Validity and Reliability of Auditory-Perceptual Scales for the Assessment of Stuttering Severity

Sima Sokolov

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Medical Speech-Language Pathology

University of Washington

2023

Committee:

Gabriel Cler

Tanya Eadie

Melissa A Kokaly

Program Authorized to Offer Degree:

Department of Speech & Hearing Sciences

University of Washington

Abstract

Validity and Reliability of auditory-perceptual Scales for the Assessment of Stuttering Severity

Sima Sokolov

Chair of the Supervisory Committee:

Gabriel Cler

Department of Speech & Hearing Sciences

The current study investigated the appropriateness of equal appearing interval scaling, direct magnitude estimation scaling, and visual analog scaling for assessing stuttering severity by determining whether the continuum of stuttering severity, as defined in this study, was prothetic or metathetic. A secondary purpose was to determine interrater reliability of all three scaling methods. The stuttering severity of 20 reading samples was each judged by three groups of 15 listeners who used the three scaling techniques. The results indicated that the sets of values were related to each other in a linear fashion, indicating that stuttering severity is a metathetic continuum, which is inconsistent with previous studies. These findings suggest that equal appearing interval scaling may be as appropriate as direct magnitude estimation or visual analog scaling for measuring stuttering severity. Future directions should include investigation of the

effect of individual training on intrarater reliability and use of expert raters on construct validity and reliability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor and mentor, Dr. Gabriel Cler, for his support through this thesis project. I am so grateful to have had this experience under his mentorship. I would like to thank the members of my committee, Dr. Tanya Eadie and Professor Melissa Kokaly, for their contributions, insight, and guidance. I would also like to thank the members of the Quantitative Imaging for Learning, Language, and Speech Lab for recruiting and running participants for this research project. Lastly, I would like to thank the participants for their time, effort, and contributions to this study.

1. INTRODUCTION

Stuttering is a motor speech disorder characterized by a disruption in the fluency, timing, and rhythm of speech. Stuttering presents as repetitions of sounds, syllables, or words, prolongations of sounds, or prolonged pauses between sounds and words often accompanied by tension (which are referred to as "blocks" by clinicians and people who stutter; Wingate, 1964). Learned physical concomitants such as eye blinking, jaw jerking, and head movements can occur as an attempt to break out of or move through moments of stuttering (Prasse et al., 2008). Developmental stuttering is the most common form of stuttering, with a gradual onset between the ages of 3 and 8 years, during a period of extensive speech and language development (Ashurst & Wasson, 2011; Yairi & Ambrose, 2013). Approximately 5% of all children experience developmental stuttering, but 75% of them undergo spontaneous remission within 4 years (Yairi & Ambrose, 2013).

Persistent developmental stuttering (PDS) is a form of developmental stuttering that has not resolved, either spontaneously or from speech therapy (Costa & Kroll, 2000). PDS impacts around 1% of the general adult population (Craig et al., 2002; Yairi & Ambrose, 2013). PDS can profoundly impact an individual’s employment, job satisfaction, personal and romantic relationships, participation in daily communication situations, and overall quality of life (Blood & Blood, 2016; Carter et al., 2017; Croft & Byrd, 2020).

Despite the prevalence and long history of study of PDS and its profound impact on quality of life, evidence-based assessment and effective treatment in adults remain elusive

(Bothe et al., 2006; Connery et al., 2021). A significant barrier in stuttering research is the lack of reliable and fast outcome measures that translate to clinical settings.

Auditory-perceptual judgments of speech allow for classification and description of a variety of communication disorders (Eadie & Doyle, 2002a). However, inappropriate use of a particular perceptual rating can lead to misclassification of persons for research and clinical purposes (Zraick & Liss, 2000). A scaling method for a particular speech parameter must be both valid and reliable in order to accurately measure it. The purpose of this study is to compare the validity and reliability of equal-appearing interval (EAI) scaling, which is the most used rating method for assessment of stuttering severity, to those of other scale ratings that are less commonly used. The findings of this study will have implications for methodologies employed for rating stuttering severity in the speech of adults who stutter in both research and clinical contexts.

1.1 Stuttering Severity

Stuttering measurement is a vital component of clinical practice and research. It allows researchers and clinicians to effectively identify and describe the nature, severity, and impact of the stuttering on a speaker's communication and overall quality of life. Additionally, it allows speech-language pathologists (SLPs) to reliably measure stuttering and efficiently track clients' progress and outcomes to support the effectiveness of treatment progress and outcome. Stuttering severity measures include the frequency, type, duration, and severity of stuttering, as well as speech naturalness, speech rate, and concomitant or associated behaviors. Currently there is no measure to capture the variability of stuttering, which can be a challenging aspect of the speaker's experience of stuttering, and a notable confounding factor in measuring severity in both

research and clinical contexts. Given that a common goal of adult behavioral stuttering treatments is reduction of stuttering during everyday speaking situations, one indicator of treatment success is reduction in the level of stuttering severity. Having a fast and reliable measurement scale clinicians and clients could use to measure the degree of change in this physical target could be helpful. Additionally, a faster measurement tool could allow for repeated measurements to capture aspects of variability if needed. However, there has been considerable debate among researchers and clinicians regarding appropriate techniques for measuring stuttering severity (Yaruss, 1997). Yaruss (1997) found that SLPs can be less comfortable diagnosing stuttering in comparison to other communication disorders due to lack of agreement on how stuttering severity should be measured.

1.2 Scales for Measuring Stuttering Severity

When measuring severity of stuttering, SLPs base their judgement on their overall impression of the speech sample, while considering both stuttering typology and frequency (O'Brian et al., 2004). Perhaps the best–known assessment of stuttering severity is the Stuttering Severity Instrument – Fourth Edition (SSI-4). It is a diagnostic assessment tool used for clinical and research purposes. It examines stuttering severity across three parameters: (1) frequency, measured by percent syllables stuttered, (2) duration, measured by the average of the three longest stuttering events, and (3) physical concomitants, measured by clinical judgments of physical and audible signs of struggle during speech (Riley, 2009). The three subtest scores provide a final severity rating of 'very mild,' 'mild,' 'moderate,' 'severe,' or 'very severe'.

However, in comparison to other severity rating measures (e.g., EAI), the SSI-4 is considered to be notoriously time-consuming and unreliable, as are other standardized

assessments that also require disfluency counts (Cordes & Ingham, 1994; Davidow, 2021; Davidow & Scott, 2017). In a recent study, Davidow and Scott (2017) examined the reliability of the SSI-4. Results of the study highlighted limitations in the use SSI-4, and the authors did not recommend it for the use of measuring stuttering severity. Furthermore, Lewis (1995) reported that the SSI-3 (a previous edition of the SSI-4) provided no additional information beyond that which could be attained using EAI.

Yet, severity measures may be beneficial to clinical decision-making for some clinicians and are vital for research in the area of stuttering (Yaruss, 1998). Given the limitations of the SSI-4, Davidow and Scott (2017) recommended the use of EAI scaling as an alternative measure for stuttering severity in clinical settings.

1.3 The Stuttering Severity Continuum

Given the need in research and in the clinic to measure stuttering, it is vital that any such measurements are valid and reliable. An issue of importance in any rendering any auditory-perceptual judgment is determining what type of rating scale may be appropriate for measuring the percept under investigation. In this case, we are interested in the continuum of stuttering severity. Stevens (1975) investigated continua that can be scaled and concluded that there are two kinds of continua: metathetic, which is substitutive and differs in terms of quality, and prothetic, which is additive and differs in degree or quantity (Stevens, 1975). A metathetic continuum can be divided into equal intervals and those interval judgments would be linearly related to direct magnitude estimates of the same stimuli, whereas a prothetic continuum cannot be divided into equal intervals. Stevens (1975) proposed that different scales should be utilized for rating prothetic versus metathetic continua: ratio scales for prothetic continua and categorical scales for metathetic ones. To distinguish prothetic and metathetic continua, Stevens (1975)

suggested regressing categorical ratings, such equal appearing interval scaling (EAI) onto ratio scale ratings, such as direct magnitude estimation (DME) and visual analog scales (VAS) and determining the line of best fit. If the fit is linear, then it is a metathetic continuum, and if the fit is curvilinear, then the continuum is prothetic since the intervals between the scale points are not being perceived as equal (Stevens, 1975).

To date, one study has investigated whether the stuttering severity continuum is prothetic or metathetic. Schiavetti et al. (1983) compared the appropriateness of direct magnitude estimation and interval scaling for assessing stuttering severity. In the study, the stuttering severity of 20 speakers who stutter was scaled by three groups of 15 listeners using interval scaling, direct magnitude estimation with standard/modulus, and direct magnitude estimation without standard/modulus. Results of the study indicated that the continuum of stuttering severity is a prothetic continuum and that interval scaling, such as EAI scales would be inappropriate for measuring stuttering severity. Ratio scale ratings, such as DME, which require listeners to assign numbers to stimuli that are proportional to the ratios of stimulus magnitudes along the continuum, would therefore have better construct validity (Schiavetti et al., 1983). Further evidence that stuttering severity is a prothetic continuum would strengthen the findings by Schiavetti et al. (1983) and reinforce the need for studies that show whether ratio scale ratings, such as DME and VAS, are as valid and reliable as interval scaling, such as EAI, to rate stuttering severity in clinical and research settings.

When comparing scales, construct validity and content validity can be examined for different purposes. Construct validity ensures that the scale measures the concept that it is intended to measure, while content validity ensures that the scale is representative of what it aims

to measure. This study will focus on measuring construct validity by determining whether the continuum of stuttering severity is prothetic or metathetic.

1.3.1 *Equal Appearing Interval Scaling*

Equal appearing interval scaling (EAI), also known as severity rating (SR) scales or global severity rating (GSR) scales, are used in research and clinical practice and involve partition scaling in which listeners assign a number to a stimulus. EAI has fixed endpoints and uses whole numbers (e.g., between 1 and n). Odd and even-numbered interval scales, such as 5-point, 7- point, 8-point, 9-point, 10-point, and 15-point, are frequently used in clinical practice and research for assessment of stuttering severity, with one end of the scale indicating the most severe stuttering perception, and the other indicating the least severe stuttering perception. The scale requires the listener to listen to video or audio-only speech samples and assign a numerical value that represents their perceived overall stuttering severity (O'Brian et al., 2004).

Some researchers have noted that EAI ratings correlate to disfluency counts (e.g., percentage of syllables stuttered) and are quick and easy to perform (O'Brian et al., 2020; Onslow et al., 2018). Advantages of EAI scales are that they are simple to use, require no equipment, can be used in isolation, and appear to need little or no training (O'Brian et al., 2004). Additionally, clients can use them for self-report of stuttering severity outside the clinic, which recognizes the situational and temporal variability that can be very common with stuttering severity measurement. This can allow clients and SLPs to communicate more easily and effectively about stuttering severity (O'Brian et al., 2004). Karimi et al. (2014) investigated whether clinician EAIs and speaker EAIs can be used interchangeably and concluded that they cannot be used interchangeably to measure temporal stuttering severity changes for an individual

client. They also found that EAI scales rated by clinicians and speakers cannot be used interchangeably to assess absolute differences within a trial. Thus, it appears that ratings using EAI scales differ depending on who makes the judgment (clinician or person who stutters) and are not reliable for measuring individual changes rendering them less than ideal for clinical or research purposes (Karimi et al., 2014).

Davidow (2021) compared the inter and intra-reliability of EAI scales when rating reading versus spontaneous speech samples. In this study, 24 speech-language pathology graduate students had taken fluency disorders coursework where they were trained in utilizing SSI-4 and EAI scales. After a training where they watched 55 speaking samples and performed the SSI-4 counts until the scores fell within a target range predetermined by experts, the participants rated four videos of PWS using SSI-4 or a 5-point EAI scale. The speakers were 21-60 years old, though no information regarding stuttering severity was provided in the methods section of the study, other than the speakers being diagnosed with developmental stuttering. The videos included both a monologue and a reading passage. The EAI scale was a 5-point, where a rating of 5 indicated very severe, 4 indicated severe, 3 indicated moderate, 2 indicated mild, and 1 indicated very mild. Davidow (2021) found that neither measure (reading or conversation) produced an acceptable level of inter-rater agreement of 80% and rater agreement for both types of speech was questionable (Davidow, 2021). The authors concluded that EAI scaling is not sufficient to replace the SSI-4 as a measure of stuttering severity. To produce an acceptable level of inter-rater agreement, additional training and the use of exemplars of standardized levels of stuttering severity, such as a modulus, are necessary (Davidow, 2021).

Moreover, there is evidence to suggest that EAI scales may be inappropriate for measuring the percept of stuttering severity because it is undecided whether individual scale

items can be considered interval-level data or ordinal data (Berry & Silverman, 1972; Schiavetti et al., 1983; Teghtsoonian et al., 1975). That is, an important consideration regarding the reliability and validity of using EAI to rate stuttering severity is whether the distance between each point on the scale is equivalent and not just “equal appearing”.

An issue of importance in determining the validity of EAI scaling is what type of percept is being rated and how that influences scale selection. Stevens (1975) proposed that different scales should be utilized for rating prothetic versus metathetic continua: ratio scales for prothetic continua and categorical scales for metathetic ones. To distinguish prothetic and metathetic continua, Stevens (1975) suggested regressing categorical (e.g., EAI) ratings onto ratio (e.g., direct magnitude estimation) scale ratings and determining the line of best fit. If the fit is linear, then it is a metathetic continuum, and if the fit is curvilinear, then the continuum is prothetic since the intervals between the scale points are not being perceived as equal (Stevens, 1975). The Schiavetti (1983) study provides evidence for stuttering severity being a prothetic continuum as discussed above, and according to Stevens (1975), ratio scales should be used for prothetic continua. Given that EAI scaling is a categorical scale, and not ratio-based scaling, it may be an inappropriate scale for the stuttering severity continuum, which is prothetic.

Additionally, EAI scaling is inappropriate for measuring a prothetic continuum because when raters try to divide the continuum into equal intervals, they subdivide the lower end of the into smaller intervals than those at the upper end (Stevens, 1975). In other words, the distances between points on the scale are not, in fact, equal. Berry and Silverman (1972) have also demonstrated that interval widths presented in EAI scales used to measure severity of stuttering are not equal because the interval points at the lower end are perceived at about half the width of the interval points at the upper end.

Given the explanations by Berry and Silverman (1972) and Stevens (1975) and the question of how appropriate and valid different scaling procedures are for different continua, such as stuttering severity, Schiavetti et al. (1983) investigated whether the continuum of stuttering severity is prothetic or metathetic to inform their recommendations about whether direct magnitude estimation (DME) and/or EAI scaling was appropriate for assessing stuttering severity. They found that stuttering severity was a prothetic continuum, and thus interval scaling, such as EAI scales would be inappropriate. However, DME, which requires listeners to assign numbers to stimuli that are proportional to the ratios of stimulus magnitudes along the continuum, would have better construct validity.

No subsequent studies since Schiavetti et al. (1983) have been conducted to determine whether the continuum of stuttering severity is prothetic or metathetic. However, studies in other communication disorders have provided evidence that severity is also a prothetic continuum. For example, Eadie and Doyle (2002) determined the validity of voice pleasantness and overall voice severity ratings of dysphonic and normal speakers using DME and EAI scaling, revealing that voice severity is a prothetic continuum (Eadie & Doyle, 2002b). While SLPs use EAI scales in clinical practice because they are accessible, easy to use and interpret, and easy to compare across patients and listeners, their validity and reliability remain controversial. As an alternative to EAI scale ratings, ratio-based methods, such as direct magnitude estimation (DME) or visual analog scaling (VAS) have been suggested.

1.3.2 *Direct Magnitude Estimation Scaling*

Direct magnitude estimation (DME) is another scaling procedure that has been used in the perceptual judgment of stuttering severity. DME can be administered either with or without a modulus (Schiavetti, 1992). In the DME with-modulus procedure (DME-M), a number is

assigned to a standard speech sample, called the modulus. Listeners then rate all subsequent stimuli relative to the magnitude of the previously presented modulus. For free-modulus DME (DME-WM), listeners assign a number to the first speech sample they hear, and then rate all subsequent stimuli with numbers that are in proportion to the first sample they rated. DME with a standard is often preferred because listeners may be uncomfortable with scaling without a modulus and the handling of the data is easier when the scale is fixed with a standard (Engen, 1971).

One advantage of DME scales is that they do not have the systematic bias that EAI scales do, regarding whether the intervals are truly equal (Berry & Silverman, 1972; Engen, 1971; Stevens, 1975). Another advantage of DME scales is that statistical methods that require ratio level data can be applied to the data. On the other hand, DME is more difficult to implement in the clinical setting as it requires more explanation, training, and raters. Also, the presentation of stimuli is more complex if a modulus is used, and significant time and statistical resources are required for analysis (Baylis et al., 2015; Whitehill et al., 2002).

Comparison of EAI versus DME ratings for the evaluation of stuttering has had mixed results. Cullinan et al. (1963) compared the use of EAI and DME scaling in stuttering and found little difference in reliability of ratings and agreement of the mean scale values between the different scaling methods. While intrajudge reliability was the same for the ratings obtained by all scaling procedures, the interjudge reliability was considerably lower for DME scaling. They called for further investigation of the usefulness of various rating procedures, especially DME scaling (Cullinan et al., 1963). Schiavetti et al. (1983) who investigated the stuttering severity continuum, also compared the reliability of EAI and DME scaling. The raters in this study were 45 speech-language pathology undergraduate students enrolled in a course on stuttering. Fifteen

of the raters used an EAI scale, 15 used a DME scale with a standard/modulus, and 15 used a DME scale without a standard/modulus. The speakers were 20 adults whose speech samples included a broad range of stuttering severity as measured by frequency of nonfluencey and rate of speech. Reading samples included the first paragraph of the Rainbow Passage. Listener training included listening to audio recordings of three speech samples which varied in frequency of stuttering in order to familiarize the listeners with the range of stuttering severity. This study did not include an explicit definition of stuttering. Unlike Cullinan et al. (1963), the authors used a modulus for the DME scaling and a stimulus material that was standardized across participants. To date, the study by Schiavetti and colleagues has been the only study that evaluated the relationship between DME and EAI scaling of stuttering severity; they found it was curvilinear, indicating that stuttering severity is prothetic continuum (Schiavetti et al., 1983). They found that EAI scaling is not a valid scaling method for stuttering severity and that DME is preferable for measuring it. The authors called for serious reconsideration to be given to the widespread and indiscriminate use of EAI scales and for further study of the distinct advantages of the use of DME for measuring speech, language, and hearing variables. Lickley et al. (2005) utilized DME rating in their study that investigated disfluency in people who stutter and people who do not (Lickley et al., 2005). However, the purpose of the study was to compare the two groups rather than add to the existing literature on the validity and reliability of DME as a scaling method for stuttering severity. Therefore, the study was not considered to inform this literature review. Like Schiavetti et al. (1983), McColl and Fucci (2006) also assessed EAI and DME scaling reliability. In this study, 20 college students with no experience with stuttering rated ten speech samples produced by the same speaker. Every other sentence in each sample included a linguistically unnecessary pause. That is, the speech samples were produced by a speaker who did not stutter,

but purposefully paused within sentences as a stimulus. The raters rated each sample using EAI and DME scaling. They were instructed to rate the samples according to their perception of how disfluent the sample was, with disfluent speech being defined as a characteristic of stuttered speech. The authors of this study found significantly high correlations for both EAI and DME scaling to measure stuttering severity, indicating that either EAI or DME scaling can be used to accurately rate disfluency severity (McColl & Fucci, 2006).

This literature review showed disparity in the findings of the limited number of studies investigating the appropriateness of using DME and EAI rating scales for stuttering severity. Additional comparison of EAI and DME scaling is required and consideration of other ratio-based scaling methods, such as visual analog scales (VAS) should be taken. Details of the studies using DME to scale stuttering severity can be referenced in table 1 below.

Table 1. Studies utilizing direct magnitude estimation to scale stuttering severity.

Study	Modulus Type	Choice of Standard	Participants	Stimulus Material	Results
Cullinan et al. (1963)	Free	None	27 PWS; different undergraduate speech-language pathology students for seven different tasks (ranged from 14 to 19 across tasks)	20 sec spontaneous speech about a future vocation, ranging from very mild to very severe stuttering (SLP judged)	Intrajudge reliability was the same for EAI and DME ratings. Interjudge reliability was considerably lower for DME.
Schiavetti et al. (1983)	Free and modulus set at 10	Unspecified judge; average stuttering frequency	20 PWS; 45 undergraduate speech-language pathology students enrolled in a course on stuttering	First paragraph of the Rainbow Passage	DME scale values related to EAI scale values in curvilinear fashion, typical of prothetic continua, suggesting that DME is preferable to EAI scaling.

Lickley et al. (2005)	Modulus set at 10	Unspecified judge; moderately disfluent	20 PWS; 20 age- and gender-matched controls who did not stutter	50 utterances from recorded dialogs between PWS; 50 utterances from dialogs between PWDNS	Utterances produced by PWS were judged as "less fluent," and PWS perceives utterances as less fluent
McColl and Fucci (2006)	Free	None	1 speaker with typical speech; 20 judges who were naïve listeners (students)	10 speech samples by one speaker who did not stutter with varying pauses within sentences	Significantly high correlations for both EAI and DME scaling. Either scaling technique can be used

1.3.3 *Visual Analog Scaling*

Visual analog scales (VAS) are typically 100-mm undifferentiated lines with labeled endpoints. In measuring stuttering severity, one endpoint may be marked "no stuttering" and the other endpoint is marked as the maximum severity (e.g., "very severe stuttering", "most severe imaginable"). A listener indicates the severity of the stuttering by marking the point on the 100-mm line that corresponds with the perceived severity, with a higher value indicating greater. To date, no studies have explored the reliability of VAS to measure stuttering severity and there are no known advantages or disadvantages in listener reliability for VAS over other scaling methods. Given the lack of information on VAS with the purpose of measuring stuttering severity, more studies are needed to show whether VAS is at least as reliable and valid as other more commonly used scales, such as EAI and DME.

Recent studies in the voice disorders literature have shown that VAS is at least as reliable within and between listeners as EAI scales (Zraick & Liss, 2000). Given that severity is known to vary nonlinearly in voice disorders and stuttering (Eadie & Doyle, 2002b; Kempster et al., 2009; Schiavetti et al., 1983), VAS appears to be a preferable method of measuring severity.

VAS has also been shown to be more sensitive than EAI scales when measuring voice disorders (Kreiman et al., 2007).

VAS offers the rater a continuum of options, which differentiates it from discrete scales such as EAI. VAS has been suggested as a viable alternative to DME in voice research, having some similar psychometric properties, yet increased ease of use and ease of data analysis (Karnell et al., 2007; Kreiman et al., 2007; Zraick & Liss, 2000). Recent comparison of EAI versus VAS ratings when analyzing pediatric voice quality has shown the ease and benefits involved in the use of VAS ratings (Kelchner et al., 2010).

Given the advantages of using VAS shown in voice assessment research, there are potential advantages to using VAS to measure stuttering severity. However, there is a need for studies that determine whether VAS, a scaling technique that is not currently proven to be valid or reliable in rating stuttering severity, can yield the same or stronger reliability than the more commonly used methods of EAI and DME scaling for ratings of stuttering severity.

1.4 STUTTERING IN THE CONTEXT OF WHO-ICF FRAMEWORK

It is important to consider the role of measuring stuttering severity (or intensity, or frequency) as part of the diagnostic and therapeutic process. The World Health Organization International Classification of Functioning, Disability and Health (WHO-ICF) model is a useful framework for considering stuttering holistically. The ICF model incorporates contextual factors (environmental and personal) to show the impact stuttering can have on a person's ability to participate in daily activities and overall quality of life (Yaruss & Quesal, 2004). Many different factors, such as negative communication attitudes, shame, embarrassment, and limitations in an individual's ability to participate in society, can contribute to the experience of stuttering, in addition to observable characteristics (such as repetitions, prolongations, and hesitations that

may characterize stuttering) (Yaruss & Quesal, 2004). While this study focuses on the observable and physical characteristics of stuttering, it is important to note that there are other internal and external factors that contribute to a speaker's overall experience of stuttering.

In addition to the measurement of disruptions of stuttering for clinical and research purposes, clinicians may find it useful to use assessments that measure the impact of stuttering on people who stutter (PWS), such as the Overall Assessment of the Speaker's Experience of Stuttering (OASES) and the Unhelpful Thoughts and Beliefs about Stuttering scale (UTBAS) (Ward et al., 2021). These are self-report measures designed to assess the impact of various aspects of stuttering on the person who stutters. To comprehensively measure the severity of a person's stuttering as well as its impact on a person, SLPs need to use all these types of measures (i.e., one is complementary to the other). This study will focus on one type of measure that captures the disruptions in the speech itself.

Results of this study have implications for stuttering assessment and treatment, as there are some situations in which it may be helpful for a speech-language pathologist and a person who stutters and their family to have a shared understanding of the degree of a client's physical experience of stuttering. While some clinicians are moving away from the use of standardized assessments (i.e., SSI-4), there will likely continue to be a need to capture the nature and severity of physical characteristics of stuttering for diagnostic and research purposes. Such measurements should be reliable and valid to be used appropriately in research and in the clinic.

1.5 Study Overview and Hypotheses

It is essential that procedures commonly utilized in the measurement of stuttering severity are shown to be both valid and reliable. The primary aim of the current study was to examine the

validity and reliability of EAI, DME, and VAS scaling for perceptual rating of stuttering severity. A comparison of EAI, DME, and VAS scaling procedures was done, using speech samples from individuals who stutter, to determine the most appropriate rating procedure in both research and clinical contexts. Prior to conducting the study, the following hypotheses were proposed:

1. The set of direct magnitude estimation scale (DME) values will relate to the equal appearing interval scale (EAI) values in a curvilinear manner, rather than a linear one, indicating that stuttering severity is a prothetic continuum. In addition, when VAS ratings are plotted against EAI ratings, results will also reveal a curvilinear relationship. Lastly, when VAS ratings are plotted against DME ratings, results will also reveal a linear relationship, as both DME and VAS may be considered ratio scales, consistent with previous research in speech science. Curvilinear relationships between EAI and DME or between EAI and VAS would indicate that stuttering severity is a prothetic continuum, supporting the findings of Schiavetti (1983), and suggesting that ratio scaling methods, such as DME and VAS, are preferable to interval scaling, such as EAI, for measuring stuttering severity.
2. DME and VAS will result in higher or at least as reliable interrater and interrater reliability in comparison to EAI, suggesting that DME and VAS should be considered as appropriate alternatives to EAI ratings for measuring listener perceptions of stuttering severity in adults.

2. METHODS

2.1 PARTICIPANTS (LISTENERS)

Forty-five individuals who are inexperienced with stuttering severity assessment (hereafter referred to as the "listeners") served as listeners in this study. Inexperienced listeners provide a model of an unfamiliar communication partner. Importantly, inexperienced listeners rating overall severity show no differences compared to speech-language pathology students, practicing SLPs or SLPs who specialize in stuttering (Amir et al., 2018). Thus, all listeners had little or no formal coursework related to fluency disorders. All listeners were native English speakers (i.e., acquired English before the age of 2) between the age of 18 and 45. Listeners passed a hearing screening at passed a pure-tone hearing screening at 25 dB hearing level (HL) for pure tones at 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz unilaterally. Listeners were primarily recruited from the University of Washington (UW) community and Seattle metropolitan area. They attended a 30–60-minute session at the Quantitative Imaging for Learning, Language, & Speech (QuILLS) lab in the Department of Speech and Hearing Sciences at the University of Washington, had the option to withdraw from the study at any time, and were compensated for their time.

All listeners completed a screening form in which information about their language background, history of childhood and/or current speech/language/reading/hearing concerns, and presence of other medical or neurological diagnoses were collected. Listeners provided written informed consent according to the protocol approved by the University of Washington IRB.

2.2 SPEECH STIMULI

Twenty previously obtained audio recordings served as stimuli for the speech ratings for this study. The speech samples were acquired from corpora of recordings from teenage and adult speakers who stutter available from FluencyBank (28 speakers aged 24-62 years). Readings were a center portion of the Friuli reading passage from the SSI-4 for each speaker of around 110 syllables. Please see Appendix A for the center portion of the reading passage. Though it is recommended to use a sample of 150 syllables (Riley, 2009), a portion of 110 syllables was chosen to optimize the total listening time and reduce likelihood of poor reliability due to listener fatigue. The samples were collected on typical clinical equipment (video cameras). Samples were downloaded as videos from FluencyBank in .mp4 format. Audio was extracted from the video with custom MATLAB software using the *audioread* function. Audio was all compressed in .mp4 format with AAC compression and recorded at a high-quality sampling rate (44.1k or 48k). Stimuli were extracted and cropped to manual audio markings with .25s hann windows on each end to avoid abrupt onset and offset. Cropped stimuli were then peak amplitude normalized and saved as .wav files. Exported stimuli were uncompressed with 16 bits per sample, with a 44.1k or 48k sampling rate as determined by the original quality. Samples were audio only to simplify ratings and correlate to acoustic measures; previous studies show that severity ratings are similar with and without video (Martin & Haroldson, 1992; Williams et al., 1963). Samples were selected from all the available samples to reflect a range of stuttering severity, ranging from very mild to severe. Samples were rated by a student graduate clinician experienced in clinical assessment of stuttering using the SSI-4. The reading percentage of stuttered syllables (%SS), reading task score, average length of three longest stutters, and duration score were obtained for each sample to gauge the severity of each sample as speaking samples were not used in this

study. The reading percentage of stuttered syllables ranged from 1.82% to 28.18% and the reading task score ranged from 2 to 9. Average length of three longest stuttering instances ranged from 0.5 to 7 seconds, with duration scores ranging from 2 to 12. These ranges of reading task and duration scores are indicative of a range of very mild to severe stuttering in the chosen speech samples. Please see Table 2 for the stimulus characteristics.

Table 2. Stimulus Characteristics of Reading Samples

Sample	Reading %ss	Reading task score	Average length of 3 longest stuttering moments	Duration score
1	3.64	5	0.5	2
2	4.55	5	0.5	2
3	9.1	7	2.67	8
4	16.36	8	3.33	10
5	4.55	5	0.67	4
6	3.64	5	1	6
7	2.73	4	0.5	2
8	2.73	4	0.67	4
9	18.18	8	3.67	10
10	23.64	9	3.33	10
11	26.36	9	3.67	10
12	13.64	8	7	12
13	11.82	7	7	12
14	1.82	2	0.5	2
15	10.91	7	0.5	2
16	28.18	9	3.67	10
17	3.64	5	0.67	4
18	19.1	8	3.33	10
19	8.18	7	6.67	12

20	1.82	2	0.5	2

2.3 RATING PROCEDURES

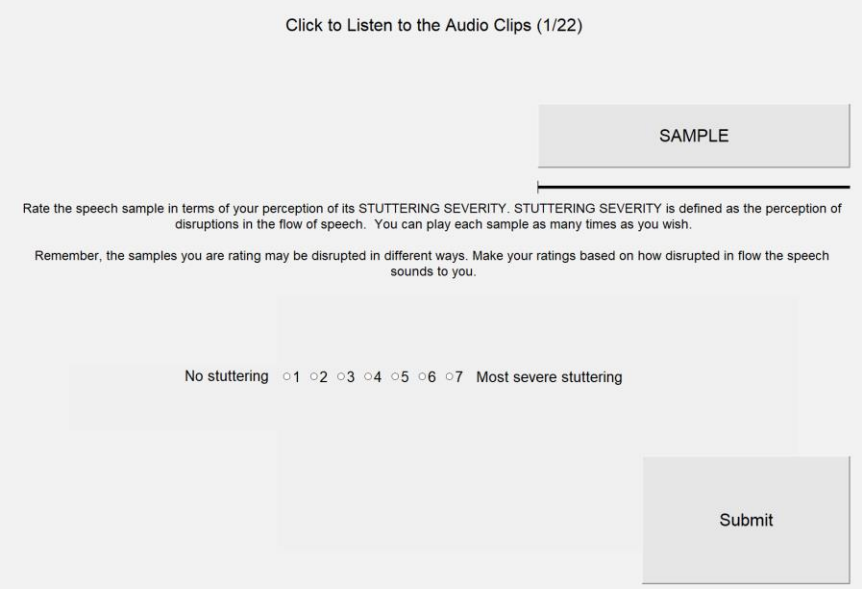
Every listener completed a brief training at the start of the session. They read definitions of stuttering and listened to examples of varying levels of severity. The definition was as follows: *stuttering is a speech disorder that involves disruptions in the flow of speech. People who stutter know what they want to say but have difficulty saying it. For example, you may hear a speaker: have trouble starting a sound (e.g., "I want *cake" *=a pause with tension), extend a sound (e.g., "I wwwwant cake"), or repeat sounds or syllables (e.g., "d d d dog" or "ba ba ba baby").* Unused samples from FluencyBank were chosen as training stimuli to prevent exposure to the rating stimuli, representing a range of stuttering severity.

Listeners received a randomized order in which to listen to the 20 samples to prevent order effects across listeners. Additionally, they listened to 3 repeated samples (15% of the total number of samples) to assess intrarater reliability. Listeners were allowed to listen to each speech sample as many times as they wished at a comfortable volume. Each listener was assigned to one of the three rating conditions. Rating conditions were assigned so that each one of the three groups had a similar mean age and number of listeners who identified as men, women, and non-binary/genderqueer people. The EAI group had a mean age of 20.4 and consisted of 4 men and 11 women. The VAS group had a mean age of 23.5 and consisted of 4 men, 10 women, and 1 genderqueer person. The DME group had a mean age of 20.9 and consisted of 4 men, 9 women, and 2 non-binary people. Listeners were then provided with task-specific training on use of the given rating scale. All ratings were automatically collected by the survey software.

2.3.1 *EAI Task*

A seven-point equal-appearing interval (EAI) scale was used (see Figure 1). Listeners listened to each speech sample and were instructed to rate the severity of stuttering on a scale where “1” equals no stuttering and “7” equals most severe stuttering. 15% of the samples were randomly presented and rated a second time by each listener for reliability. Listeners could listen to each speech sample as many times as needed.

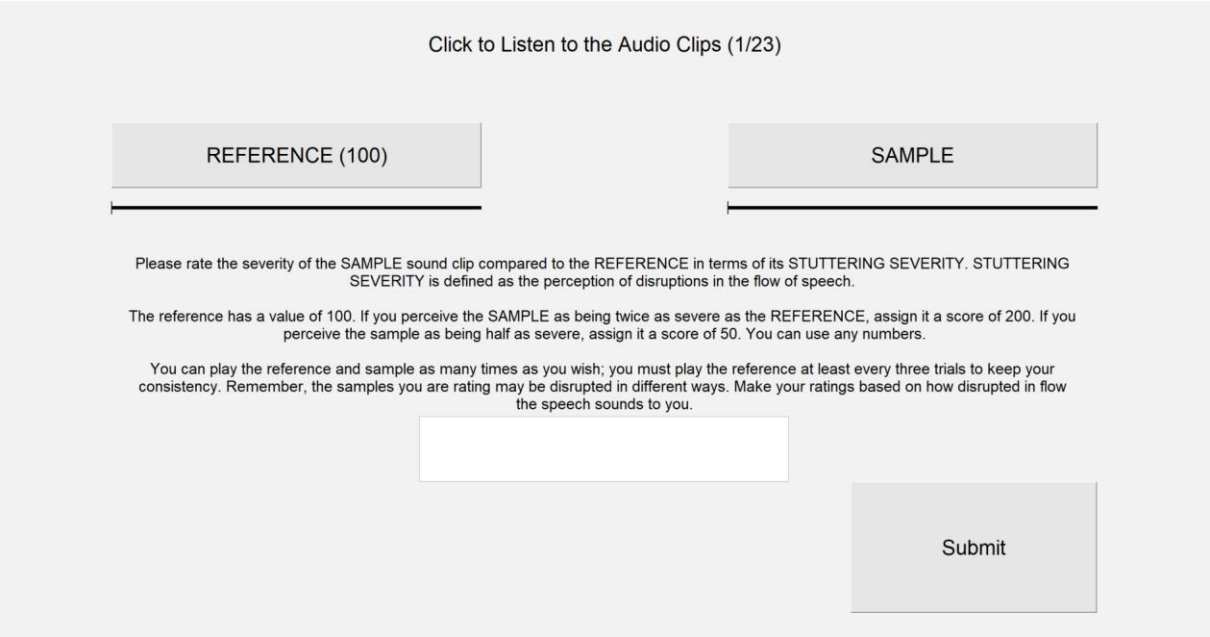
Figure 1: EAI condition: Listeners were asked to rate a sample audio sample on 7-point equal appearing interval scale.



2.3.2 *DME Task*

For the DME condition (see Figure 2), listeners were instructed to listen to the modulus stimulus first and rate each speech sample in relation to a standard, which was assigned a modulus value of 100. The standard was chosen by the author of this thesis to represent mild-moderate stuttering severity (5.94 percent stuttered syllables) and was played at the beginning of the task and then after every three samples to maintain consistency. Listeners could listen to the standard at any time throughout the task. Next, participants listened to each speech sample and assigned a rating based on its comparison with the modulus. They were instructed that if they perceived a speech sample twice as severe as the modulus, they should rate the sample 200. If they perceived it to be half as severe, they should rate it 50. They were told that they could assign any number to the sample. 15% of the samples were rated twice by each listener for reliability. Raters listened to the modulus and the sample as many times as they wished.

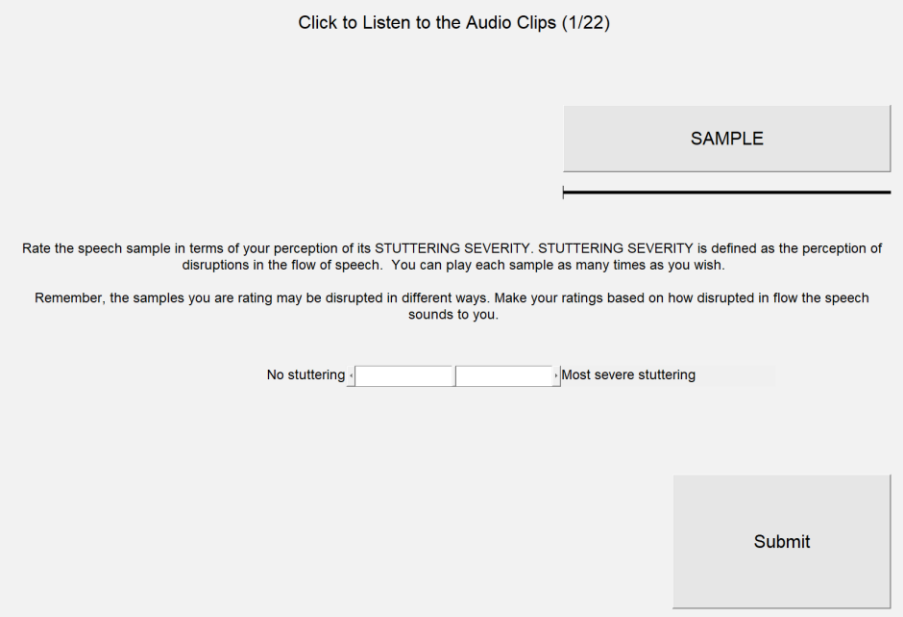
Figure 2: DME condition: Listeners were asked to rate a sample audio sample in comparison to a modulus set at 100. The DME score was unrestricted.



2.3.3 *VAS Task*

For the VAS ratings (Figure 3), listeners entered their rating of the stuttering severity by using their computer mouse to drag a sliding bar along the scale, which reflected a 100-mm line on the screen. Listeners were instructed to drag the bar along the scale to indicate their rating for severity of stuttering; the endpoints of the bar were labeled “no stuttering” and “most severe stuttering”. They were instructed to use the ends of the scale when appropriate.

Figure 3: VAS condition: Listeners were asked to rate a sample audio sample by dragging a sliding bar along a scale with the ends labeled “no stuttering” and “most severe stuttering”.



2.4 STATISTICAL ANALYSES

All ratings were automatically collected by the software and were saved locally for further analysis into Excel spreadsheets for each listener. Initial data cleaning and statistical data analyses were performed in R (4.0.2).

To determine which of the three rating methods resulted in higher reliability, interrater and intrarater reliability were computed for the three rating conditions. Pearson's correlations were used for intrarater reliability. Intraclass correlation coefficients (ICC) were calculated, with two-way mixed average measures consistency ICC(3,k) calculated for interrater reliability (Shrout & Fleiss, 1979). Given that Shrout and Fleiss (1979) stated that ICC(3,k) is appropriate when listeners rate the same n targets because listeners are viewed as fixed effects, ICC(3,k) was used in this study to measure interrater reliability (Shrout & Fleiss, 1979).

Next, EAI mean ratings were plotted as a function of the VAS mean ratings and as a function of the DME geometric mean ratings. VAS mean ratings were also plotted as a function of the DME geometric mean ratings according to the procedures of Stevens (1975) and

Schiavetti et al. (1983). Mean ratings were plotted to determine whether the values of each scale are related in a linear fashion, indicating a metathetic continuum, or in a curvilinear fashion, indicating a prothetic continuum (Schiavetti et al., 1983). The three scatter plots were visually inspected to determine whether the relationship between EAI vs. DME, EAI vs. VAS, and VAS vs. DME was linear or curvilinear. A linear regression and polynomial regression were completed for each of the three plots to determine which model accounted for more variance. Then, an ANOVA was done to compare the two models to determine which resulted in a better fit.

3. RESULTS

Forty-five listeners rated 20 samples with an additional three repetitions (15% of total samples) using one of three rating scales: EAI, VAS, or DME. Scales are compared in terms of intrarater reliability, interrater reliability, and construct validity by examining the relationships between EAI and VAS both compared to DME.

3.1 INTRARATER RELIABILITY

Pearson's correlations were computed per listener based on the three repeated samples (see Figure 4). The Pearson's correlation across the samples ranged from -0.87 to 1.0 for EAI ratings, with an overall average correlation of 0.2. Pearson's *r* could not be calculated for 7 of the EAI ratings as listeners gave them the same rating for all three samples (either originally or at repetition, despite the severity range in chosen repetitions). The intrarater reliability across the samples ranged from -0.98 to 0.92 for DME ratings, with an average reliability of -0.09 overall. Pearson's *r* could not be calculated for 3 of the DME ratings. The reliability ranged from -0.98 to 1.00 for VAS ratings, with an average reliability of -0.09 overall. The reliability varied between raters regardless of rating scale, from positive large association (VAS: 1.0, DME: .92, EAI: 1.0) to negative large association (-0.87 to -0.98 for all scales).

The reliability for all three scales had a very large range and mean near 0. We suggest that this is likely caused by the low number of repeated samples (3 samples per listener) leading to unstable estimates of intrarater reliability.

Figure 4. Pearson's *r* values for EAI, VAS, and DME

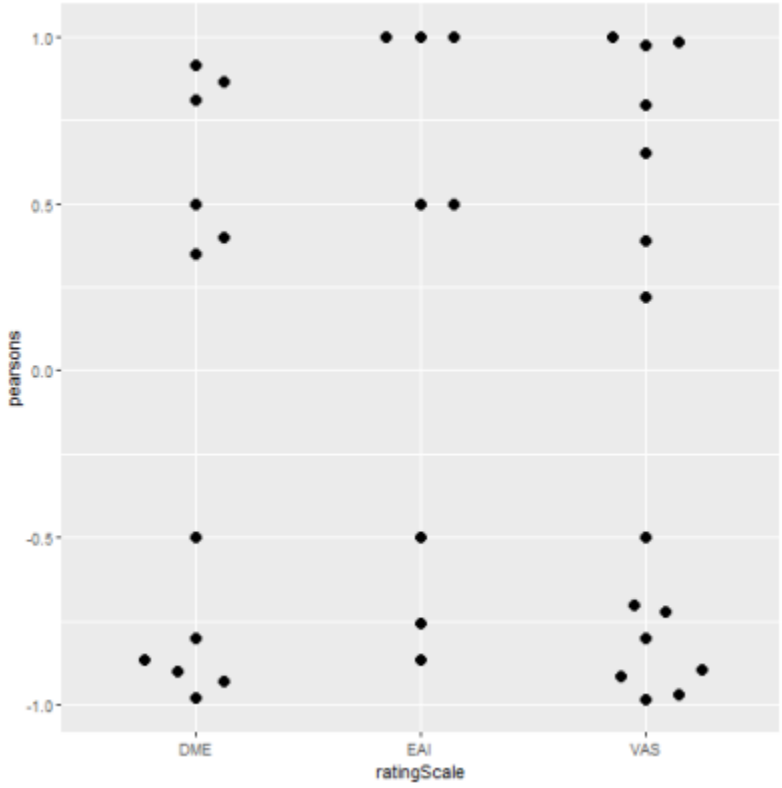


Table 3. Summary of Statistical Analysis of Intrarater Reliability (Pearson's correlation) of EAI, VAS, and DME Ratings of Stuttering Severity

Scale	Mean	Minimum	Maximum	Number of samples with null correlation**
EAI	0.2	-0.087	1.0	7
VAS	-0.09	-0.98	1.0	0
DME	-0.09	-0.98	0.92	3

**Could not be calculated due to lack of variation in ratings

3.2 EXCELLENT INTERRATER RELIABILITY

Single-measures and average-measures ICCs were computed to measure interrater reliability to compare to Schiavetti et al. (1983). As anticipated, interrater reliability for the EAI and VAS ratings were higher than those obtained for the DME ratings, although all reliability measures fell within the excellent level of agreement, except for the DME single-measures ICC (EAI single-measures ICC = 0.84, average-measures ICC = 0.99; VAS single-measures ICC = 0.85, average-measures ICC = 0.99; DME single-measures ICC = 0.51, average-measures ICC = 0.94). See Table 4 for results.

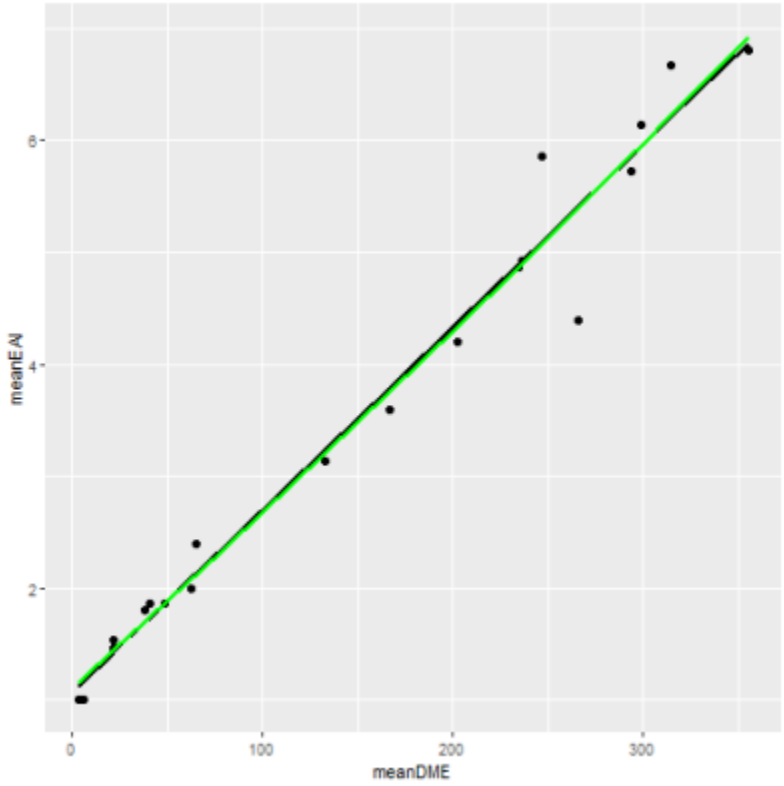
Table 4. Summary of Statistical Analysis of Interrater Reliability of EAI, VAS, and DME Ratings of Stuttering Severity

Scale	Average-Measures ICC	Level of Agreement	Single-Measures ICC	Level of Agreement
EAI	0.99	excellent	0.84	excellent
VAS	0.99	excellent	0.85	excellent
DME	0.94	excellent	0.51	fair

3.3 Model of Best Fit – EAI versus DME

Figure 5 shows the EAI data plotted against the DME data to determine the model of best fit. Visual inspection reveals a linear relationship between the two rating scale measures. The linear model was $y= 0.016 x+1.074$ with an $R^2 = 0.9702$. The curvilinear model ($y = .00003x^2 +.015x +1.107$) accounted for the same amount of variance ($R^2 = 0.9686$). An ANOVA comparing these two models did not result in a significantly better fit for the curvilinear model, above and beyond the variance predicted in the linear model ($F= 0.0955$; $p= 0.761$). Thus, we conclude that the linear relationship is the better fit.

Figure 5. Scatterplot of geometric mean of DME ratings versus mean EAI ratings of stuttering severity. Linear model trendline shown in black and polynomial fit shown in green.

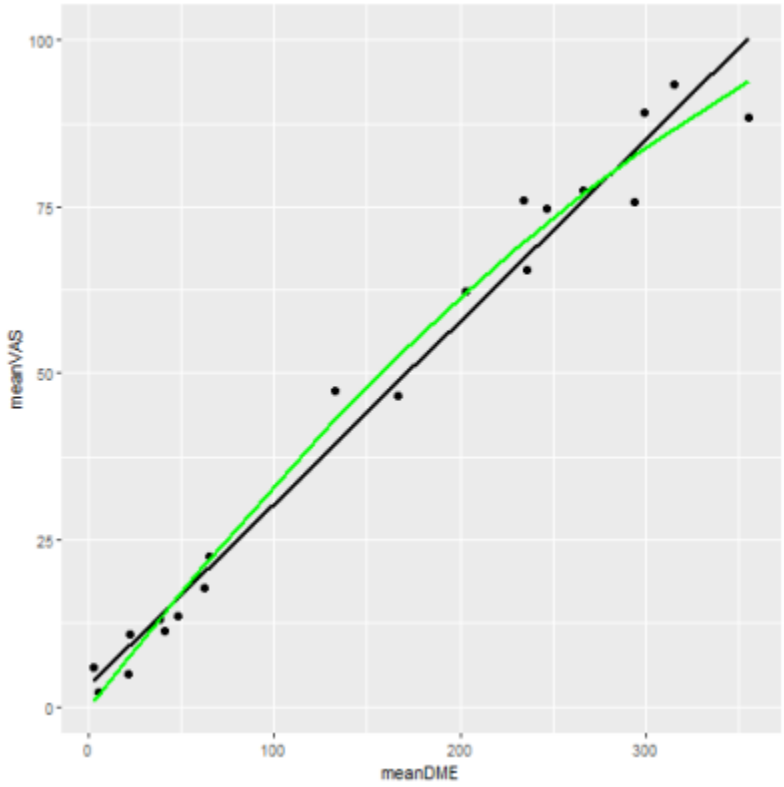


3.4 Model of Best Fit – VAS versus DME

Figure 6 shows the VAS data plotted against the DME data to determine the model of best fit. Visual inspection reveals a linear relationship between the two rating scale measures. R^2 values for the models were similar (linear model y = 0.274x + 3.109, R^2 = 0.976; curvilinear model y = -0.334x2 + 0.363x - 0.334; R^2 = 0.9816). An ANOVA comparing the two models was significant (F = 6.4, p < 0.022). An additional model including a cubic factor revealed the equation y= 2.47+.238x- .0006x2- .0000016x3; R^2= 0.99. The ANOVA comparing the models was not significant (F=2.48; p>0.13). We conclude that the second order curvilinear model

provides a slightly better fit than the linear model (~0.6% improvement in variance than the linear).

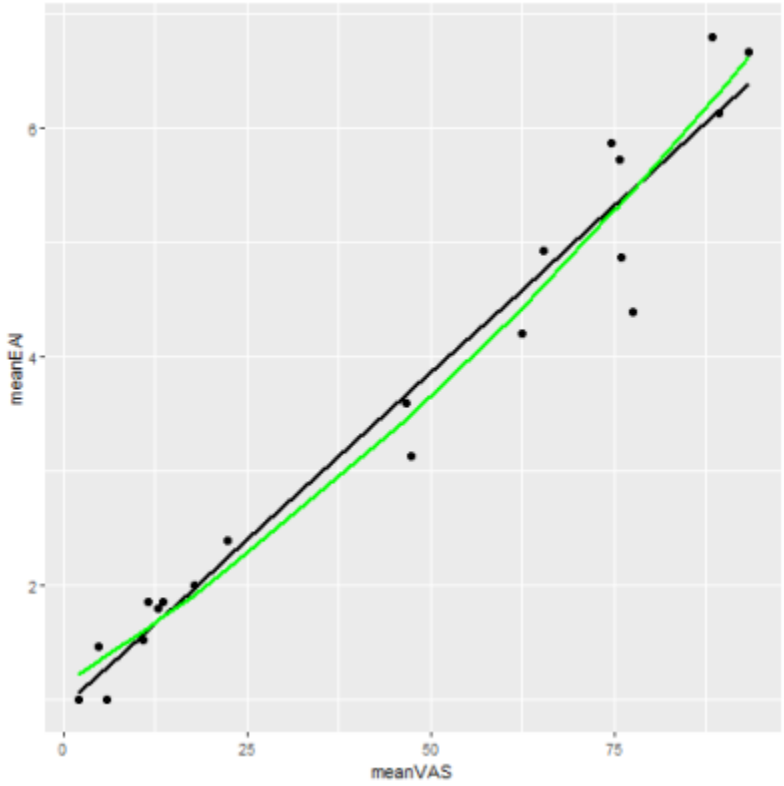
Figure 6. Scatterplot of geometric mean of DME ratings versus mean VAS ratings of stuttering severity. Linear model fit shown in black and polynomial fit shown in green.



3.5 Best Fit – VAS versus EAI

Figure 7 shows the VAS data plotted against the EAI data to determine the model of best fit. Visual inspection reveals a linear relationship between the two rating scale measures. R^2 values for the models were similar (linear model $y = 0.059x + 0.938$, $R^2 = 0.9553$; curvilinear model $y = 0.0002x^2 + 0.04x + 1.14$, $R^2 = 0.9572$). An ANOVA comparing the two models was not significant ($F= 1.78$, $p=0.2$). We conclude that the linear model is a better fit.

Figure 7. Scatterplot of mean VAS ratings versus mean EAI ratings of stuttering severity. Linear model fit shown in black and polynomial fit shown in green.



4. DISCUSSION

The current study examined the validity and reliability of EAI, DME, and VAS scaling for perceptual rating of stuttering severity to determine the most appropriate rating procedure in research and clinical contexts. The results are interpreted in the following section. Implications for both research and clinical practice, limitations of the current study, and future directions are discussed.

4.1 IN CONTRAST TO SCHIAVETTI ET AL. (1983)

A linear model fit the data for EAI versus VAS and EAI versus DME, and a curvilinear model fit the data for VAS versus DME (the curvilinear model provided a slightly better fit than the linear model with a ~0.6% improvement in the predicted variance). Thus, we conclude that stuttering severity, as it was defined in this study (i.e., disruption in the flow of speech), is consistent with a metathetic continuum (EAI vs DME as well as VAS vs EAI ratings). As such, our study indicates that direct magnitude estimation, visual analog scaling, and equal appearing interval scaling may all be appropriate for measuring stuttering severity. The results of the current study are inconsistent with the results of Schiavetti et al. (1983) in regards to construct validity. They found that stuttering severity is a prothetic continuum, as the DME scale values were related to EAI values curvilinearly, with a nonlinear regression procedure accounting for 95.70% of the variance in the data for EAI versus DME with standard/modulus. Thus, they concluded that EAI was not an appropriate scaling mechanism. Here we found that EAI and DME ratings are related in a linear-fashion, suggesting a metathetic continuum. These results from Schiavetti et al. (1983) suggested that DME is preferable to interval scaling for measuring stuttering severity, whereas the current study found that EAI may be valid as well.

Both the current study and the Schiavetti et al. (1983) study resulted in high interrater reliability. The intraclass correlation procedure values for Schiavetti et al. (1983) for the mean ratings of each group were 0.98 for EAI scaling and 0.96 for DME with standard/modulus. Our results showed high interrater ICCs as well, with 0.99 for EAI and = 0.94 for DME; VAS was also high, at 0.99.

In terms of intrarater reliability, Schiavetti et al. (1983) had high intraclass correlation values at 0.75 for EAI scaling, 0.61 for DME with standard/ modulus, and 0.65 for DME without standard/modulus. Our intrarater reliabilities were not interpretable, since they were unstable due to too few repeated samples.

Possible explanations for the discrepancies between these studies may be found in the methodologies of the studies. For subjects, Schiavetti et al. (1983) recruited listeners who were enrolled in an undergraduate course in stuttering, while the listeners in the present study were naïve and had no education in stuttering. Previous studies have shown that naïve listeners are capable of judging stuttering with comparable intrarater and interrater reliability results, though rating differences may exist when severity differences are mild and perceptually less noticeable to a naïve listener (Amir et al., 2018). However, there is a lack of research on the reliability of student ratings in comparison to naïve or expert listeners. Thus, listeners in this study and the Schiavetti et al. (1983) study were relatively heterogenous with respect to their knowledge of stuttering, likely limiting the ability to effectively compare the reliability results from each study. Additionally, another difference in the methodologies may have been the definition of stuttering. This study defined stuttering as *the disruption of flow of speech*. On the other hand, Schiavetti et al. (1983) did not provide an explicit definition of stuttering, though they referred to the frequency of nonfluencey and rate of speech when discussing the range of severity. The current

study's definition of stuttering is more similar to the definition provided by McColl and Fucci (2006), which was disruption in speech. This study's construct may be more similar to the construct measured in the study by McColl and Fucci, who also found that stuttering severity is a metathetic continuum. Since this study's definition of stuttering focused on speech flow, which was similar to that of McColl and Fucci (2006), the results of this study are consistent with that previous study.

Additionally, over the past forty years, since the Schiavetti et al. (1983) study was completed, quality of audio recordings has improved and the ability to hear subtle differences in speech that may present in stuttering has been enhanced, allowing us to record speech samples that are of a higher quality. Lastly, public awareness of stuttering has increased over the past forty years, since the Schiavetti et al. (1983) study and overall knowledge and understanding of stuttering has become less limited. These differences in perception over the past forty years may have shifted the auditory perception of stuttering severity, with listeners rating samples as less severe now than they may have forty years ago.

4.2 Interpretation of relationship between VAS and DME

Although our main hypothesis was related to replicating the Schiavetti EAI vs DME result, this study added a VAS scale as an additional ratio scale (like DME) that is simpler than DME for raters to use. We hypothesized that stuttering severity measured on these two ratio scales would be linearly related regardless of whether stuttering severity was found to be prothetic or metathetic, given that both ratio scales should scale either type of continuum. As expected, we did find a strong linear fit between VAS and DME (R^2=0.976). However, we also

found that a curvilinear fit was statistically significantly better (R^2=0.9816). This indicated the curvilinear model was 0.6% better than the linear fit.

There are two possible interpretations of these results. The first is that in this case, VAS does not act as a ratio scale, but more as an ordinal scale. The curvilinear relationship with DME would indicate then that stuttering severity is a prothetic continuum and only DME is appropriate for scaling. Evidence countering this interpretation is that EAI and DME were linearly related, suggesting that stuttering severity is in fact metathetic; other evidence also indicates that VAS is usually considered ratio scaling in other auditory-perceptual ratings (Stevens, 1975). The other possible interpretation of this evidence is that while a .6% difference in variance might indicate a statically significant improvement from the linear model, it is not of functional consequence. Thus, we would conclude that the linear fit still characterizes the continuum. Future research is needed to reach a conclusion.

4.3 Clinical and research Implications

There are some potential clinical implications of the current study related to the assessment of stuttering severity in clinical and research settings. Given the results of the current study, it appears that when listeners are asked to judge stuttering severity in a research setting, mean values for each speaker may be used across listeners when using any type of rating scale: EAI, VAS, or DME. Given the inconsistency between this study and previous studies regarding construct validity of EAI and whether stuttering severity is a prothetic or metathetic continuum, more research is needed to determine the continuum of stuttering severity and how to best measure it.

The results of this study showed that naïve listeners demonstrated wide ranges of intrarater reliability, with means near 0, regardless of the rating scale. In other words, when rating the

stuttering severity of a speaker twice, the second ratings were inconsistent and did not agree with the first ratings. A problem with our methodology was the limited number of repeated samples; 15% of the samples was only 3 data points. As such, the intrarater reliability estimations are likely unstable and impossible to interpret. In any case, intrarater reliability was universally poor across rating scales. In future studies, we would recommend repeating a larger percentage of samples.

However, given the reliability *between* listeners (i.e., interrater reliability), it is likely that the subjective ratings are reasonable estimates of "ground truth" of stuttering severity according to naïve listeners. Therefore, in research contexts, a group of raters may use more simple rating scales (i.e., either EAI or VAS) rather than the more complicated DME procedures to produce reliable and accurate measures of listener-perceived stuttering severity.

It should be considered that the current study's raters were naïve listeners, and only a short training period was provided to the listeners prior to rating the samples. Further, we are inherently measuring the reliability of *averaged* estimates. We did not ascertain the reliability or accuracy of single estimates (either by clinicians or naïve listeners) for measures of interrater reliability.

4.4 Limitations of the Current Study

This study has several methodological limitations that should be considered before findings can be generalized to other research and clinical settings. First, there are limitations with the speech stimuli. Though it is recommended to use a sample of 150 syllables (Riley, 2009), a portion of 110 syllables was used in this study and only 20 speech samples were used to optimize total listening time and reduce likelihood of poor reliability due to listener fatigue. Though the speaking samples were uniformly distributed across mild, moderate, and severe stuttering severity, no samples were included from speakers who are typically fluent. When inspecting the figures, there are data gaps from about 20 to 50 in the mean VAS data. Although one trained

rater's measurements of percent dysfluency were used to select samples across a range of severity, these results suggest that naïve listeners rated samples as more towards the ends of the range rather than in the middle.

Additionally, samples were audio-only to simplify ratings and correlate to acoustic measures in future studies. Although previous studies show that severity ratings are similar with and without video, audio-visual samples can provide additional information that is not available with audio only samples, such as onset of prolongations of sounds, prolonged pauses between sounds and word, and physical concomitants like eye blinking, jaw jerking, and head movements (Martin & Haroldson, 1992; Williams et al., 1963). Because the reading speech samples were taken from an existing corpus, limited information was provided on the demographics of the speakers, other than age and sex. More information on the demographics of the speakers may have resulted in increased generalization to a wider variety of clinical or research population. Additionally, following the methods of Schiavetti et al. (1983), listeners spoke English as their first language and had no previous education in fluency disorders. We chose this as a proxy for communicating with an unfamiliar listener over the phone. Therefore, our raters may not act as a model for practicing speech-language pathologists, despite some previous evidence that naïve listeners and SLPs are similarly reliable when rating stuttering severity (Amir et al., 2018). This likely means that they are similarly reliable when compared to other naïve listeners, but they might have been rating the stuttering severity as more or less severe than practicing speech-language pathologists would. Additionally, this study used reading samples only, similarly to Schiavetti et al. (1983). Given that clinicians and researchers use both spontaneous speaking samples and reading samples, generalization may be limited given the discrepancies of stuttering severity between speaking and reading samples that can exist within each PWS, with some PWS

more fluent in one of the tasks, and some more fluent in the other. Thus, further study is needed to determine generalization to spontaneous speaking samples.

Second, as stated earlier in the discussion, no extensive training was provided to the listeners in this study in order to optimize the listening time and avoid fatigue which may have influenced the reliability. While EAI scaling does not require training as most listeners are familiar with this type of scaling, VAS and DME scaling require more explanation and training. It is possible that more extensive training of listeners utilizing the EAI, DME, and VAS scales in the current study may produce a more acceptable level of intrarater reliability, similarly to other studies that implemented extensive training and education prior to the task (Bainbridge et al., 2015). However, the current study likely had an issue with measuring intrarater reliability given that all scales had similarly poor reliability, so it is unclear what the "true" level of intrarater reliability was here, and how training would affect that in future.

Third, the use of 15listeners for each scaling procedure may translate to research contexts, but may not be generalizable to a clinical setting, where it is not feasible to have 15 listeners assess stuttering severity. Thus, it is necessary to determine the minimum number of listeners that are needed for an acceptable level of reliability in order to inform the use of the different scaling procedures in clinical settings.

This study took place in a controlled environment which may not simulate clinical environments. Listeners listened to speech samples in a comfortable volume in a sound booth using headphones, which may not be practical for clinicians in clinical settings. This is another limitation which may limit generalization of the results of the current study to clinical settings.

Finally, although this study contributes to the continued need to capture the nature and severity of physical characteristics of stuttering in clinical and research settings, it is crucial to

consider stuttering within the context of how it may impact a person's ability to participate in daily activities and their overall quality of life. In order to holistically assess stuttering in adults, both standardized assessments that capture the severity of stuttering and measurements that holistically inform treatment decisions and give insight into the impact of stuttering on a person's quality of life should be utilized in clinical and research settings.

4.5 Future Directions

Future directions are based on the inconsistency between the results of the current study and previous studies regarding the construct validity of each rating scale. Schiavetti et al. (1983) found that stuttering severity is a prothetic continuum and DME or other ratio scaling methods are needed, while the current study found that stuttering severity is a metathetic continuum and all rating scales (i.e., EAI, DME, and VAS) are appropriate for rating it. Given that these results are based on ratings derived from naïve listeners, future research should focus on investigating the validity and reliability of these scales using expert listeners, such as speech-language pathologists and researchers in the field of fluency disorders to determine whether results are similar or different. Additionally, given the poor intrarater reliability for all rating scales, future research should include more repeated samples to provide more accurate estimates of reliability. Then a training period may be needed to achieve higher reliability. Lastly, future research should investigate the minimum number of listeners required for acceptable interrater reliability to ensure that the use of each of the rating scales is feasible for assessment in clinical and research settings where the number of raters is limited.

5. CONCLUSION

This study provides an investigation of the validity and reliability of EAI, DME, and VAS scaling for perceptual rating of stuttering severity by comparing the three scaling procedures using speech samples from PWS. It provides additional evidence for determining appropriate rating procedures in both research and clinical contexts and addressing the disparity in the findings reported in previous research on the topic in an effort to potentially improve current clinical practices. Reading samples were acquired from corpora of recordings from teenage and adult speakers who stutter available from FluencyBank, and listeners rated the samples using a randomly-assigned scaling procedure. Results demonstrated good interrater reliability for all three rating procedures. Regression models indicated that linear relationships significantly fit the data; curvilinear relationships also fit the data though not significantly above and beyond the linear fits, providing evidence that stuttering severity, as it was defined in this study, was consistent with a metathetic continuum. For metathetic percepts, a variety of scales can be used to validly capture the construct, including EAI scales. Results suggest that in research contexts, having a group of raters use either EAI or VAS scales will result in excellent interrater reliability and thus accurate measures of listener-perceived stuttering severity. Additional research utilizing trained expert listeners with a background of knowledge in fluency disorders will provide results generalizable to the clinic that an increased number of clinicians and researchers will be able to use in research and clinical practice in a valid and reliable manner.

References

Amir, O., Shapira, Y., Mick, L., & Yaruss, J. S. (2018). The Speech Efficiency Score (SES): A time-domain measure of speech fluency. *Journal of Fluency Disorders*, *58*, 61–69. https://doi.org/10.1016/j.jfludis.2018.08.001

Ashurst, J. V., & Wasson, M. N. (2011). Developmental and persistent developmental stuttering: An overview for primary care physicians. In *Journal of the American Osteopathic Association* (Vol. 111, Issue 10).

Bainbridge, L. A., Stavros, C., Ebrahimian, M., Wang, Y., & Ingham, R. J. (2015). The efficacy of stuttering measurement training: Evaluating two training programs. *Journal of Speech, Language, and Hearing Research*, *58*(2). https://doi.org/10.1044/2015_JSLHR-S-14-0200

Baylis, A., Chapman, K., & Whitehill, T. L. (2015). Validity and reliability of visual analog scaling for assessment of hypernasality and audible nasal emission in children with repaired cleft palate. *Cleft Palate-Craniofacial Journal*, *52*(6), 660–670. https://doi.org/10.1597/14-040

Berry, R. C., & Silverman, F. H. (1972). Equality of intervals on the Lewis-Sherman scale of stuttering severity. *Journal of Speech and Hearing Research*, *15*(1). https://doi.org/10.1044/jshr.1501.185

Blood, G. W., & Blood, I. M. (2016). Long-term Consequences of Childhood Bullying in Adults who Stutter: Social Anxiety, Fear of Negative Evaluation, Self-esteem, and Satisfaction with Life. *Journal of Fluency Disorders*, *50*, 72–84. https://doi.org/10.1016/J.JFLUDIS.2016.10.002

Bothe, A. K., Davidow, J. H., Bramlett, R. E., & Ingham, R. J. (2006). Stuttering treatment research 1970-2005: I. Systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. In *American Journal of Speech-Language Pathology* (Vol. 15, Issue 4). https://doi.org/10.1044/1058-0360(2006/031)

Carter, A., Breen, L., Yaruss, J. S., & Beilby, J. (2017). Self-efficacy and quality of life in adults who stutter. *Journal of Fluency Disorders, 54*. https://doi.org/10.1016/j.jfludis.2017.09.004

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4). https://doi.org/10.1037/1040-3590.6.4.284

Connery, A., Galvin, R., & McCurtin, A. (2021). Effectiveness of nonpharmacological stuttering interventions on communication and psychosocial functioning in adults: A systematic review and meta-analysis of randomized controlled trials. *Journal of Evidence-Based Medicine*, *14*(1). https://doi.org/10.1111/jebm.12408

Cordes, A. K., & Ingham, R. J. (1994). The reliability of observational data: II. Issues in the identification and measurement of stuttering events. In *Journal of Speech and Hearing Research* (Vol. 37, Issue 2).

Costa, D., & Kroll, R. (2000). Stuttering: An update for physicians. In *CMAJ* (Vol. 162, Issue 13).

Craig, A., Hancock, K., Tran, Y., Craig, M., & Peters, K. (2002). Epidemiology of Stuttering in the Community Across the Entire Life Span. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1097–1105. https://doi.org/10.1044/1092-4388(2002/088)

Croft, R. L., & Byrd, C. T. (2020). Self-compassion and quality of life in adults who stutter. *American Journal of Speech-Language Pathology*, *29*(4). https://doi.org/10.1044/2020_AJSLP-20-00055

Cullinan, W. L., Prather, E. M., & Williams, D. E. (1963). Comparison of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research*, *6*, 187–194. https://doi.org/10.1044/jshr.0602.187

Davidow, J. H. (2021). Reliability and Similarity of the Stuttering Severity Instrument-Fourth Edition and a Global Severity Rating Scale. *Speech, Language and Hearing*, *24*(1). https://doi.org/10.1080/2050571X.2020.1730545

Davidow, J. H., & Scott, K. A. (2017). Intrajudge and interjudge reliability of the stuttering severity instrument–fourth edition. *American Journal of Speech-Language Pathology*, *26*(4), 1105–1119. https://doi.org/10.1044/2017_AJSLP-16-0079

Eadie, T. L., & Doyle, P. C. (2002a). Direct Magnitude Estimation and Interval Scaling of Naturalness and Severity in Tracheoesophageal (TE) Speakers. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1088–1096. https://doi.org/10.1044/1092-4388(2002/087)

Eadie, T. L., & Doyle, P. C. (2002b). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America, 112*(6). https://doi.org/10.1121/1.1518983

Engen, T. (1971). Psychophysics II. Scaling methods. In *Woodworth and Schlossberg's experimental psychology* (pp. 47–86). Holt, Rinehart, & Winston.

Karimi, H., Jones, M., O'Brian, S., & Onslow, M. (2014). Clinician percent syllables stuttered, clinician severity ratings and speaker severity ratings: Are they interchangeable? *International Journal of Language and Communication Disorders*, *49*(3). https://doi.org/10.1111/1460-6984.12069

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) Documentation of Voice Disorders. *Journal of Voice, 21*(5), 576–590. https://doi.org/10.1016/j.jvoice.2006.05.001

Kelchner, L. N., Brehm, S. B., Weinrich, B., Middendorf, J., deAlarcon, A., Levin, L., & Elluru, R. (2010). Perceptual Evaluation of Severe Pediatric Voice Disorders: Rater Reliability Using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice*, *24*(4), 441–449. https://doi.org/10.1016/j.jvoice.2008.09.004

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. *American Journal of Speech-Language Pathology*, *18*(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *Citation: The Journal of the Acoustical Society of America*, *122*, 1867. https://doi.org/10.1121/1.2770547

Lickley, R. J., Hartsuiker, R. J., Corley, M., Russell, M., & Nelson, R. (2005). Judgment of disfluency in people who stutter and people who do not stutter: Results from magnitude estimation. *Language and Speech*, *48*(3). https://doi.org/10.1177/00238309050480030301

Martin, R. R., & Haroldson, S. K. (1992). Stuttering and Speech Naturalness. *Journal of Speech, Language, and Hearing Research*, *35*(3), 521–528. https://doi.org/10.1044/jshr.3503.521

McColl, D., & Fucci, D. (2006). Measurement of speech disfluency through magnitude estimation and interval scaling. *Perceptual and Motor Skills*, *102*(2), 454–460. https://doi.org/10.2466/PMS.102.2.454-460

O'Brian, S., Heard, R., Onslow, M., Packman, A., Lowe, R., & Menzies, R. G. (2020). Clinical trials of adult stuttering treatment: Comparison of percentage syllables stuttered with self-reported stuttering severity as primary outcomes. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1387–1394. https://doi.org/10.1044/2020_JSLHR-19-00142

O'Brian, S., Packman, A., & Onslow, M. (2004). Self-Rating of Stuttering Severity as a Clinical Tool. *American Journal of Speech-Language Pathology*, *13*(3), 219–226. https://doi.org/10.1044/1058-0360(2004/023)

Onslow, M., Jones, M., O'Brian, S., Packman, A., Menzies, R., Lowe, R., Arnott, S., Bridgman, K., de Sonneville, C., & Franken, M. C. (2018). Comparison of percentage of syllables stuttered with parent-reported severity ratings as a primary outcome measure in clinical trials of early stuttering treatment. *Journal of Speech, Language, and Hearing Research*, *61*(4), 811–819. https://doi.org/10.1044/2017_JSLHR-S-16-0448

Prasse, J. E., Ma, C.-S., Hospital, S., Stamford, G. E., & Kikano, C. (2008). *Stuttering: An Overview*. www.aafp.org/afp.

Riley, G. D. (2009). *Stuttering Severity Instrument–Fourth Edition*. Pro-Ed.

Schiavetti, N. (1992). Scaling Procedures for the Measurement of Speech Intelligibility. In R. D. Kent (Ed.), *INTELLIGIBILITY IN SPEECH DISORDERS: THEORY, MEASUREMENT AND MANAGEMENT* (pp. 11–34). The Netherlands: John Benjamins Publishing Company.

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). https://doi.org/10.1044/jshr.2604.568

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations : Uses in Assessing Rater Reliability. In *Psychological Bulletin* (Vol. 86, Issue 2).

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

Teghtsoonian, R., Stevens, S. S., & Stevens, G. (1975). Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects. *The American Journal of Psychology*, *88*(4). https://doi.org/10.2307/1421904

Ward, D., Miller, R., & Nikolaev, A. (2021). Evaluating three stuttering assessments through network analysis, random forests and cluster analysis. *Journal of Fluency Disorders*, *67*. https://doi.org/10.1016/j.jfludis.2020.105823

Whitehill, T. L., Lee, A. S. Y., & Chun, J. C. (2002). Direct Magnitude Estimation and Interval Scaling of Hypernasality. *Journal of Speech, Language, and Hearing Research*, *45*(1), 80–88. https://doi.org/10.1044/1092-4388(2002/006)

Williams, D. E., Wark, M., & Minifie, F. D. (1963). Ratings of stuttering by audio, visual, and audiovisual cues. *Journal of Speech and Hearing Research*, *6*, 91–100. https://doi.org/10.1044/jshr.0601.91

Wingate, M. E. (1964). A Standard Definition of Stuttering. *Journal of Speech and Hearing Disorders*, *29*(4), 484–489.

Yairi, E., & Ambrose, N. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, *38*(2). https://doi.org/10.1016/j.jfludis.2012.11.002

Yaruss, J. S. (1997). Clinical Measurement of Stuttering Behaviors. *Contemporary Issues in Communication Science and Disorders*, *24*(Spring). https://doi.org/10.1044/cicsd_24_s_27

Yaruss, J. S. (1998). Real-Time Analysis of Speech Fluency: Procedures and Reliability Training. *American Journal of Speech-Language Pathology*, *7*(2). https://doi.org/10.1044/1058-0360.0702.25

Yaruss, J. S., & Quesal, R. W. (2004). Stuttering and the International Classification of Functioning, Disability, and Health (ICF): An update. *Journal of Communication Disorders*, *37*(1), 35–52. https://doi.org/10.1016/S0021-9924(03)00052-2

Zraick, R. I., & Liss, J. M. (2000). A Comparison of Equal-Appearing Interval Scaling and Direct Magnitude Estimation of Nasal Voice Quality. In *Journal of Speech, Language, and Hearing*

Research (Vol. 43, Issue 4, pp. 979–988). American Speech-Language-Hearing Association. https://doi.org/10.1044/jslhr.4304.979

Appendix A: a portion of the SSI-4 Friuli Passage used as the reading stimulus

Occupying the extreme northeast corner of Italy, Friuli's scenery ranges from rugged coastline along the eastern border to placid plains in the west and the majestic Alps in the north, where Italy butts up against Austria. Directly to the south is Venice, just a little more than an hour and a half away. Though off the beaten tourist track, Friuli is hard in the path of history. (101 syllables)