Procedures used for assessment of stuttering frequency and stuttering duration

LEANNE JANI, MARK HUCKVALE, & PETER HOWELL

Department of Cognitive, Perceptual and Brain Sciences, University College London, London, UK

(Received 4 May 2013; revised 24 May 2013; accepted 27 May 2013)

Abstract

Frequency of stuttered syllables and their durations were assessed using different procedures. The experiment examined overall syllable counts, counts of stuttered syllables and measures of stutter durations when they were made simultaneously or successively. Samples of speech with associated syllable, stuttered syllable and duration measurements of stuttering events were employed in reference transcriptions. Samples contained a minimum of 200 syllables. Ten participants assessed these samples for syllables, stuttered syllables and duration in an experiment. The responses of these participants were stored in alignment with the speech recordings for analysis. Performance was significantly more accurate (relative to transcriptions) for measures other than duration when the successive procedure was used as opposed to the simultaneous procedure. Although the successive method was more accurate, accuracy of stutter event identification was low for most participants. The procedure that allowed listeners to replay a speech sample and count the syllables, stuttered syllables and durations in three passes yielded more accurate syllable and stuttered syllable counts than procedures that required these judgments to be made in one pass.

Keywords: Developmental stuttering, stuttering duration, stuttering frequency, stuttering severity

Introduction

Most stuttering assessments involve judgments made whilst attending to multiple aspects of speech simultaneously. However, in many cases, the procedures for making such judgments have not been evaluated for reliability and validity. This study evaluated computer-based instruments for the assessment of stuttered speech. The guidance for use of these instruments often permits speech information to be collected in different ways. The question that arises is whether different ways of making real-time judgments (simultaneously or successively) with the same instrument affect the results (Howell, Soukup-Ascencao, Davis, & Rusbridge, 2011). If so, caution would need to be exercised when results using different procedures are compared. Some assessment procedures made with computer and electronic instruments are reviewed in the next section. The computer- and electronic-based instruments themselves are described in the section that follows. An experiment that evaluates whether there are differences between two commonly...
used assessment procedures, making such judgments simultaneously or successively, is then reported.

Assessment of stuttering

This study concerned procedures that count stuttering symptoms in discrete syllables, so procedures such as time-interval analysis that are not syllable-based are not considered in detail (Cordes, Ingham, Frank, & Ingham, 1992). The Stuttering Severity Instrument versions 3 and 4 (SSI-3 and SSI-4) (Pro-ed, Austin, TX) provide the most widely used syllable-based procedure for assessing stuttering symptoms (Riley, 1994, 2009). SSI-3 and SSI-4 combine percentage of stuttered syllables (%SS) and average duration of the three longest stuttering symptoms, with indications about physical concomitants to stuttering observed at the time of symptom assessment, to give an overall severity score (Riley, 1994, 2009). Riley (1994) indicated that SSI can be used: (1) as part of diagnostic evaluation; (2) for tracking changes in severity during and after treatment (Cook, Donlan, & Howell, 2013; Miller & Guitar, 2009); (3) to describe the severity distribution in experimental groups that include people who stutter (Howell, Davis, & Williams, 2008) and (4) to validate other stuttering measures (Howell, Davis, & Williams, 2009). In addition to the uses Riley (1994) identified, it has also been reported that SSI-3 can be used to estimate the statistical risk whether an 8-year-old child who stutters will persist in or recover from their stutter by teenage (Howell & Davis, 2011), and it can be used to distinguish groups of children who stutter from their fluent peers (Howell, 2012). It should be noted that no version of SSI has a category of ‘‘Normal fluency’’ or ‘‘Not-stuttering’’ and there is no ‘‘non-stutterer’’ reference group in the norms (Howell & Davis, 2011), so the assessments themselves do not provide a diagnosis. However, SSI-3 scores can be obtained for fluent children (Davis, Shisca, & Howell, 2007), and Howell (2013) estimated that an SSI-3 score of 13, made according to specific procedures, reliably separated fluent children from children who stutter. One reason for the popularity of SSI-3 is that it was the only norm-referenced severity tool available for many years. Another reason is that it permitted variations in procedures so it can be used in such diverse settings as clinics and research laboratories (Howell et al., 2011). The basic principles remain the same in the permitted variants of the procedure. Riley (2009) argued that the norms that were obtained for SSI-3 also apply to SSI-4 (the SSI-3 norms are used in SSI-4). At present, it is not known whether the different procedures lead to different SSI-3 scores.

Disfluency-based analyses (as exemplified in Riley’s work) are one way of making judgments, and other groups make assessments in other ways. For example, Cordes et al. (1992) had people judge whether fixed length extracts of speech are or are not stuttered. This allows other speech-related information (beyond disfluency type) in the material that is judged to be taken into account. This is also a feature of Yaruss, Max, Newman, and Campbell’s (1998) work, which developed assessment procedures to obtain frequency and duration measures of stuttering and fluency. They compared the accuracy of transcript-based and live procedures using the Systematic Disfluency Analysis (SDA) technique (Campbell & Hill, 1987). SDA separates the assessment of stuttering from assessment of disfluency. Orthographic transcriptions were made of the audible and physical aspects of speech disturbances. Clinicians also tallied the number of syllables as well as the number and type of both stuttered and non-stuttered disfluency while they listened to a sample of speech. The transcription technique was more sensitive than the live procedure at identifying complex stuttering events but the overall stutter frequency measures were similar for both techniques. Consequently, it was suggested that the transcription technique should be used during diagnostic assessment where fine detail is needed and the live technique should be employed to measure change during treatment (Yaruss et al., 1998).
The technique usually used to assess the frequency of stuttered syllables in connection with the Lidcombe program requires a clinician to count syllables and stuttered syllables using the TrueTalk electronic device (Lincoln & Harrison, 1999), which is described in the following section. The Lidcombe procedure allows the clinician to assess a child’s speech while he or she is present (live), instead of from a recording. It would be less problematic to make the assessments from a recorded sample than live to avoid placing high demand on clinicians when they are dealing with a client who stutters (Bakker, Brutten, & McQuain, 1995). The clinician would be free to concentrate on ancillary behaviors, such as physical concomitants, if the assessments were made later. At present, no studies have compared procedures where the speech is assessed live (Lincoln & Harrison, 1999) versus those made from a pre-recorded sample even though these would appear likely to give different results (Howell et al., 2011; Yaruss et al., 1998).

Instrumentation for real-time judgments

Yaruss (1999) developed a computer program to count the number of fluent and stuttered syllables, with different types of stutter being indicated by different key strokes. Work aiming to assess the reliability and validity of that program is in progress (J. Yaruss, personal communication).

Another instrument based on similar principles is TrueTalk, used in Lidcombe work to date (Lincoln & Harrison, 1999). TrueTalk is a dedicated electronic device used to measure speech fluency (http://www.synelec.com.au/synergy/). It utilizes two buttons, one of which is used to indicate when a fluent syllable is heard and the other to indicate when a stuttered syllable is heard.

The Computerized Scoring of Stuttering Severity version 2 (CSSS-2.0) program (Pro-ed, Austin, TX) (Bakker, 2009) is supplied with SSI-4. It is used to count syllables and the frequency and duration of stuttered syllables. Syllables and stuttered syllables are indicated by clicking the left and right buttons of a mouse, respectively (Riley, 2009). Duration of stutters is measured by holding down the key used to indicate stuttering over the length of the stutter. Reliability and validity have not been published for use of the CSSS-2.0 but K. Bakker (personal communication) has confirmed that this has been done. The SSI-4 manual recommends that users should practice before they use the program in the clinic but it does not indicate how much practice is needed nor does it give any advice about how to monitor performance.

Ingham, Bakker, Moglia, and Kilgo (1999) provided the Stuttering Measurement System (SMS) software that allows a clinician to measure stuttering frequency, speech rate and speech naturalness (http://www.speech.ucsb.edu/allFaculty.php?view=roger). SMS requires the listener to indicate fluent and stuttered sections using a mouse. Speech naturalness was rated between 1 and 9 using the numeral keys. An advance made by Ingham et al. (1999) was that they provided a step-by-step training program to accompany the software. The training program takes the user through stages in which difficulty and complexity are increased (http://sms.id.ucsb.edu/downloads/SMS_Manual.pdf) using data for training that are accessed through the program (the latest version of the SMA training program is currently its 2011 edition). As well as the speech samples provided, expert syllable and stutter measurements are given for these samples so that trainees can monitor their achievement level. Ingham et al.’s (1999) method of monitoring accuracy during training with the SMA program compared a user’s scores with the expert scores that were calculated using a transcript method.

iPhone applications are available for counting stuttering frequency. Smarty Ears’ The Disfluency Index Counter is dedicated stuttering frequency counter that can be used on an iPhone or on an iPad. (http://itunes.apple.com/us/app/disfluency-index-counter/id366359722?mt=8#, accessed on 03/04/2011). B.J. Danitz’s The Duo Counter is a general device for the
iPhone or iPad that counts two responses at once (http://itunes.apple.com/gb/app/duo-counter/id312957534?mt=8#, accessed on 03/04/2011). It can be used to obtain stuttering frequency. These apps can be used to count fluent and stuttered, syllables in a pre-recorded sample or in current speech whilst a speaker is present at assessment. Such tools are convenient but there is no indication as to whether the performance of the devices has been checked nor about how users should employ them, or any indication of whether training is necessary. Extra possible sources of variability for this class of devices over and above those considered earlier arise, mostly due to the fact that they employ touch screens (e.g. touch responses may bounce off the surface of the screen).

All the programs and devices that are available to assess stuttering frequency work in a similar way, although they differ in the exact procedures they employ. Some procedures require the listener to use one hand when making two responses (e.g. when counting syllables and stuttered syllables by using a computer mouse), e.g. the CSSS-2.0 (Bakker, 2009). Generally speaking, the assessment instruments do not have a facility for storing responses and comparing them against the original speech files. Other procedures allow the user to employ different hands to make these responses (e.g. the keyboard-based assessments used in the Disfluency Frequency Counter (Yaruss, 1999) and in dedicated counting devices such as TrueTalk).

Rationale for study on assessments made simultaneously or successively

The question of whether making responses regarding overall syllables, stuttered syllables and duration of stutters simultaneously or successively led to different performance was examined in the current study. In both conditions, participants counted syllables (responding with one finger) and stuttering events and durations of the stutters (responding with a different finger). The responses were made either both at the same time (i.e. simultaneously) or a single response was made in each of three sessions (i.e. successively). Based on Howell et al. (2011) and Yaruss et al. (1998), the hypothesis was that scores for all three parameters (syllables, stuttered syllables and stutter durations) would differ when the counts were made successively than when they were made simultaneously. Recommendations concerning how results for these different procedures can be compared are made.

Method

Participants

Ten female speech-language pathology (SLP) students, aged between 22 and 25 years, took part in the study. Each of them had clinical experience as students over the course of their 4-year SLP degree. They had all received the same listener training by SLPs from a London clinic that specializes in childhood stuttering in the year prior to the study. They had assessed people who stutter in clinics using the types of assessment instruments and procedures reviewed in the introduction (including computer-based and counting devices).

Speech samples

Speech samples from seven speakers who stutter were used. The speech samples were from children and adolescents who were diagnosed as having a stutter by trained SLPs. They comprised five boys and two girls aged between 11 and 19 years. Stuttering severity scores ranged from 28 to 38 on SSI-3. The samples all contained 200 or more syllables, which is the length criterion used
in SSI-3 and SSI-4 (Riley, 1994, 2009). Table 1 contains information on gender, age, SSI-3 score and, from the reference transcriptions, number of stuttered syllables and overall syllables. The speech samples were all spontaneous monologues as described in Howell, Davis, and Bartrip (2009).

Transcriptions of speech samples

The original transcriptions were made by trained researchers (who did not participate in this study). This was done using Speech Filing System (SFS, http://www.phon.ucl.ac.uk/resource/sfs/) facilities to transcribe the entire speech samples with annotations of syllables and stuttered syllables that have been assessed for reliability (Howell et al., 2009). SFS can display the speech waveform aligned against wideband and narrowband spectrograms. Any section of the waveform can be selected and zoomed, and the selected section played repeatedly. Once a user makes a judgment about the transcription for the sample, permanent annotations can be created that are displayed aligned in time alongside the waveform. The transcriptions were used in this study for examining accuracy of responses of participants to different speakers.

Procedure for experimental participants

Audindex software was used to assess the simultaneous and successive procedural variants. Audindex software records keypress responses that participants make as they listen to speech. Audindex mimics procedures used in SSI-4 (Riley, 2009) and CSSS-2.0 (Bakker, 2009). The audio is replayed using the Windows API for waveform input/output. Keyboard presses are captured as windows messages. When a key press arrives in the program’s message loop, the current audio position (in samples) of the waveform output device is queried. This position is converted to time and stored. There is some delay and some indeterminacy in the time measurement caused by the use of the Windows message loop to process key presses. This was estimated as less than 20 ms. The speech samples were played individually to participants over Sennheiser HD212Pro headphones (Sennheiser, Wedemark, Germany).

The seven speech samples were assessed by each participant according to simultaneous and successive procedures. Order that participants received the simultaneous and successive conditions was counterbalanced across participants. Syllables, stuttered syllables and duration of stutters were assessed simultaneously on each of the seven samples in the simultaneous procedure. Participants listened to a sample and responded with one index finger on a specified key for each syllable heard and responded on another key with the other index finger when

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Gender</th>
<th>Age</th>
<th>SSI-3</th>
<th>Number of stuttered syllables/all syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female</td>
<td>12 years 4 months</td>
<td>28</td>
<td>38/201</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>11 years 6 months</td>
<td>34</td>
<td>54/256</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>11 years 0 months</td>
<td>33</td>
<td>60/364</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>13 years 0 months</td>
<td>36</td>
<td>25/283</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>19 years 1 months</td>
<td>36</td>
<td>16/284</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>14 years 1 months</td>
<td>38</td>
<td>34/310</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>12 years 0 months</td>
<td>32</td>
<td>30/223</td>
</tr>
</tbody>
</table>
a stuttered syllable was heard (the index finger mapping to responses was counterbalanced across participants). The key used to indicate stuttered syllables was held down until the end of a stutter in order to obtain the duration of the stuttering event. Audindex monitored and stored the syllable, stuttered syllable and duration responses.

Audindex was also used to count the syllables, stuttered syllable and durations for the seven speech samples in the successive condition. Participants responded as in the simultaneous condition but responses were made either about syllables, stuttered syllables or stutter durations each time the speech file was played. Order that participants made the different responses was counterbalanced. They heard each sample three times, indicating in turn syllables, stuttered syllables or stutter durations (order randomized across participants). Nothing was displayed on the screen in either the simultaneous or the successive condition (so participants focused on listening to the audio-samples and concentrated on making keypress responses).

Procedure for analysis

The keypresses detected by Audindex that represent the different judgments were stored in SFS files so that they could be displayed aligned in the position in time where they had occurred relative to the original speech waveform and transcriptions. Files in SFS format can display multiple analyses of the speech in alignment (e.g. the original speech, the syllable and the stutter transcriptions and each of the Audindex responses). The responses from Audindex were uploaded into SFS and displayed against the original speech and the reference transcriptions. The syllable and stuttered syllable scores were counted by determining the number of key strokes. The duration score used in SSI-3/CSSS-2.0 was calculated by averaging the three longest key strokes (Riley, 1994). The %SS was calculated from the syllable and stuttered syllable counts from participants responses (Riley, 1994). Comparison was made between syllable and stutter counts, and duration made simultaneously or successively by related t-tests. The steps in the procedure and resultant extracts from SFS files after performing transcription, simultaneous (one plot) and successive (three plots) conditions on these extracts are indicated schematically in Figure 1.

Results

The differences between the simultaneous and successive conditions were analyzed by separate related t-tests for syllable counts, stuttered syllable counts and the estimated duration of the three longest stutters. The differences between counts across the successive and simultaneous conditions were significant for syllable ($t = 2.884, p = 0.018$) and stuttered syllable ($t = 2.545, p = 0.031$), the number of degrees of freedom was 9 in both analyses. In both cases, means for the successive condition had higher counts than the simultaneous condition (syllables: successive mean = 1848.6, SE = 60.95 and simultaneous mean = 1696.1, SE = 43.65; stutters: successive mean = 179, SE = 10.69; simultaneous mean = 157.5, SE = 8.70). These statistics are summarized in row 1 of Table 2. The average counts across the seven speakers for the successive, simultaneous and reference transcriptions are given in the following row. From the averages, it can be seen that the syllable and stutter counts by the judges were always lower than those in the transcriptions. Relative to the reference transcriptions, judges detected 69.8% of the stuttered syllables in the successive condition and 61.3% of the stuttered syllables in the simultaneous condition.

The average duration of the three longest key strokes was not significantly different between the successive and simultaneous conditions ($t = 0.091, p = 0.930$). For the successive condition, mean = 2.38 and SE = 0.226, and for the simultaneous condition, mean = 2.37 and SE = 0.188. The duration scores varied across participants with some of the counts being overestimated and
The average duration of the three longest stutters was 81.8% of those in the reference transcriptions for the successive condition and 81.4% of those in the reference transcriptions for the simultaneous condition.

A qualitative analysis of speakers showed that speaker 7 had quick syllable and part-word repetitions and blocks. These were sometimes so brief that the flow of speech in that segment was

Figure 1. Summary of results from the procedures for transcriptions, simultaneous and successive conditions on speech extracts. The speech extracts are in SSF format. Transcription results in one SFS item that contains syllables, stutters and their durations. The successive procedure results in three separate SFS files with the requisite information for the response made. The simultaneous procedure produces one SFS file with all information coded in one item.

<table>
<thead>
<tr>
<th></th>
<th>Successive</th>
<th>Simultaneous</th>
<th>Reference transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of syllables</td>
<td>1848.6</td>
<td>1696.143</td>
<td>1696.143</td>
</tr>
<tr>
<td>Av = 264.1</td>
<td>Av = 242.3</td>
<td>Av = 274.4</td>
<td>SD = 20.6</td>
</tr>
<tr>
<td>Number of stutters</td>
<td>179.0</td>
<td>157.5</td>
<td>157.5</td>
</tr>
<tr>
<td>Av = 25.6</td>
<td>Av = 22.50</td>
<td>Av = 36.7</td>
<td>SD = 5.90</td>
</tr>
<tr>
<td>% Syllables stuttered</td>
<td>9.81</td>
<td>9.30</td>
<td>9.30</td>
</tr>
<tr>
<td>Av = 13.63</td>
<td>Av = 12.55</td>
<td>Av = 2.91</td>
<td>SD = 2.11</td>
</tr>
<tr>
<td>Duration</td>
<td>2.38 s</td>
<td>2.37 s</td>
<td>2.37 s</td>
</tr>
<tr>
<td></td>
<td>0.226</td>
<td>0.188</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>Av = 2.91</td>
<td>Av = 2.91</td>
<td>SD = 0.401</td>
</tr>
</tbody>
</table>

Av, average counts for the seven speakers are given in the second row beneath ‘‘Number of syllables’’ and ‘‘Number of stutters’’. These can be compared directly with the counts in the reference transcriptions (far right columns).

some being underestimated (the majority) relative to those in the corresponding transcriptions. The average duration of the three longest stutters was 81.8% of those in the reference transcriptions for the successive condition and 81.4% of those in the reference transcriptions for the simultaneous condition.

A qualitative analysis of speakers showed that speaker 7 had quick syllable and part-word repetitions and blocks. These were sometimes so brief that the flow of speech in that segment was
not noticeably disrupted and was not easy to judge when the sample was listened to in real time. The speech also contained many filler words such as ‘‘er’’ which themselves are not symptoms of stuttering but do disrupt the flow of the speech (Jiang, Lu, Peng, Zhu, & Howell, 2012). The fillers may, therefore, mask the occurrence of stutter symptoms, thus affecting judgments regarding speech fluency. From this analysis, it seems that stutter type and speaking style may play a part in stuttering frequency assessment.

Discussion

Overall syllables, stuttered syllables and their durations were examined in the experiment because they are assessed in other widely used procedures (Bakker, 2009; Riley, 1994, 2009; Yaruss, 1999) and counting devices such as TrueTalk. Results indicated that using a successive procedure gives different scores compared with using the simultaneous procedure for syllables and stuttered syllables. The patterns of performance showed that, in general, participants gave higher syllable and stuttered syllable counts in the successive condition than in the simultaneous condition. The successive procedure was more time consuming than the simultaneous procedure. As the same speech sample results in different stuttered syllable estimates when different procedures are employed, then separate norms are required for the different procedures. This does not imply that one method is preferred over the other as both are useful for different applications and sometimes both may be required in the same application. For instance Howell (2013) outlined a screening procedure that is applied to children in schools for fluency problems that scores syllables, stutters and durations. It would be impractical to employ the more time-consuming successive procedure to do this with all the children given that stuttering has a low incidence in such samples. One possibility would be to assess all children with the simultaneous procedure and to re-examine children detected by this procedure with the successive procedure. According to this suggestion, some children would be examined by both procedures and so it is necessary to know the relationship between the scores which could be done if separate norms were available for each of them.

In the experiment, it was observed that counting syllables, stuttered syllables and stutter durations separately, led to different results than assessing them together. Stuttering assessment instruments should provide training packages that include practice samples with baseline syllable, stuttered syllable and duration counts available for comparing results and monitoring progress according to different procedures (Ingham et al., 1999). Audindex could also be used for these purposes, with users’ own materials. Training may be time-consuming but it helps to ensure that assessments carried out in clinics are not only rapid but also accurate and comparable across clinics. Inter- and intra-rater reliability also need attention. For instance, it was noted that duration scores varied across participants (some participants’ duration estimates were longer than those in the transcription and some were shorter).

A limitation of this study was the small number of participants who judged speakers. It would be useful to extend this study to a larger sample of SLPs who were required to judge a larger number of speakers. Other issues to address are whether accuracy of frequency and duration scores varies between audio samples and video samples, and whether the type and combination of stutters and speech style of a speaker affect the accuracy of assessment.

Declaration of interest

The authors report no conflict of interest. This research was supported by a grant from the Domenic Barker Trust, Ipswich, UK to Peter Howell.
References


Ingham, R. J., Bakker, K., Moglia, R., & Kilgo, M. (1999). Stuttering measurement system (SMS) [computer software]. Santa Barbara, CA: University of California.


