Journal of Communication Disorders 74 (2018) 23–34

Contents lists available at ScienceDirect

Journal of Communication Disorders

journal homepage: www.elsevier.com/locate/jcomdis

Effects of measurement method and transcript availability on inexperienced raters' stuttering frequency scores

Nalanda Chakraborty*, Kenneth J. Logan

University of Florida, United States

ARTICLE INFO

Keywords:
Stuttering
Measurement accuracy
Measurement efficiency
Reliability
Clinical training

ABSTRACT

Purpose: To examine the effects of measurement method and transcript availability on the accuracy, reliability, and efficiency of inexperienced raters' stuttering frequency measurements.
Method: 44 adults, all inexperienced at evaluating stuttered speech, underwent 20 min of preliminary training in stuttering measurement and then analyzed a series of sentences, with and without access to transcripts of sentence stimuli, using either a syllable-based analysis (SBA) or an utterance-based analysis (UBA). Participants' analyses were compared between groups and to a composite analysis from two experienced evaluators.
Results: Stuttering frequency scores from the SBA and UBA groups differed significantly from the experienced evaluators' scores; however, UBA scores were significantly closer to the experienced evaluators' scores and were completed significantly faster than the SBA scores. Transcript availability facilitated scoring accuracy and efficiency in both groups. The internal reliability of stuttering frequency scores was acceptable for the SBA and UBA groups; however, the SBA group demonstrated only modest point-by-point agreement with ratings from the experienced evaluators.
Conclusions: Given its accuracy and efficiency advantages over syllable-based analysis, utterance-based fluency analysis appears to be an appropriate context for introducing stuttering frequency measurement to raters who have limited experience in stuttering measurement. To address accuracy gaps between experienced and inexperienced raters, however, use of either analysis must be supplemented with training activities that expose inexperienced raters to the decision-making processes used by experienced raters when identifying stuttered syllables.

1. Introduction

Speech-language pathologists commonly measure stuttering frequency as a means of describing the fluency of people who stutter. With this measure, the clinician's main goal is to quantify how often observable instances of stuttering-related behavior occur in a speech sample. Measurement of stuttering frequency involves data collection, data analysis, and data interpretation, and it is essential for clinicians who are new to working with people who stutter to develop competency in each of these areas. The present study focuses on the second of these areas – data analysis – and, more specifically, the effects that analysis method and transcript availability have on the accuracy, reliability, and efficiency of stuttering frequency measurements that inexperienced raters make.

In contemporary clinical practice, stuttering identification is based largely on perceptual judgment of acoustic information. That is, a clinician must determine whether a specific segment of the speech signal contains behavior that sounds as if it is symptomatic of stuttered speech (Bloodstein & Bernstein Ratner, 2008; Smith & Kelly, 1997). Beyond its contribution to the diagnosis of stuttering,

* Corresponding author.
E-mail address: nalanda.c@ufl.edu (N. Chakraborty).

https://doi.org/10.1016/j.jcomdis.2018.04.006
Received 6 October 2017; Received in revised form 7 April 2018; Accepted 22 April 2018
Available online 24 April 2018
0021-9924/

stuttering frequency analysis provides a sense for the severity of a speaker's fluency impairment as well as the extent to which a speaker's fluency functioning varies across daily situations (e.g., Gillam, Logan, & Pearson, 2009; Logan, Byrd, Mazzocchi, & Gillam, 2011; Prosek, Walden, Montgomery, & Schwartz, 1979; Riley, 1994, 2009).

Stuttering judgment seems to be a multi-dimensional process, in that clinicians can consider multiple factors when deciding whether or not a particular segment of speech is symptomatic of stuttering (Bloodstein & Bernstein Ratner, 2008; Logan, 2015). Examples of speech-based factors that may enter into the identification of stuttered speech include the following: (a) the type of disfluency that has occurred; (b) disfluency duration; (c) speech prosody (e.g., sound segment duration, stress and rhythm patterns); (d) speaking effort, and (e) speech naturalness. When visual information about a speaker is available, other aspects of behavior, such as extraneous body movements, may contribute to the identification of stuttered speech, as well.

When quantifying stuttering frequency, clinicians can choose from among several data analysis methods (cf., Cordes, Ingham, Frank, & Ingham, 1992; Riley, 2009; Ryan & Ryan, 1995; Williams, Darley, & Spreistersbach, 1978). The traditional approach involves elicitation of several hundred consecutive words or syllables during conversation, narration, or oral reading (e.g., Riley, 2009; Williams et al., 1978). Linguistic units in the speech sample – usually words or syllables – then are coded for the presence or absence of stuttering-related behavior. Data analysis can take place either while the client is speaking (i.e., "live" or "real-time" analysis) or some time after the speech sample has been elicited (i.e., "off-line" analysis) by replaying a recording of the client's speech (Logan, 2009; Yaruss, Max, Newman, & Campbell, 1998). Clinicians also may elect to develop written transcripts of what the client has said. Transcription can be done either live or through replay of an audio recording. Either way, the transcript results in a permanent record of the client's performance, and when used in combination with an audio recording, it provides the opportunity to review the content of client's speech as well as the ability to document speech sample components such as the number and nature of syllables or words spoken, grammatical aspects of the spoken utterance, and the number of syllables that feature stuttering-related behavior.

An alternative to syllable- and word-based stuttering analyses is to analyze longer linguistic units such as entire utterances, entire speaking turns, or multi-word spans within utterances that are located in utterance positions that are prone to containing stuttered speech. Although the latter approach has not been used as often as the syllable- or word-based approach, it nonetheless has appeared in both disfluency-related research studies (e.g., Bernstein Ratner & Sih, 1987; Gaines, Runyan, & Meyers, 1991; Gordon & Luper, 1989; MacLachlan & Chapman, 1988) and in assessment tools such as the *Test of Childhood Stuttering* (Gillam et al., 2009) and the *Stocker Probe* (Stocker & Goldfarb, 1995). With such approaches, linguistic units larger than a word are coded based on the presence or absence of behaviors that a clinician perceives to be related to stuttering.

Because utterances, phrases, and speaking turns typically consist of more than one word, these analyses usually entail less coding than either word- or syllable-based analyses. As with syllable-based analyses, analyses that focus on linguistic units longer than a word have proven capable of reliably differentiating speakers who stutter from typical speakers and can yield estimates of the severity of fluency impairment (Bernstein Ratner & Sih, 1987; Gaines et al., 1991; Gillam et al., 2009). Thus, both types of measures warrant investigation in studies of the quality with which stuttering frequency analyses are performed.

To date, much has been written about the types of analyses that are available to clinicians for documenting stuttering frequency. However, there is less information available on factors that affect the implementation of these analyses, particularly with respect to the accuracy, reliability, and efficiency of their use. Most of the research that deals with this aspect of fluency assessment has related to the intra- and inter-judge reliability of stuttering judgments. Such research has led to the conclusion that overall stuttering frequency score agreement is higher than point-by-point agreement (i.e., whether specific syllables are stuttered or not) in intra-judge, inter-judge, and inter-clinic contexts (Cordes & Ingham, 1994; Kully & Boberg, 1988; Lewis, 1994; Yaruss et al., 1998), and, not unexpectedly, that inexperienced raters tend to be less reliable in their stuttering measurements than experienced raters are, with much of the difference between the two rater groups apparently resulting from the tendency for inexperienced raters to under-identify instances of stuttering-related behavior (Brundage, Bothe, Lengeling, & Evans, 2006; Young, 1975). The tendency toward under-identification of stuttered syllables has been attributed to factors such as unfamiliarity with a speaker's stuttering patterns and rater reluctance to wrongly label non-stuttered disfluencies as stuttering (Brundage et al., 2006), as well as to observers' momentary lapses in attention, the frequency and speed with which stuttering behaviors occur in a speech sample, and challenges associated with detecting stutters that span word boundaries (Young, 1975). Although some studies have reported the overall inter-judge and intra-judge agreement for stuttering measurements by inexperienced raters to be as high as 90% and similar to that of experienced raters (Brundage et al., 2006; Cordes & Ingham, 1994; Ingham, Cordes, & Gow, 1993), other studies have reported much lower point-by-point agreement by inexperienced raters, and in some cases, the mean agreement has been as low as 50% (see Costello & Ingham, 1984; Young, 1975), with point-by-point agreement being particularly low when stuttering instances occur both between and within words (Curlee, 1981).

Given the issues that have been identified with the accuracy and reliability of stuttering-related measurements, it is not surprising that some researchers have examined factors associated with the training of stuttering measurement skills. Overall, it has been demonstrated that the reliability of stuttering measurements can be improved when evaluators (i.e., "raters") receive training in identifying stuttering (Cordes & Ingham, 1994; Costello & Hurst, 1981; Ingham et al., 1993). Cordes and Ingham (1994) examined the effects of various training stimuli on inexperienced raters' ability to perform time-interval-based measurements of stuttering. They concluded that use of training exemplars that were highly agreed upon by experts as being stuttered were more effective than training exemplars that were poorly agreed upon by experts. Yaruss et al. (1998) compared disfluency analyses of inexperienced student raters under two-conditions: transcript-based and real-time. Their results showed that the transcript-based and real-time conditions yielded similar frequency scores for both "more typical" (i.e. revisions, interjections, hesitations, repetition of phrases) and "less typical" (e.g., repetitions of sounds or syllables, prolongations, and blocks) types of speech disfluency; however, similar to findings in other