the largest specificity, sensitivity, and accuracy values from the ROC analyses (except for specificity for the moderate group being higher for intelligible speaking rate; see Table 7), indicating superior distinguishability between clinician-rated severity groups. The strong performance of listener effort in differentiating groups demonstrates the need to consider joint contribution between a speaker and a listener when measuring severity (Olmstead et al., 2020). Using listener effort as an outcome measure thereby begins to address the level of participation in the International Classification of Functioning, Disability and Health (World Health Organization, 2002), which is a critically important domain to consider when stratifying severity groups in treatment efficacy studies, for example. In addition, the outstanding performance of listener effort in its association with clinician severity ratings may indicate that listener effort significantly influences clinicians' ratings of speech severity.

Although listener effort has the disadvantage of being subjective and, therefore, highly reliant on listener characteristics, our findings demonstrate its utility as a valid, reliable, and sensitive severity stratification measure, at least for ratings as completed by SLPs who have experience in serving patients with dysarthria. In contrast to intelligibility, listener effort was sensitive to early-stage severity differences (i.e., normal vs. mild). In contrast to speaking rate and intelligible speaking rate, listener effort was also sensitive to late-stage severity differences (i.e., moderate vs. severe vs. profound) likely because patients in the later stages of disease progression may not adjust their rate as dramatically if it no longer enhances clarity. However, as speech function inevitably deteriorates, listeners exert more effort to understand the speaker's message. Overall, our findings provide preliminary support for ratings of listener effort as a useful severity stratification tool. However, despite the relative ease of acquiring ratings of listener effort (i.e., effort ratings are less time and labor consuming than orthographic transcription), this measure has not yet been investigated as extensively as intelligibility or speaking rate.

Limitations and Future Directions

Findings from this work may not generalize to populations with speech impairment other than ALS or to assessors other than experienced SLPs. However, similar methods could be used for future work in different patient populations, dysarthria subtypes, and etiologies beyond ALS to develop a severity tool that can be applied across patients and diagnoses. Similarly, this work may not generalize to listeners outside of the experienced clinicians used in this study. The cutoff values calculated in this study should be applied with caution to other listener groups, such as naive listeners, as ratings of speech severity, intelligibility, and listener effort may be judged differently by SLPs and other types of listeners (i.e., Dagenais et al., 1999). Examining the assignment of adjectival severity labels by naive listeners may be an interesting area of future research. In the current study, transcription from a single research assistant was used solely for ensuring our cohort included a diverse range of severities. The wide range of severity ratings scored by the SLPs (as reported in the results) suggested that this approach was effective. The small number of SLP raters for each individual speech sample (i.e., $n = 2$) may also be strengthened in the future by recruiting more listeners. However, the reliability statistics indicated excellent reliability both within and

between listeners and do not support the need for a larger number of listeners (at least for this group of experienced SLPs). The four severity-surrogate measures examined were chosen based on previous work demonstrating evidence for their utility in patients with ALS; however, theoretically, a number of other measures could also serve to stratify severity groups, such as comprehensibility (Hustad, 2008; Yorkston et al., 1996), patient-reported outcomes (Baylor et al., 2009), or communication staging models (Allison et al., 2019; Yorkston & Beukelman, 1999). Future work could employ similar procedures to the ones used in the current study to explore the validity and reliability of other metrics to stratify speakers into speech severity categories. Lastly, because our study specifically asked clinicians to classify the speakers into one of five severity groups, the analyses performed in this study cannot directly address the problem of how many categories is adequate or optimal for defining dysarthria severity. Thus, while our findings provide evidence for the validity and reliability of using five categories, further research should investigate the classification performance of adjectival ratings and other severity-surrogate metrics while allowing listeners to group speakers into any number of severity groups they see fit (see Lansford et al., 2014). Furthermore, although the severity-surrogate measures individually demonstrated high sensitivity, specificity, and accuracy for distinguishing between the severity groups, future work could examine multivariate combinations of the variables examined in the current study (e.g., intelligible speaking rate *and* listener effort) given the advantages and disadvantages of each (see, e.g., J. Lee et al., 2018, who used both speech intelligibility and speaking rate as proxies for severity). We did not conduct multivariate analyses in the current study given the high sensitivity, specificity, and accuracy of the current cutoff points, which suggested that adding multiple variables was unlikely to significantly increase diagnostic power.

Conclusions

This study is the first, to our knowledge, to (a) validate clinician-based adjectival ratings as a severity stratification measure and (b) provide a systematic and empirical classification scheme for dysarthria severity by examining how severity surrogates align with clinician ratings of severity in a well-curated data set of speakers with ALS and neurologically healthy controls. In addition to demonstrating strong support for adjectival ratings as a valid and reliable severity stratification method, we provided cutoff points for the four severity-surrogate measures (i.e., speech intelligibility, speaking rate, intelligible speaking rate, and listener effort) used to test the validity of adjectival labels. Ultimately, we found that listener effort performed the best for stratifying the five speech severity groups and that intelligibility and intelligible speaking rate may also be useful for stratifying between some severity groups. However, for researchers or clinicians who have an existing data set and access to only a subset of the measures analyzed in this article, we provided recommendations to guide decision making when selecting a measure for severity stratification. The findings from the current work fulfill a critical need for a scientifically validated scale of speech severity in dysarthria for both clinical and research purposes.

Acknowledgments

This study was supported by National Institute on Deafness and Other Communication Disorders Grants R01DC009890 (PIs: Jordan R. Green and Yana Yunusova), R01DC013547 (PI: Jordan R. Green), R01DC017291 (PIs: Yana Yunusova and Jordan R. Green), and K24DC016312 (PI: Jordan R. Green). This work was conducted with support from Harvard Catalyst, The Harvard Clinical and Translational Science Center (National Center for Advancing Translational Sciences, National Institutes of Health Award UL 1TR002541) and financial contributions from Harvard University and its affiliated academic health care centers. This work is based on the master's thesis of the second author. The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, or the National Institutes of Health. The authors would like to thank all the participants (speakers and speech-language pathologists) for their participation in this study and Brian Richburg (Speech and Feeding Disorders Lab, Boston, MA) for his assistance with data collection.

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

Allison, K. M., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., & Green, J. R. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 18*(5–6), 358–366. https://doi.org/10.1080/21678421.2017.1303515

Allison, K. M., Yunusova, Y., & Green, J. R. (2019). Shorter sentence length maximizes intelligibility and speech motor performance in persons with dysarthria due to amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology, 28*(1), 96–107. https://doi.org/10.1044/2018_AJSLP-18-0049

Ball, L. J., Beukelman, D. R., & Pattee, G. L. (2002). Timing of speech deterioration in people with amyotrophic lateral

sclerosis. *Journal of Medical Speech-Language Pathology, 10*(4), 231–236.

Ball, L. J., Willis, A., Beukelman, D. R., & Pattee, G. L. (2001). A protocol for identification of early bulbar signs in amyotrophic lateral sclerosis. *Journal of the Neurological Sciences, 191*(1–2), 43–53. https://doi.org/10.1016/S0022-510X(01)00623-2

Barnett, C., Green, J. R., Marzouqah, R., Stipancic, K. L., Berry, J. D., Korngut, L., Genge, A., Shoesmith, C., Briemberg, H., Abrahao, A., Kalra, S., Zinman, L., & Yunusova, Y. (2019). Reliability and validity of speech & pause measures during passage reading in ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 21*(1–2), 42–50. https://doi.org/10.1080/21678421.2019.1697888

Baylor, C. R., Yorkston, K. M., Eadie, T. L., Miller, R. M., & Amtmann, D. (2009). Developing the Communicative Participation Item Bank: Rasch analysis results from a spasmodic dysphonia sample. *Journal of Speech, Language, and Hearing Research, 52*(5), 1302–1320. https://doi.org/10.1044/1092-4388(2009/07-0275)

Blaney, B., & Hewlett, N. (2007). Dysarthria and Friedreich's ataxia: What can intelligibility assessment tell us? *International Journal of Language & Communication Disorders, 42*(1), 19–37. https://doi.org/10.1080/13682820600690993

Brooks, B. R., Miller, R. G., Swash, M., & Munsat, T. L. (2000). El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders, 1*(5), 293–299. https://doi.org/10.1080/146608200300079536

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics, 15*(3), 181–193. https://doi.org/10.1080/02699200010003378

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., & Nakanishi, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences, 169*(1–2), 13–21. https://doi.org/10.1016/S0022-510X(99)00210-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Connaghan, K. P., & Patel, R. (2017). The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research, 60*(1), 38–50. https://doi.org/10.1044/2016_JSLHR-S-15-0291

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America, 119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Cote-Reschny, K. J., & Hodge, M. M. (2010). Listener effort and response time when transcribing words spoken by children with dysarthria. *Journal of Medical Speech-Language Pathology, 18*(4), 24–35.

Dagenais, P. A., Watts, C. R., Turnage, L. M., & Kennedy, S. (1999). Intelligibility and acceptability of moderately dysarthric speech by three types of listeners. *Journal of Medical Speech-Language Pathology, 7*(2), 91–95.

dos Santos Barreto, S., & Ortiz, K. Z. (2015). Protocol for the evaluation of speech intelligibility in dysarthrias: Evidence of reliability and validity. *Folia Phoniatrica et Logopaedica, 67*(4), 212–218. https://doi.org/10.1159/000441929

Doyle, P. C., Leeper, H. A., Kotler, A.-L., Thomas-Stonell, N., O'Neill, C., Dylke, M.-C., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development, 34*(3), 309–316.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management* (3rd ed.). Elsevier Health Sciences.

Gordon-Brannan, M., & Hodson, B. W. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology, 9*(2), 141–150. https://doi.org/10.1044/1058-0360.0902.141

Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., Zinman, L., & Berry, J. D. (2013). Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 14*(7–8), 494–500. https://doi.org/10.3109/21678421.2013.817585

Gurevich, N., & Scamihorn, S. L. (2017). Speech-language pathologists' use of intelligibility measures in adults with dysarthria. *American Journal of Speech-Language Pathology, 26*(3), 873–892. https://doi.org/10.1044/2017_AJSLP-16-0112

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research Electronic Data Capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hustad, K. C. (2006a). A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology, 15*(3), 268–277. https://doi.org/10.1044/1058-0360(2006/025)

Hustad, K. C. (2006b). Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica, 58*(3), 217–228. https://doi.org/10.1159/000091735

Hustad, K. C. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatrica et Logopaedica, 59*(6), 306–317. https://doi.org/10.1159/000108337

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51*(3), 562–573. https://doi.org/10.1044/1092-4388(2008/040)

Hustad, K. C., & Gearhart, K. J. (2004). Listener attitudes toward individuals with cerebral palsy who use speech supplementation strategies. *American Journal of Speech-Language Pathology, 13*(2), 168–181. https://doi.org/10.1044/1058-0360(2004/017)

Hustad, K. C., Sakash, A., Broman, A. T., & Rathouz, P. J. (2019). Differentiating typical from atypical speech production in 5-year-old children with cerebral palsy: A comparative analysis. *American Journal of Speech-Language Pathology, 28*(2S), 807–817. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0108

Kalra, S., Khan, M., Barlow, L., Beaulieu, C., Benatar, M., Briemberg, H., Chenji, S., Clua, M. G., Das, S., Dionne, A., Dupré, N., Emery, D., Eurich, D., Frayne, R., Genge, A., Gibson, S., Graham, S., Hanstock, C., Ishaque, A., . . . Zinman, L. (2020). The Canadian ALS Neuroimaging Consortium (CALSNIC)—A multicentre platform for standardized imaging and clinical studies in ALS. *Amyotrophic Lateral Sclerosis, 14,* 11–23. https://doi.org/10.1101/2020.07.10.20142679

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 16*(3), 222–234. https://doi.org/10.1044/1058-0360(2007/027)

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

King, J. M., Watson, M., & Lof, G. L. (2012). Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. *Journal of Medical Speech-Language Pathology, 20*(1), 1–17.

Koch Fager, S., & Burnfield, J. M. (2015). Speech recognition for environmental control: Effect of microphone type, dysarthria, and severity on recognition results. *Assistive Technology, 27*(4), 199–207. https://doi.org/10.1080/10400435.2015.1024349

Kuruvilla, M. S., Green, J. R., Yunusova, Y., & Hanford, K. (2012). Spatiotemporal coupling of the tongue in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 55*(6), 1897–1909. https://doi.org/10.1044/1092-4388(2012/11-0259)

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., & Steen, N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology, 16*(4), 408–416. https://doi.org/10.3109/17549507.2014.927922

Lansford, K. L., Liss, J. M., & Norton, R. E. (2014). Free-classification of perceptually similar speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 57*(6), 2051–2064. https://doi.org/10.1044/2014_JSLHR-S-13-0177

Lee, J., Bell, M., & Simmons, Z. (2018). Articulatory kinematic characteristics across the dysarthria severity spectrum in individuals with amyotrophic lateral sclerosis. *American Journal of Speech-Language Pathology, 27*(1), 258–269. https://doi.org/10.1044/2017_AJSLP-16-0230

Lee, Y. W., Kim, G. H., & Kwon, S. B. (2020). The usefulness of auditory perceptual assessment and acoustic analysis for classifying the voice severity. *Journal of Voice, 34*(6), 884–893. https://doi.org/10.1016/j.jvoice.2019.04.013

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders, 48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Miracle Ear. (n.d.). *Hearing test*. https://www.miracle-ear.com/hearing-test

Nagle, K. (2015). Effect of intelligibility and speech rate on perceived listener effort. *The Journal of the Acoustical Society of America, 137*(4), 2433. https://doi.org/10.1121/1.4920878

Nagle, K. F., & Eadie, T. L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders, 45*(3), 235–245. https://doi.org/10.1016/j.jcomdis.2012.01.001

Nagle, K. F., & Eadie, T. L. (2018). Perceived listener effort as an outcome measure for disordered speech. *Journal of Communication Disorders, 73,* 34–49. https://doi.org/10.1016/j.jcomdis.2018.03.003

Olmstead, A. J., Lee, J., & Viswanathan, N. (2020). The role of the speaker, the listener, and their joint contributions during communicative interactions: A tripartite view of intelligibility in individuals with dysarthria. *Journal of Speech, Language, and Hearing Research, 63*(4), 1106–1114. https://doi.org/10.1044/2020_JSLHR-19-00233

Paja, M. S., & Falk, T. H. (2012). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association* (pp. 3–6).

Picou, E. M., Moore, T. M., & Ricketts, T. A. (2017). The effects of directional processing on objective and subjective listening effort. *Journal of Speech, Language, and Hearing Research, 60*(1), 199–211. https://doi.org/10.1044/2016_JSLHR-H-15-0416

R Development Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rong, P. (2019). The effect of tongue–jaw coupling on phonetic distinctiveness of vowels in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 62*(9), 3248–3264. https://doi.org/10.1044/2019_JSLHR-S-19-0058

Rong, P., Yunusova, Y., & Green, J. R. (2015). Speech intelligibility decline in individuals with fast and slow rates of ALS progression. *Interspeech,* 2967–2971.

Rong, P., Yunusova, Y., Wang, J., & Green, J. R. (2015). Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach. *Behavioural Neurology, 2015,* 1–11. https://doi.org/10.1155/2015/183027

Rong, P., Yunusova, Y., Wang, J., Zinman, L., Pattee, G. L., Berry, J. D., Perry, B., & Green, J. R. (2016). Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems. *PLOS ONE, 11*(5), Article e0154971. https://doi.org/10.1371/journal.pone.0154971

Schiavetti, N., Sacco, P. R., Metz, D. E., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering severity. *Journal of Speech and Hearing Research, 26*(4), 568–573. https://doi.org/10.1044/jshr.2604.568

Shellikeri, S., Green, J. R., Kulkarni, M., Rong, P., Martino, R., Zinman, L., & Yunusova, Y. (2016). Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 59*(5), 887–899. https://doi.org/10.1044/2016_JSLHR-S-15-0238

Shriberg, L. D., & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders, 47*(3), 256–270. https://doi.org/10.1044/jshd.4703.256

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics, 21*(20), 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

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

Stipancic, K. L., Yunusova, Y., Berry, J. D., & Green, J. R. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 61*(11), 2757–2771. https://doi.org/10.1044/2018_JSLHR-S-17-0366

Stipancic, K. L., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., & Green, J. R. (2021). Two distinct clinical phenotypes of bulbar motor impairment in amyotrophic lateral sclerosis. *Frontiers in Neurology, 12,* 664713. https://doi.org/10.3389/fneur.2021.664713

Strand, E. A., & Yorkston, K. M. (1994). Description and classification of individuals with dysarthria: A 10-year review. In J. A. Till, K. M. Yorkston, & D. R. Beukelman (Eds.), *Motor speech disorders: Advances in assessment and treatment* (pp. 37–56). Brookes.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *Journal of Speech, Language, and Hearing Research, 57*(3), 779–792. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research, 47*(4), 766–783. https://doi.org/10.1044/1092-4388(2004/058)

Torchiano, M. (2020). *effsize—A package for efficient effect size computation*. Zenodo. https://doi.org/10.5281/zenodo.1480624

Turner, M. R., Bowser, R., Bruijn, L., Dupuis, L., Ludolph, A., McGrath, M., Manfredi, G., Maragakis, N., Miller, R. G., Pullman, S. L., Rutkove, S. B., Shaw, P. J., Shefner, J., & Fischbeck, K. H. (2013). Mechanisms, models and biomarkers in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 14*(Suppl. 1), 19–32. https://doi.org/10.3109/21678421.2013.778554

Wang, J., Kothalkar, P. V., Kim, M., Bandini, A., Cao, B., Yunusova, Y., Campbell, T. F., Heitzman, D., & Green, J. R. (2018). Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples. *International Journal of Speech-Language Pathology, 20*(6), 669–679. https://doi.org/10.1080/17549507.2018.1508499

Weismer, G., Laures, J. S., Jeng, J. Y., Kent, R. D., & Kent, J. F. (2000). Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica, 52*(5), 201–219. https://doi.org/10.1159/000021536

Whitehill, T. L., & Wong, C. C.-Y. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14*(4), 335–341.

Wilson, C., Page, A. D., & Adams, S. G. (2020). Listener ratings of effort, speech intelligibility, and loudness of individuals with Parkinson's disease and hypophonia. *Canadian Journal of Speech-Language Pathology & Audiology, 44*(2), 33–49.

World Health Organization. (2002). *Towards a common language for functioning, disability and health: ICF—The International Classification of Functioning, Disability and Health*.

Xue, W., Ramos, V. M., Harmsen, W., Cucchiarini, C., Van Hout, R. W. N. M., & Strik, H. (2020). Towards a comprehensive assessment of speech intelligibility for pathological speech. *INTERSPEECH,* 3146–3150. https://doi.org/10.21437/Interspeech.2020-2693

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders, 11*(6), 499–512. https://doi.org/10.1016/0021-9924(78)90024-2

Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders, 46*(3), 296–301. https://doi.org/10.1044/jshd.4603.296

Yorkston, K. M., & Beukelman, D. R. (1999). Staging interventions in progressive dysarthria. *SIG 2 Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders, 9*(4), 7–12. https://doi.org/10.1044/nnsld9.4.7

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

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech. *American Journal of Speech-Language Pathology, 5*(1), 55–66. https://doi.org/10.1044/1058-0360.0501.55

Yorkston, K. M., Strand, E. A., Miller, R., Hillel, A., & Smith, K. (1993). Speech deterioration in amyotrophic lateral sclerosis: Implications for the timing of intervention. *Journal of Medical Speech-Language Pathology, 1,* 35–46.

Yunusova, Y., Graham, N. L., Shellikeri, S., Phuong, K., Kulkarni, M., Rochon, E., Tang-Wai, D. F., Chow, T. W., Black, S. E., Zinman, L. H., & Green, J. R. (2016). Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). *PLOS ONE, 11*(1), Article e0147573. https://doi.org/10.1371/journal.pone.0147573

Yunusova, Y., Green, J. R., Greenwood, L., Wang, J., Pattee, G. L., & Zinman, L. (2012). Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica, 64*(2), 94–102. https://doi.org/10.1159/000336890

Yunusova, Y., Green, J. R., Lindstrom, M. J., Ball, L. J., Pattee, G. L., & Zinman, L. (2010). Kinematics of disease progression in bulbar ALS. *Journal of Communication Disorders, 43*(1), 6–20. https://doi.org/10.1016/j.jcomdis.2009.07.003

Yunusova, Y., Green, J. R., Wang, J., Pattee, G., & Zinman, L. (2011). A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (ALS). *Journal of Visualized Experiments, 48,* e2422. https://doi.org/10.3791/2422