Predicting Stuttering Severity Ratings by Timing and Tallying Dysfluencies Using Praat Software

Elizabeth S. Hasseltine
Shannon F. Black
Tayler M. Corcoran
Danika L. DiPalma
Susan E. Dixon
Anne T. Gooch
Lauren M. Hurlburt
Ashton B. Murray
Kathryn B. Potts
Anna C. Schnizler
Caitlin Secrist
Roma Marisa Shickel
Filip Loncke
University of Virginia, Charlottesville

Paul Corthals
University College Ghent, Belgium

ABSTRACT: Purpose: The goal of this study was to examine the relationship between objective descriptors of stuttering behavior and perceptions of stuttering severity. Classification systems for speech dysfluencies are numerous; this study sought to find a less complicated, yet accurate, predictor of stuttering severity.

Method: This study used a taxonomy of stutters outlined by Teesson, Packman, and Onslow (2003). Using this taxonomy, recorded speech samples were annotated using Praat software, and the type of stuttering symptom and its duration was determined. The power of dysfluency types and duration as predictors for stuttering severity was examined by means of a regression analysis. Raters evaluated 1-min speech samples for perceived stuttering severity.

Results: Timing parameters yielded more significant severity predictors than tallying parameters did. A concise equation for predicting stuttering severity was established that accounted for the duration of fluent speech in conversation.

Conclusion: Although the timing parameter predictor of objective stuttering severity is only 1 component of the assessment of a person who stutters, this simplified focus on percentage of time fluent will compensate for the lack of consensus regarding stuttering classification systems. These quantitative data can be used in addition to an assessment of a person's stuttering experience to better understand and treat fluency disorders holistically (Yaruss & Quesal, 2006).

KEY WORDS: fluency, stuttering
For many years, practitioners and researchers have disagreed on the most effective methods of classifying speech dysfluencies because stuttering is so variable across settings and situations (Yaruss, 1997). Following a comprehensive review of the literature, Yaruss (1997) stated that the different approaches for classifying these dysfluencies are so vast that finding a cohesive and comprehensive classification system is extremely difficult.

Yaruss (1997) outlined the many classification systems for stuttering that have been used and modified during the last several years by researchers and clinicians. Johnson (1961) developed one of the early classification systems, focusing on interjections, repetitions, revisions, broken words, incomplete phrases, and prolonged sounds. Since then, a multitude of scholars have worked to categorize dysfluencies through different classification systems either to become more specific or to focus on different aspects of fluency (Yaruss, 1997). Although many of these classification systems have striking similarities, the approaches differ in the nuances of their groupings. Meyers (1986), for example, developed a system to differentiate between *stutter-type dysfluencies* such as prolongations, tense pauses, and broken words and *normal-type dysfluencies*, which include revisions, interjections, and incomplete phrases. Similar to Meyer’s classification, the “stuttering-like dysfluency” grouping differentiated between stuttering instances like word repetition versus interjections or phrase revisions (Yairi, 1996; Yaruss, 1997). More recently, Einarsdottir and Ingham (2005) demonstrated these complexities by outlining more than 20 different possible labels that a particular speech dysfluency could have. Some of the terms used in this labeling process are practically synonymous, however, conveying different but only vaguely defined nuances (e.g., “tense pause” and “blocks,” or “broken words” and “part-word repetition”). Moreover, it is not entirely clear what domain(s) descriptors of stuttering should cover. Whereas some domains refer to units of meaning (e.g., interjections), other domains refer to units of syntax (e.g., phrase repetitions) or other categories (e.g., sound prolongation), creating even more confusion and controversy during the labeling process.

Ways of counting the instances of stuttering also vary greatly from clinic to clinic. Clinicians differ in the criteria they use for classifying and categorizing stuttering and stuttering behaviors (Yaruss, 1997). To contrast the popular tallying method of instances of dysfluencies as noted by the aforementioned yet far from exhaustive classification systems above, researchers Cordes and Ingham (1994) turned their attention to timing stuttering intervals (Yaruss, 1997). Cordes and Ingham further illustrated variation of labeling by highlighting the great inconsistency in counting stuttering events. Although they found decent interrater reliability in timing methods, they did not see how this knowledge could be applied in a clinical setting. In other words, these studies emphasize that stuttering is determined by a multitude of factors that may escape descriptive classifications.
Most methods to assess stuttering rely on listener judgments. For example, the Stuttering Severity Instrument (Riley, 1972) is a norm-referenced tool that measures stuttering severity in four areas of speech behavior: frequency, duration, physical concomitants, and naturalness of the individual’s speech. Various types of stuttering behaviors within a speech sample are judged as well as global severity ratings. Preferably, the assessments combine to form a consensus score that is based on individual ratings by several judges. This approach implies the use of an agreed-on dysfluency typology, whereas the literature offers several taxonomies of stuttering-like dysfluencies. Additionally, interjudge and intrajudge agreement levels vary significantly when judges have not received training in identifying the behaviors corresponding to the stuttering-like behaviors (Cordes, 2000). Collapsing dysfluency descriptors into a limited number of self-defining categories would likely enhance inter- and intrajudge agreement (Teesson et al., 2003). Quantifying the proportion of nonstuttered fragments may also help to capture what observers are able to agree on.

The purpose of the current study was to develop a dysfluency assessment protocol and juxtapose it to subjective listener perceptions of stuttering severity. Specifically, we examined the feasibility of using various stuttering symptoms extracted from short, recorded speech samples as predictors of perceived stuttering severity and then developed an equation to predict perceived severity. Using the taxonomy developed by Teesson et al. (2003), speech samples were annotated with Praat software (Boersma & Weenink, 2013), marking both the duration (timing) and the type (tallying) of each dysfluency. In order to obtain ratings for stuttering severity, we presented the samples to a panel of listeners and drew correlations between the average listener ratings of an individual and the duration, number, and type of dysfluent behaviors in the different samples. We hypothesized that listeners’ perceived stuttering severity can be predicted in a valid way using a simple and straightforward taxonomy by taking into account the duration and number of stuttering instances and fluent fragments. We also hypothesized that timed measures would be more important than tallied measures in predicting a listener’s perception of stuttering severity.

**METHOD**

**Participants**

The participants included 120 undergraduate students and family members who were selected by convenience sampling. The raters ranged in age from 14 to 62 years, and all were English speakers. None of the participants had any professional training in speech and hearing sciences.

**Materials**

We used 36 speech samples from the University College London’s Archive of Stuttered Speech (UCLASS; University College London, 2004). These speech samples are available to the public. All of the samples were in English. To reduce the total number of speech samples each rater was tasked with listening to and rating, we divided the 36 speech samples into three subgroups of 12. Per the listing order on the UCLASS database, samples listed 1–12 constituted subgroup 1, samples listed 13–24 constituted subgroup 2, and samples listed 25–36 constituted subgroup 3. The raters were randomly split into one of three groups. The 40 raters in group one listened to the same set of 12 speech samples, the 40 raters in group two listened to a different set of 12 speech samples, and the 40 raters in group three listened to the remaining set of 12 speech samples. Therefore, each of the 36 speech samples in this study was rated by 40 people.

**Design**

This study presents a correlational research design. It seeks to find the relationship between the timing and tallying of various stuttering dysfluencies in the speech of a person who stutters and a rater’s perceived stuttering severity level of their stuttering behavior. For the purposes of objective analysis in this study, we used the Teesson et al. (2003) dysfluency taxonomy to develop a system of mnemonics for labeling observable speech motor events. This taxonomy, shown in Figure 1, is divided into three main categories: repeated movements, fixed postures, and superfluous behaviors.

The first category, *repeated movements*, includes three subcategories: syllable repetition, incomplete syllable repetition, and multisyllable repetition. *Syllable repetition* can be defined as the repetition of a single syllable (e.g., the word “on” in the phrase “on-on-on the bus,” or a single-syllable part of a word such as in “un-un-under the table”). *Incomplete syllable repetition* occurs when one speech sound (less than a syllable) reoccurs multiple times in a row. An example of this is the fricative /s/ reoccurring multiple times at the beginning of the production of the word “sun” (e.g., “s-s-s-s-s-sun”). *Multisyllable repetition* occurs when a set of more than one syllable or a series of words recurs, as in the recurrence of the first two syllables in the production of

**Figure 1**
the word “together” (e.g., “toge-toge- together”) or in the repetition of the first two words in a phrase or sentence (e.g., “And then, and then, and then we went to the mall.”).

The second category of the outlined taxonomy is fixed postures. The first subcategory here is prolongation. Prolongation occurs when a continuous speech sound (e.g., a vowel or a fricative) is produced for a longer duration than normal (with a “normal” occurrence defined as lasting approximately one tenth of a second) and the sound includes audible airflow. The other subcategory of fixed postures is called a block. In this case, the person speaking struggles to produce a sound and is unable to do so. As a result, this type of fixed posture does not include audible airflow in the way that a prolongation does. This tense stop occurs unnaturally and in the middle of words, not simply during the natural pauses in between words.

The last category of the taxonomy for speech dysfluencies is classified as superfluous behaviors. This category includes both verbal and nonverbal behaviors. Verbal superfluous behaviors include interjections and revisions during speech. For example, in the phrase, “This, um, is, um a dog,” both instances of the interjection “um” would be labeled as interjections because they are unnecessary and unnatural in the flow of speech. A revision, on the other hand, would be a case in which the speaker corrects him- or herself before continuing on with the sentence or train of thought, such as the revision from “this” to “that” in “This, that is a dog.” A nonverbal superfluous behavior would include visible and/or audible movements such as twitching, squirming, lip smacking, throat clearing, and so on. It is important to note that in the present study, only audible nonverbal superfluous behaviors could be detected via recorded speech samples, and therefore, nonaudible, nonverbal superfluous behaviors were not included in the analysis.

**Procedure**

Each of the 36 samples was listened to and rated by 40 people. To avoid possible order-effect biases, raters in each group were exposed to their 12 samples in a different sequence. The raters were asked to listen to approximately the first minute of each sample and to rate each sample on a Likert scale (1 = mild
to 5 = severe) for the severity of stuttering. To calibrate the responses, the raters listened to the same sample of a “moderate” stutter (as determined unanimously by the researchers) before beginning their rating session. The distribution of mean severity ratings on the Likert scale showed that all degrees between 1 and 5 were represented by the speech samples.

We annotated the same first minute of each of the 36 samples that the raters listened to using Praat software (Boersma & Weenink, 2013). The annotations consisted of short mnemonics on a text grid corresponding to the oscillogram of the sound sample. Figure 2 outlines some of the mnemonic codes that represent the taxonomy used in the analysis. The mnemonics were Rs (repeated syllables), Ri (incomplete syllable repetitions), Rm (multisyllabic repetitions), Fp (audible fixed postures, also known as prolongation), Fb (inaudible fixed postures, also known as blocks), and the three types of superfluous behaviors—Sr (revisions), Sn (nonverbal behaviors), and Si (interjections). In addition to the mnemonic codes, the number 1 was used to indicate stretches of fluent speech, and the number 0 was used to indicate silence not classified as a dysfluent behavior. Annotations in Praat software also delineate the duration of each item on the text grid. In order to obtain maximum intersubject reliability, we first independently annotated samples and then discussed any discrepancies.

We used a Praat script developed by one of the authors to analyze the frequency of occurrence of each subtype of dysfluency from the text grid annotations, as well as the mean duration of each subtype and its relative duration (as a percentage of the sample duration). SPSS version 19 (IBM, 2010) was used to calculate correlations and regression equations.

The strength of the relationship between the mean severity rating and timed or tallied parameters of every dysfluency type was analyzed with Pearson product–moment correlation coefficients. Because several significant relationships existed, we calculated a multiple regression equation to determine the relative importance of each dysfluency type as a severity predictor. Finally, we calculated a one-predictor regression equation to predict perceived stuttering severity starting from an annotated 1-min recording.

**RESULTS**

**Fluent Fragments Parameters**

Regarding parameters pertaining to fluent fragments of speech, classified as stretches of speech void of all stuttering instances, the analysis revealed a significant negative correlation for the two timing behaviors: mean duration of fluent fragments ($r = –.734, p <$
.01) and percentage of time of fluent fragments \( (r = -.847, p < .01) \). This means that the tallied number of fluent stretches \( (r = -.126, p > .05) \) in the recordings did not correlate significantly with the listeners' perceptions.

**Repeated Movement Parameters**

The analysis of repeated movement parameters showed many significant correlations between the timing and tallying behaviors. Specifically, syllable repetitions had a significant correlation between perceived severity of stuttering and both timing measurements: mean duration \( (r = .616, p < .01) \) and percentage of time \( (r = .492, p < .05) \). Incomplete syllable repetitions showed significant correlation for all three measurements: number of instances \( (r = .440, p < .05) \), mean duration \( (r = .519, p < .01) \), and percentage of time \( (r = .548, p < .01) \). Likewise, all subtypes combined also showed significant correlations for all three measurements: number of instances \( (r = .743, p < .05) \), mean duration \( (r = .654, p < .01) \), and percentage of time \( (r = .904, p < .05) \). However, there were no significant correlations for any of the multisyllable repetition parameters. These results show that both timing behaviors—mean duration and percentage of time of syllable repetitions—had a significant positive correlation with listeners' perceived stuttering severity for two of the three parameters, whereas number of instances was only correlated with one parameter. Therefore, both timing and tallying behaviors for incomplete syllables appear to be important in influencing a rater's perceived stuttering severity; of these two, the timing behaviors showed more significant results.

**Fixed Posture Parameters**

Next, for parameters pertaining to fixed postures (blocks), the results showed a significant positive relationship between perceived stuttering severity and both the number of blocks \( (r = .365, p < .05) \) and the percentage of time of blocks \( (r = .497, p < .01) \). This means that as the number and length of fixed postures in the speech sample increased, the rater was more likely to perceive a higher stuttering severity. This finding suggests that both the timing and tallying of this dysfluency behavior have a significant relationship in determining a rater's perceived severity of stuttering if fixed postures are present in the speech sample.

**Superfluous Behavior Parameters**

For parameters pertaining to superfluous behaviors, the number of revisions (tallying behavior) showed a significant negative correlation \( (r = -.581, p < .05) \). This seems to suggest that as the number of instances that a speaker makes revisions in a speech sample increases, the more likely the rater is to perceive a lower stuttering severity. In addition, parameters pertaining to superfluous behaviors showed a significant positive correlation between the mean duration of nonverbal superfluous behaviors \( (r = .368, p < .05) \) and raters' perceived stuttering severity. Interestingly, the strongest correlation occurred when all of the subtypes of superfluous behaviors were combined, yielding a significant positive correlation \( (r = .582, p < .01) \).

The results of the stepwise multiple regression analysis to determine relative predictor importance revealed that the top five predictors from most important to least important were (a) percentage of time fluent, (b) mean duration of fluent fragments, (c) percentage of time of incomplete syllable repetitions, (d) percentage of time blocks (fixed postures), and (e) mean duration of incomplete syllable repetitions. Because the strongest correlation occurred between the percentage of time speech was fluent and listeners' perceived stuttering severity \( (r = -.847, p < .01) \), we used this measurement to create an equation to predict stuttering severity perception. This linear regression of percentage of time fluent is shown in Figure 3. The equation we developed to help raters perceive stuttering severity is: Rating = 5.4 – (0.048 × time %Fluent). This equation can explain a relatively high amount of variance (\( R^2 = 71.8 \)), meaning that the equation predicts listeners' severity ratings with considerable accuracy.

**Figure 3.** Mean rating as a function of percentage of time speaking fluently.
**DISCUSSION**

As hypothesized, the timing behaviors yielded more significant correlates and predictors with more weight for predicting listeners’ perceived stuttering severity than the tallying behaviors did. Of all of the significant correlates of perceived severity, the strongest were those pertaining to the amount of time that speech was fluent (speech void of all instances of stuttering) and those pertaining to the amount and duration of repetitions, as seen in Table 1. The data show a −.847 correlation for the relative duration of fluent fragments and a −.904 correlation for the relative duration of all types of repetitions combined. These data reflect that overall, the timing of behaviors produces more and greater significant correlations than the tallying of data.

One might argue that the categories are divided unfairly because timing behaviors include two categories (i.e., mean duration of speech behavior and percent time of dysfluency behavior) over the entire utterance, and this is in contrast to the single approach to tallying. Even in breaking up these parameters, however, the correlation of one approach to timing is still more significant than tallying. Tallying behaviors may seem the easiest way to quantify stuttering severity; however, when using text grid annotations and a preprogrammed script to perform the calculations, timing behaviors can also be done in a simple and efficient way and may actually yield results that are more consistent with listeners’ perceived stuttering severity.

We annotated the first minute of each recording exhaustively, using mnemonics chosen from the full taxonomy as well as the codes 1 (fluent fragment) and 0 (silence not classified as dysfluency). The output of the Praat script offers a survey of all annotated behaviors (timed and tallied). However, the regression to predict the listeners’ perceived stuttering severity only takes one type of behavior into account, namely, percentage time of fluent speech. Therefore, if the only goal is to assess severity, one only needs to mark fluent fragments (using code 1 exclusively, without marking the other Lidcombe Behavioral Data measures) and express their cumulative duration as a percentage of the total duration of the sample being analyzed (i.e., 1 min). Used as a predictor variable in the regression equation, this percentage yields a prediction of perceived severity with considerable accuracy. This outcome of the regression equation is a prediction of the average of all individual severity scores that a group of naive listeners would have given. This means that this formula has face validity as an outcome measure. To monitor the effects of treatment eliminating bias as much as possible, quantitative and verifiable indices are needed because fluency remains a critical goal of stuttering treatment.

Annotations of dysfluencies are not necessary to use the regression equation, but the annotation tier containing the marked fluent fragments can be saved and completed later on by adding mnemonics to signal particular dysfluencies. Once completed, the breakdown of observed phenomena (as seen in Table 1) will offer a profile of the person who stutters, highlighting the most prominent symptoms.

<table>
<thead>
<tr>
<th>Dysfluency subtype</th>
<th>Number of instances</th>
<th>Mean duration</th>
<th>Percentage of time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluent fragments</td>
<td>−.126</td>
<td>−.734**</td>
<td>−.847**</td>
</tr>
<tr>
<td>Fixed postures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audible (prolongations)</td>
<td>.064</td>
<td>.375</td>
<td>.162</td>
</tr>
<tr>
<td>Inaudible (blocks)</td>
<td>.365*</td>
<td>.277</td>
<td>.497**</td>
</tr>
<tr>
<td>Repeated movements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syllable repetitions</td>
<td>.062</td>
<td>.616**</td>
<td>.492*</td>
</tr>
<tr>
<td>Incomplete syllable repetitions</td>
<td>.440*</td>
<td>.519**</td>
<td>.548**</td>
</tr>
<tr>
<td>Multisyllable repetitions</td>
<td>.138</td>
<td>.377</td>
<td>.410</td>
</tr>
<tr>
<td>All subtypes combined</td>
<td>.743*</td>
<td>.654*</td>
<td>.904*</td>
</tr>
<tr>
<td>Superfluous behaviors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interjections</td>
<td>−.106</td>
<td>−.164</td>
<td>−.142</td>
</tr>
<tr>
<td>Revisions</td>
<td>−.581</td>
<td>.150</td>
<td>−.397</td>
</tr>
<tr>
<td>Nonverbal</td>
<td>.027</td>
<td>.368*</td>
<td>.304</td>
</tr>
<tr>
<td>All subtypes combined</td>
<td>.046</td>
<td>.349</td>
<td>.582*</td>
</tr>
</tbody>
</table>

*Correlation significant at the 0.05 level, two-tailed. **Correlation significant at the 0.01 level, two-tailed.
Limitations

One major limitation of our study is the length of the samples used. Due to the limits of using 1-min speech samples, only a reasonably small number of phrases and utterances could be formulated in each sample. Although significance was found for the tallying of superfluous revisions, we believe that the raters’ perceptions might have correlated with more types of superfluous behaviors if longer speech samples were used. For this reason, we believe that future research could benefit from giving some subgroups of raters the 1-min clips while others listen to longer versions of the same samples in order to test if the sample length influences perception as well.

Despite this limitation, we feel that the data were more than sufficient to show the relative contribution of each predictor to the prediction of perceived stuttering severity. A look at the top five predictors in the study shows that the greatest contribution to perceived severity came from timed behaviors as all five of the top predictors were from timed events. Given the importance of timing behaviors, the data again make an argument in favor of the need for a tool such as Praat that can make accurate timings for diagnostic and therapeutic purposes. Because timing behaviors seem to hold the most weight in affecting listeners’ ratings of stuttering severity, analyzing samples using the protocol in this study could help therapists set goals more accurately. Furthermore, the amount of variance explained ($R^2 = 71.8\%$) shows that just one behavioral timing measure, percentage time of fluent speech, is needed in order to consistently predict listeners’ severity ratings.

The relative duration of all types of repetitions combined also correlates very strongly with average severity ratings, but the benefit to using the percentage time of fluent speech in the equation is that the evaluator does not need a vast and detailed understanding of the different taxonomies of stuttering in order to use this method. The only observational task of the listener is to be able to recognize and time fluent speech, which is a simple parameter with face validity. Because of its simplicity, the equation based on the percentage time of fluent speech could likely be used as a screening tool to determine whether an individual should be further evaluated for a fluency disorder. Therefore, if trained in how to time for fluency, users with little formal knowledge of stuttering could use the formula to better understand the stuttering severity of a person who stutters.

Obtaining severity indices pertaining to speech fluency is not enough to completely describe the multifaceted nature of stuttering, although it does help. The World Health Organization’s International Classification of Functioning, Disability and Health (ICF) provides a more comprehensive framework, including functioning, disability, and contextual factors (Yaruss & Quesal, 2004). The ICF does not currently mandate that a particular classification system be used when judging the severity of stuttering behaviors. However, a rating of fluent speech fits in the ICF model as it quantifies an important component of a person’s functional status, that is, communication.

The ICF model looks to describe the experience of a person who stutters both with strengths and weaknesses, so understanding their patterns of fluent speech will help clinicians to focus on both functioning and impairment (Yaruss & Quesal, 2004). In addition to using stretches of fluent speech to classify the stuttering severity of a person who stutters, it is imperative to keep in mind the perspective of a person who stutters because this ultimately influences the overall judgment of stuttering severity. How severe a person perceives his or her own stuttering behaviors to be as well as the reactions they receive from others in various environments regarding communication certainly plays a role in both their confidence and their attitude toward social situations requiring speaking (Yaruss & Quesal, 2006).

Conclusion

To revisit the research question of whether there is a relationship between perceived severities of stuttering and timed or tallied behaviors, our study has shown that timing and tallying stuttering behaviors by annotating relatively short stuttered audio speech samples yields significant correlates of listeners’ perceived stuttering severity. This corroborates the validity of the protocol that was used in this study. The hypothesis that timing behaviors yield predictors with more weight in perceived severity than tallying behaviors yield was supported by the results of this study. An equation for predicting perceived stuttering severity was then established and tested for validity.

From the study of mean severity rating of nonspecialized participants and the analysis of the taxonomy patterns of each sample, we showed that the percentage of time fluent factor can be used as a single predictor of perceived severity through our equation to systematically assess the severity of stuttering. This predictor of objective severity is only one component of stuttering severity. However, this simplified focus on the percentage of time fluent can compensate for the lack of consensus regarding stuttering classification systems. This piece of quantitative data can be used in addition to an assessment of
a person’s experience of stuttering to better understand and treat fluency disorders in a holistic manner (Yaruss & Quesal, 2006).

REFERENCES


Contact author: Elizabeth Hasseltine, 417 Emmett Street South, Charlottesville, VA 22904. E-mail: esh7rk@virginia.edu