Scores on Riley’s Stuttering Severity Instrument Versions Three and Four for samples of different length and for different types of speech material

HELENA TODD¹, AVIN MIRAWDELI¹, SARAH COSTELLOE², PENNY CAVENAGH², STEPHEN DAVIS¹, & PETER HOWELL¹

¹Division of Psychology and Language Sciences, University College London, London, UK and ²School of Health and Human Sciences, University Campus Suffolk, Suffolk, UK

(Received 1 April 2014; revised 15 May 2014; accepted 19 May 2014)

Abstract
Riley stated that the minimum speech sample length necessary to compute his stuttering severity estimates was 200 syllables. This was investigated. Procedures supplied for the assessment of readers and non-readers were examined to see whether they give equivalent scores. Recordings of spontaneous speech samples from 23 young children (aged between 2 years 8 months and 6 years 3 months) and 31 older children (aged between 10 years 0 months and 14 years 7 months) were made. Riley’s severity estimates were scored on extracts of different lengths. The older children provided spontaneous and read samples, which were scored for severity according to reader and non-reader procedures. Analysis of variance supported the use of 200-syllable-long samples as the minimum necessary for obtaining severity scores. There was no significant difference in SSI-3 scores for the older children when the reader and non-reader procedures were used. Samples that are 200-syllables long are the minimum that is appropriate for obtaining stable Riley’s severity scores. The procedural variants provide similar severity scores.

Keywords: Developmental stuttering, percent syllables stuttered, screening, stuttering severity instrument

Introduction
The aim of this paper was to assess some aspects of Riley’s widely used Stuttering Severity Instrument versions Three and Four, SSI-3 and SSI-4 (Riley, 1994, 2009). This instrument provides an estimate of stuttering severity based on speech recordings, and the instrument has been evaluated statistically. There are separate versions of the instrument for readers (standardized using spontaneous and read speech samples) and non-readers (standardized using spontaneous speech samples alone). In some applications where the SSI instrument is used, only a short time is
available for conducting the test and, in these situations, assessing a short speech sample would be advantageous. Although there are separate forms for readers and non-readers, the relationship between the scores for the two forms has not been examined and this is necessary in longitudinal work when young children who cannot read are followed up to later ages where the reader form is appropriate. The issues of minimum sample length and equivalence of reader/non-reader procedures were addressed in the empirical work reported below. The procedures used with SSI-3 and SSI-4 are described in the next section and then an assessment of the instrument is made before the empirical work is reported.

**Riley’s SSI-3 and SSI-4 instruments**

All versions of the SSI derive scores from measures made on both a speech sample (percentage syllables stuttered, %SS, and duration of the three longest stutters) and observations of physical concomitants made at the time of the recording. There are tables that convert %SS and duration measures into scores that are combined with the raw physical concomitant score to give the total overall SSI score. While it is frequently emphasized that assessment of stuttering involves aspects over and above counts of symptoms in speech samples, this does not mean that speech measurements are unimportant. The importance of speech assessment is supported by the observations that SSI-3 has been used to report details of people who stutter in more than 350 publications and as it has been translated into other languages (Bakhtiar, Seifpanahi, Ansari, Ghanadzade, & Packman, 2010).

**Important design features**

The procedures for making SSI-3 and SSI-4 estimates are detailed in the method. Four design features that are strengths of the instrument are highlighted in this section. The first feature is that the events that count as stutters are stated explicitly. The symptoms that are considered as stutters are identical in SSI-3 and SSI-4 and are described in detail in the SSI-3 manual (Riley, 1994), where the following events are included as stutters: “repetitions or prolongations of sounds or syllables (including silent prolongations)” (Riley, 1994, p. 4). Riley (1994) also notes some of the events which are not counted as stutters: “Behaviors such as rephrasing, repetition of phrases or whole-words, and pausing without tension are not counted as stuttering. Repetition of one-syllable words may be stuttering if the word sounds abnormal (shortened, prolonged, staccato, tense, etc.); however, when these single-syllable words are repeated but are otherwise spoken normally, they do not qualify as stuttering using the definition just stated” (Riley, 1994, p. 4). The counts of syllables stuttered are represented as a percentage of all syllables.

The second feature is that there is some flexibility in the procedures that can be used to obtain speech samples. Provision is made in the SSI instrument for the assessments to be made in clinical, home, or laboratory settings.

Third, the minimum required sample length is specified. The minimum sample length given in Riley (1994, p. 9) is 200 syllables. Riley reduced the minimum sample length to 150 syllables for SSI-4 (2009, p. 6). No investigations have been reported on whether 150- or 200-syllable-long samples give reliable SSI-3 or SSI-4 estimates.

Fourth, separate reader and non-reader assessment procedures are given. This allows further flexibility in terms of the range of ages over which the instrument can be employed. Readers provide read and spontaneous samples while non-readers provide a spontaneous sample only.
Statistical assessment

Riley (1994) reported on the reliability and validity of SSI-3. No new statistical assessments were made when SSI-4 appeared and the norms are identical to those used in SSI-3. Riley (1994) assessed the intra-, and inter-judge reliability of SSI-3 component scores (i.e. after conversion for %SS and duration components). Intra-judge reliability concerns how reproducible the results are for an individual, and this was assessed by five judges each of whom estimated %SS and duration twice on 17 samples. Mean agreements ranged from 75 to 100%. Inter-judge reliability between 15 judges was estimated for all three components that make up SSI-3. Agreement ranged from 94.6% to 96.8% for %SS, from 58.1% to 87.2%, for duration, from 59.8% to 97.5% for physical concomitants, and from 71% to 100% for overall scores. Intra- and inter-judge reliabilities were deemed by Riley to be satisfactory.

Checks were made for criterion validity (independent measures that should be related to SSI-3) and construct validity (assessment of the internal components in SSI-3). In the check on criterion validity, Riley (1981) showed that SSI-3 scores correlated with scores from his Stuttering Prediction Instrument. For the assessment of construct validity, Riley (1994) showed that total SSI-3 scores correlated significantly with each of its components (frequency, duration, and physical concomitants). Riley (1994) concluded that SSI-3 achieved acceptable standards of validity.

Lewis’s (1995) statistical evaluation of SSI-3 was less favorable than that of Riley: she first replicated the agreement levels on tabled scores and reported that these were close to those given by Riley (1994) and were deemed satisfactory. However, raw pre-conversion scores led to markedly lower agreement levels. She also reported that SSI-3 scores were closely related to subjective estimates of severity. She concluded that a particular SSI-3 designation (mild, moderate, etc.) could reflect a considerable range of stuttering behaviors. Overall, she suggested that the SSI-3 does not represent a reliable or valid measure.

The main applications of the instrument

Riley (1994) indicated that the SSI-3 could be used as a part of diagnostic evaluations that it could assist in tracking changes in severity during and following treatments and that it could be used to validate other assessment instruments.

Evaluation

While it has been noted earlier that Riley (1994) designed his SSI-3 so that it was flexible, the flexibility is limited by the fact that the norms and statistical evaluation (subsequently ‘standards’ refers to both of these) only apply when the same procedures used in norming are employed. As the same standards are applied in SSI-4, only those procedures allowed in SSI-3 can be used if the norms are employed. Several changes in the procedure are suggested in SSI-4. However, if these are implemented, either the test needs to be restandardized or it has to be shown empirically that the changes do not affect the scores obtained. If any restandardization is performed, care should be taken to ensure that this is conducted on a representative sample of speakers of the target language (Riley’s test was standardized on a sample of participants in California).

One major change that Riley (2009) introduced in SSI-4 compared with SSI-3 was in the method used to count stutters; a software counter that automatically indexes syllables and stutters and stutter durations was employed. One key of a mouse has to be tapped to count syllables,
whereas another key counts stutters. The key for stutters has to be kept pressed down for as long as each stutter goes on, thus providing an estimate of stutter durations. Using the counter is difficult and it affects the severity measure (Jani, Huckvale, & Howell, 2013). In turn, if this procedure was used as advised in the SSI-4 manual, the norms, which were obtained according to the SSI-3 procedures, would not apply.

In order to ensure that the standards apply, there can be no flexibility in interpreting the guidelines in the SSI-3 handbook concerning which symptoms to include. In their research, different authors do or do not include whole-word repetitions in the counts of %SS. Although there are pros (Yairi, Ambrose, Watkins, & Paden, 2001) and cons (Brocklehurst, 2013) concerning whether whole-word repetitions should be included as stuttering symptoms, Riley’s (1994, 2009) ruling has to be adhered to, otherwise the SSI-3 and SSI-4 scores would not be correct and the norms would not apply.

Even alterations that seem advisable on a priori grounds should not be made when using the instrument unless the test is restandardized. For instance, while Riley (2009) suggests that audiovisual recordings may be substituted for audio recordings, as they would allow objective assessment of physical concomitants, only audio recordings were used when the standards were established; Riley (1994) does not mention that audiovisual recordings were used when the normative data were analyzed. Rousseau, Onslow, and Packman (2008) reported that stuttering frequency and severity were higher with audiovisual than audio-alone material. Therefore, an audiovisual recording would affect %SS and duration estimates as well as physical concomitants, making comparison with reports where the scores were obtained from audio records impossible. Another example where changes are prohibited concerns physical concomitants. It has been questioned as to whether they should be included as they are the most problematic aspect of SSI-3 (Bakhtiar et al., 2010; Lewis, 1995). The important point for the current discussion is that they have to be retained for the standards to apply.

Additional elements cannot be added after a test has been normed. While it may seem self-evident that including more types of speech samples is an advantage in getting a more representative impression of a client’s speech (Riley, 2009), the standards only apply for the types of speech used in norming. The norms for readers are based on read and spontaneous samples alone (Howell, 2013). As the standards were obtained with these types of speech alone, SSI-3 scores are only correct when just these forms of speech are used. Consequently, forms of speech not used in standardization should not be included when making the SSI-3 scores (in contrast to what Riley, 2009, recommends).

SSI-3 (Riley, 1994) specifies 200 syllables and SSI-4 (Riley, 2009) specifies 150 syllables as the minimum sample duration required to obtain a stuttering severity score. The minimum duration should not have been changed for SSI-4 as the instrument was not restandardized. Consequently, 200 syllables should be the minimum sample length when SSI-4 is used.

Arguments have been presented about the need for longer samples. For example, short samples may miss some stuttering (Logan & Haj-Tas, 2007; Roberts, Meltzer, & Wilding, 2009; Sawyer & Yairi, 2006; Yairi & Ambrose, 2005; Yairi & Seery, 2011). Riley also suggests samples of 300–500 syllables in length are more reliable, but give no statistical support for this. The maximum sample length is obviously an issue when precise profiling of an individual’s stuttering severity is required. However, not all applications where SSI-3 and SSI-4 are used require this level of detail. Presumably, this is why Riley (1994) specified the minimum length of a sample required for his instrument to be used. One example where using minimum length samples may be useful is in clinics where time is precious. Another example is where SSI-3 is used in studies to specify details of samples of speech from people who stutter where the focus of the study is not speech (and so a representative sample of all stuttering events may not be required).
A final example is when SSI-3 is used for screening large samples of children for stuttering (Howell, 2013). Most children in these cohorts are fluent. If 200-syllable-long samples provide a reliable standardized SSI-3 score, samples of this length could be obtained on all children. A threshold value of SSI-3 can then be employed (Howell, 2013; Howell & Davis, 2011) to determine which children need to be referred for further investigation. The use of a short sample, providing it gives a reliable estimate, would be efficient for work in schools where little time can be spared from other core activities. If it can be shown that SSI-3 based on 200-syllable samples provides reliable estimates, then there would be no reason to take longer samples when screening children for fluency in schools when this standardized instrument is used. It is important to bear in mind that establishing what minimum sample length can give a reliable SSI-3 estimate for this and the other applications mentioned above does not carry implications about whether longer samples and ancillary information are needed for more detailed clinical examinations of children who stutter (Logan & Haj-Tas, 2007; Roberts et al., 2009; Sawyer & Yairi, 2006; Yairi & Ambrose, 2005, Yairi & Seery, 2011). For example, the Sawyer and Yairi (2006) study that used a %SS estimate that counted different symptoms to Riley (1994, 2009), looked at variability of stuttering across longer speech samples in 20 children who had been diagnosed as stuttering previously. They noted differences in the rate of their %SS measure across the length of the sample (beginning/end) and detected shifts in %SS as sample length increased. Here, the question is whether Riley’s stipulated minimum is sufficient for a reliable SSI-3 or SSI-4 score to be obtained, not whether longer samples are required for clinical assessment of children already known to stutter.

Summary and research questions

Despite some limitations, SSI-3 performs well, as witnessed by its widespread use by many authors, being a brief and versatile test to conduct. Its brevity is in part due to the fact that it only requires samples of 200 syllables in length at minimum. It is versatile insofar as it allows procedures that can be used in clinics or research laboratories, and there are forms for children who cannot read and for older children who can read.

Nevertheless, it is not apparent why the minimum length of the sample was set at 200 syllables and whether this is sufficient to obtain a stable SSI-3 score. As mentioned, authors who have been taking speech samples for clinical assessment of children already known to stutter have considered that longer samples are required (Sawyer & Yairi, 2006; Yairi & Ambrose, 2005). Of course, it does not necessarily follow that a sample of speech 200 syllables in length is too short to obtain a reliable SSI-3 score estimate. Consequently, one thing tested in the experiment reported below is whether a 200-syllable-long sample provides a stable SSI-3 score.

It was pointed out that only types of samples that were used during standardization can be used to obtain valid SSI-3 and SSI-4 scores. A related matter concerns the relationship between scores that used one type of material (spontaneous material with children who cannot read) or two (read and spontaneous), for which separate conversion tables are supplied (Riley, 1994, 2009). When screening takes place when children enter schools, they can only be assessed with the non-reader form. If the children who do not pass the screen are followed up, they would be assessed at later ages with the reader form. Consequently, it is necessary to know whether or not the scores from the two forms give corresponding values in longitudinal work in which children may start with the one-sample non-reader form and progress when older to the two-sample reader form. Otherwise, changes in SSI-3 scores over ages where two forms are used may arise either because speech behavior changes (because intervention or spontaneous recovery has been effective) or
because the change in procedure itself gives rise to different scores (spurious effects attributed to intervention or recovery processes). To rule out the second possibility, it is necessary to establish whether the reader and non-reader forms produce equivalent results. This is the second purpose in the work reported below.

Methods

Participants and recordings

There were 23 children in the younger age group, who were recruited by a speech–language therapist based in Suffolk (SC) and had been diagnosed as stuttering. There were 18 males and 5 females with ages ranging from 2 years 8 months to 6 years 3 months (mean age: 4 years 7 months, SD: 1 year 0 month). None of the children in the younger age group were able to read; a spontaneous recording was made of these children. Topics chosen to use in the spontaneous recording were ones that children could readily talk about. Examples were hobbies, family activities, favorite TV programs, etc. These were conducted in a quiet room and the speech was recorded on a Sony DAT recorder using a Sennheiser K6 condenser microphone. Physical concomitants were rated according to Riley (1994) at the time of the recording. Collection of spontaneous speech samples and assessment of physical concomitants were either conducted by the fifth author or he trained one of the other authors how to do these tasks. All the researchers who collected data and scored the speech were supervised by the corresponding author.

There were 31 children in the older age group. They attended a specialist clinic dealing with stuttering in London. There were 22 males and 9 females and the age range was 10 years 0 months to 14 years 7 months (mean age: 13 years 0 month, SD: 1 year 1 month). Spontaneous and read speech samples were made for these children. Spontaneous samples were obtained as with the younger group of children. All children in this age group were able to read the age-appropriate material supplied in Riley (1994). Thus, the reading materials for 10–11 year olds, 12–13 year olds and for speakers older than 13 years were used. Physical concomitants were noted at the time of the recording as with the younger children. The physical concomitant scores were made in the same way as with the younger children (they were not made by clinicians).

Pre-processing of speech samples and reliability assessment

All audio recordings were uploaded and coded using Speech Filing System (SFS) software (Huckvale, 2013) obtainable as free download from http://www.phon.ucl.ac.uk/resource/sfs/. The recordings were transcribed orthographically in a format that allowed a syllable count to be made (Howell, Davis, & Williams, 2008). Two hundred and fifty syllables were transcribed for each recording and, in the case of the older children, for both types of material.

Interjudge reliability measures were obtained by getting two trained transcribers to independently assess the speech materials from eight speakers chosen at random. The agreement between the two judges on fluency was 96% and on syllables was 93%, giving kappa coefficients of 0.92 and 0.89 which show agreement is well above chance (Fleiss, 1971). The three stutters selected as having the longest durations were identical for the two judges (SFS displays times on the x-axis). The average durations of the three longest stutters for the eight speech samples did not differ significantly by the t test across the judges. Interjudge reliability scores were not available for physical concomitants as they were made at the time of the recording when only one person was present.
Procedures

The SSI-3 procedure is described for the older children who can read. Then the modifications made when the test is administered to the younger non-reader children are described. For readers, procedures for obtaining frequency, duration, and physical concomitant scores are given. Following that the way in which the samples were divided to provide shorter samples for the length analyses are detailed.

**Administration of the SSI-3 to readers**

*Procedure for obtaining frequency scores and counting total syllables*

Separate counts were made of all syllables spoken and those syllables that were stuttered. Syllable counts were obtained directly from the transcriptions in the SFS files. Non-word fillers such as ‘‘erm’’ were counted as words, so were thus included in the total syllable count, consistent with examples given in the SSI-3 manual (Riley, 1994).

*Procedures for counting stuttered syllables*

The symptoms defined in the SSI-3 manual were counted (Riley, 1994). Each stuttering episode was counted as one stutter, so in the following example, there is one stutter out of five syllables:

(a) A a a a a and I want that one
(b) 1 2 3 4 5

Each repeated syllable in whole-word repetitions was included in the syllable count unless any of the repetitions had other signs of stuttering, so, in the below example, there are eight syllables in total and no stutters.

(a) And and and and I want that one
(b) 1 2 3 4 5 6 7 8

For read and spontaneous samples, %SS was calculated and converted to a task score using tables provided in the SSI-3 manual (Riley, 1994).

**Duration score**

Duration of each stutter was obtained using the SFS display and replay facilities, the duration of a stuttering episode being the time from the start of the stutter to the end of the final release of the syllable involved in the stutter. For readers, the three longest stutters were obtained separately for the spontaneous and read samples. Then the durations of the three longest stutters of the six selected stutters (irrespective of sample) were taken, averaged, and converted using the tables in Riley (1994).

**Physical concomitants**

Physical concomitants were scored at the time of the recordings according to the SSI-3 criteria (Riley, 1994), and comprised distracting non-speech sounds, facial grimaces, head movements, and movements of the extremities, each given a score from 0 (none) to 5 (severe and painful looking). The ratings for the four criteria were then summed which allowed a maximum possible score of 20 for physical concomitants.
Total overall score and differences when non-reader scores were obtained

The raw SSI-3 scores were used throughout in this study, not severity categories. These were obtained separately for all sample lengths for both groups of children. The 250-syllable sample was divided into five successive non-overlapping 50-syllable sub-sections as described below, and separate SSI-3 scores were calculated for each of them. SSI-3 scores were based on spontaneous and read material for readers, and on spontaneous material alone for non-readers. The total overall SSI-3 score for a reader was obtained by adding together the task scores for frequency and duration and the raw score for physical concomitants.

The modifications for non-readers were as follows: (1) the tables provided by Riley (1994) were used to convert the %SS from the single spontaneous sample score to a task score; (2) the average of the three longest durations was based on the spontaneous sample alone; (3) physical concomitants were based on the observations on the spontaneous sample alone.

For the readers, SSI-3 scores were also calculated by ignoring the read sample and estimating the scores (as if these older children were non-readers) to establish whether these procedures gave different results when applied to the same children. Riley (1994) gives different conversion tables for %SS (but not duration or physical concomitants) for use with participants who cannot read (non-reader form) and for those who can read (reader form).

Sub-division of the recordings into sections with different numbers of syllables

The 250-syllable section of each recording that was transcribed (for both younger and older children) was divided into sub-sections that included the first 50, 100, 150, 200, and 250 syllables. SSI-3 scores were calculated using %SS and duration in the extract and the physical concomitant score. The entire 250-syllable section was also divided into five successive 50-syllable sub-sections. These were used to make additional SSI-3 assessments using %SS and duration within each sub-section and the physical concomitant score to see whether there were any positional effects across the 250-syllable extract.

Analyses

IBM SPSS Statistic 21 (SPSS Inc., Chicago, IL) was used to conduct all analyses. Data from the two age groups were analyzed separately because the scoring procedures and factors in the analysis of variance (ANOVA) differed.

Results

Younger children

The SSI-3 scores for different sample lengths were compared by ANOVA. The five SSI-3 estimates for the 50-syllable extracts from the overall transcribed section were also compared by ANOVA.

Sample length

A one-way ANOVA was performed on the SSI-3 scores for the younger children with length of sub-section as the factor (50, 100, 150, 200, or 250 syllables). Mauchly’s test indicated that sphericity had been violated. Consequently, the degrees of freedom were adjusted using the
Greenhouse–Geisser estimates. There was a main effect of sample length ($F_{2.354, 51.784} = 8.515$, $p < 0.001$). The means and 95% confidence intervals are shown for each sample length in Figure 1.

Figure 1 reveals that SSI-3 scores decreased as the sample size decreased with the exception that the 250-syllable sample had similar SSI-3 scores to the 200-syllable sample. The post hoc $t$-tests (summarized in Table 1) supported these impressions. Significant differences occurred between the scores for 200-syllable sample lengths and the scores for samples with 100 or fewer syllables after Bonferroni correction was made (a $p$ value of 0.01 was used).

**Sample position**

A further one-way ANOVA was performed on the five 50-syllables samples’ SSI-3 scores. There was no significant difference across sample positions ($F_{3.87, 85.11} = 1.112$, $p = 0.355$).
The data are shown in Figure 2, where it can be seen that SSI-3 scores increased monotonically over sample positions (although this was not significant).

**Older children**

*Sample length and procedure type*

An ANOVA with two within-group factors was carried out using SSI-3 scores as the dependent variable. The factors were sample length (50, 100, 150, 200, and 250 syllables) and procedure (scored according to the reader procedure using both spontaneous and read samples versus scored according to the non-reader procedure using spontaneous samples alone). Mauchly’s test showed that sphericity was violated for the main effect of sample length and for the interaction between sample length and procedure, so degrees of freedom were adjusted with the Greenhouse–Geisser estimates of sphericity. The main effect of procedure ($F_{1, 30} = 0.40, p = 0.531$) and the interaction between sample length and procedure ($F_{2.13, 63.83} = 0.48, p = 0.635$) were not significant. However, the main effect of sample length was significant ($F_{2.15, 64.38} = 10.08, p < 0.001$).

Figure 3 shows that, again, there was a trend for SSI-3 scores to decrease as the sample size decreased (with the exception of the longest sample sizes) and here this applied to both procedures. Post hoc $t$ tests were carried out for selected sample-size lengths and procedures and the results are presented in Table 2. Bonferroni corrections were made as with the younger children. The differences between the SSI-3 scores of 250-syllable samples and
Figure 3. Older group of children, mean SSI-3 scores, and ±1 standard errors for different sample lengths. SSI-3 scores obtained with the reader procedure are shown on the left (white rectangles) and for the non-reader procedure on the right (diagonal stripes).

Table 2. t-tests for the older children that compared results for the reader and non-reader procedure SSI-3 scores for a 200-syllable-long sample against other sample lengths.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>t</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250/200</td>
<td>0.34</td>
<td>0.738</td>
</tr>
<tr>
<td>200/150</td>
<td>3.09</td>
<td>0.004*</td>
</tr>
<tr>
<td>200/100</td>
<td>3.79</td>
<td>0.001*</td>
</tr>
<tr>
<td>200/50</td>
<td>4.88</td>
<td>0.001*</td>
</tr>
<tr>
<td>Non-reader procedure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250/200</td>
<td>0.57</td>
<td>0.72</td>
</tr>
<tr>
<td>200/150</td>
<td>1.43</td>
<td>0.163</td>
</tr>
<tr>
<td>200/100</td>
<td>2.50</td>
<td>0.018</td>
</tr>
<tr>
<td>200/50</td>
<td>3.04</td>
<td>0.005*</td>
</tr>
</tbody>
</table>

The results for the reader and non-reader procedures are given at the top and bottom, respectively. Degrees of freedom were 29 in all cases. *Significant at $p<0.01$. 
200-syllable samples were not significant for either the reader or the non-reader procedure. There was a significant difference between the SSI-3 scores for 200-syllable samples and 150-syllable (and shorter) samples for the reader, but not the non-reader, procedure. In the non-reader procedure, the 200-syllable SSI-3 scores differed from the 50-syllable SSI-3 scores.

Sample position and procedure type
The ANOVA that compared SSI-3 scores across sample positions had the additional factor of procedure. Neither procedure ($F_{1, 30} = 0.026, p = 0.873$) and position ($F_{3.046, 91.393} = 0.874, p = 0.459$) were significant nor was the interaction between these two factors ($F_{3.138, 94.137} = 0.558, p = 0.652$). There was only about a two-point change in SSI-3 scores across sample position.

Discussion
Considering the younger children, first, 200-syllable long samples gave the same SSI scores as samples that were 250 and 150 syllables in length. This suggests that SSI scores are stable for 150–250-syllable-long samples. Riley (1994) recommended 200-syllable samples as the minimum and this appears to be appropriate for obtaining stable SSI estimates. Samples less than 150 syllables in length are not adequate, as SSI scores decreased significantly with shorter sample lengths. Second, there was no significant difference between 50-syllable samples across utterance positions (having examined five positions). However, there was a monotonic, non-significant, trend for SSI scores to increase from early to later positions. This did not occur with the older speakers.

With the older speakers, the reader and non-reader procedures were examined in addition to the sample length. There was no effect of procedure (main effect or interaction), but there was a main effect of sample length. The lack of any effect of procedure is consistent with the view that the pattern for reader and non-reader procedures gave similar SSI scores. Looking at the main effect of sample length (collapsed across the two procedures), 250-syllable and 200-syllable samples had the same SSI scores, but all other lengths had significantly lower scores. This showed that 200-syllable samples provided stable SSI scores and would support the use of such samples as the minimum. In a second analysis, there were no differences in SSI scores with respect to position of the sample or procedure for the 50-syllable extracts. The absence of differences across procedures in both analyses with older speakers supports the validity of the non-reader procedure for SSI scores; it gives scores that are not statistically distinguishable from the reader procedure.

The implications that the current results on sample length and reader/non-reader procedures have in general are considered next. When the sample length is discussed, some differences that were observed across age groups are considered.

Length of speech sample and age effects
The conclusion that, for both age groups, a 200-syllable-long sample is adequate as the minimum length sample for making an SSI-3 score does not necessarily mean that 200-syllable-long samples are appropriate for other purposes. Speech sample size is an issue that Yairi and Ambrose (2005) and Sawyer and Yairi (2006) have considered when addressing variability of stuttering across speech samples of different lengths. Also, they critique authors who have used short samples in clinical studies for this purpose: “Speech sample size used in research,
however, has varied greatly across studies, as well as among subjects in the same study. Johnson et al. (1959) included samples that ranged in length from 31 to 2044 words, whereas Schwartz and Conture (1988) used 85–650 words. Many studies in the past two decades were based on the samples of 300–350 words (e.g. Conture & Kelly, 1991). Some samples have been even smaller, with Yaruss (1997) employing 200-syllable samples, and Onslow, Costa, and Rue (1990) using samples as short as 1 minute” (Sawyer & Yairi, 2006). In this quote, Sawyer and Yairi (2006) imply that long samples should be the rule for a range of purposes whereas we consider that different sample sizes may be required for different purposes (in clinics, when the focus of the study is not speech and screening for fluency in schools). Specifically, 200-syllable-long samples appear to be long enough to obtain stable SSI scores. An SSI score provides a common standard that helps healthcare professionals, research groups, and service providers to communicate severity of cases. As discussed in the Introduction, symptom-based assessments such as SSI are not the only feature of stuttering that needs to be measured for detailed case assessment in clinics and research laboratories. A further point, often overlooked, is that SSI is not simply a %SS measure; it includes duration and physical concomitant scores as well.

The minimum sample length necessary for obtaining an SSI score for adults who stutter has not been examined in the current study; the oldest speaker was 14 years 7 months. However, the effect of longer sample lengths on %SS scores has been investigated in two studies which have reported results that conflict with those obtained with children (Sawyer & Yairi, 2006). Logan and Haj-Tas (2007) showed, using the same symptoms as Sawyer and Yairi (2006), that there was no change in %SS over successive 300 syllable-long samples up to 1800 syllables at maximum. Roberts et al. (2009), who used a more extensive symptom set, also reported that there was no effect of sample length in adults who stutter. The reason why children need longer samples than adults merits further investigation.

Procedure

There was no significant difference between SSI-3 scores made according to the reader and non-reader procedures. The results may also suggest that a spontaneous sample scored according to the non-reader form would be sufficient to obtain an SSI estimate whether or not older people who stutter can read or not. This is not something we would advocate, but a spontaneous sample alone could be used with caution when only this material is available or when cross-age group comparisons with the same form are necessary.

Riley (2009) advocated taking a range of samples across situations. Basing assessment on more than one sample is thought to be beneficial in that it gives a view of stuttering in various situations, in which severity may vary (Yaruss, 1997), and so it may be valuable for other purposes. Also, the read sample may be useful for assessing whether speakers are using avoidance strategies. Avoidance may be more likely in a spontaneous sample, resulting in a reduced severity score. Based on this assumption, Ward (2013) argued that the issue of avoidance may be overlooked in young children as they cannot provide a read sample. Here, however, the comparison between SSI scores obtained with reader and non-reader procedures did not differ, so there was no support for this view. As noted when the effects of sample-length was considered, additional types of samples may be helpful for other purposes, but a read and spontaneous sample (for older children) or spontaneous sample alone (for younger children) is sufficient for obtaining a stable SSI score. A further point supporting the use of read and/or spontaneous samples is that these were used to provide the standards for the older speakers in SSI (Howell, 2013).
Limitations

We have only addressed the minimum duration of a sample for obtaining an SSI-3 score. Although Riley suggests that longer samples give more reliable SSI scores, there has been no research that supports this. Examination of longer samples merits study, particularly in the light of the discrepancy between the effects of this variable in children (Sawyer & Yairi, 2006) and adults (Logan & Haj-Tas, 2007; Roberts et al., 2009). At the same time, some investigation of sample length and severity scores, not the %SS that previous authors have addressed, would be useful. Roberts et al. (2009) and Logan and Haj-Tas (2007) used different symptoms and the effects this has on %SS and SSI measures would be worth examining. Finally, further comparison between read and spontaneous procedures is required. In this case, any differences between minimum and more extensive sample lengths, age groups, severity versus %SS measures, and symptom type should be examined so that a complete picture can be presented of how stuttering samples vary with these factors.

Some observers consider that an adult speaker may stutter less during reading than during conversational speech. Hence, the different procedures (“reader” analysis versus “non-reader” analysis) might make more of a difference for adults. Currently, there is little evidence of differences between material types, particularly across age groups.

A further limitation is the difficulty if making objective assessments of the physical concomitant and duration measure of SSI-3. Assessing physical concomitants is subjective as it lacks operational definition about how can one assure reliability for judgments such as 1 and 2: “not noticeable unless looking for it” versus “barely noticeable to casual observer”, and 3, 4, 5 “distracting, very distracting and severe and painful looking. Also, these visual concomitants cannot be reassessed on audio recordings. If they are obtained in real time (e.g. in clinics), this does not allow clinicians (especially beginning clinicians) to double check, and does not allow for intra- or inter-rater reliability. A specific issue in this study (mentioned in the method) is that no reliability measures were obtained for physical concomitants as there was only one observer, and audio recordings were obtained. Similarly, the duration measure as made in clinics is rudimentary unless one has accurate acoustical analysis software, which many clinicians do not.

Conclusions

This study showed that Riley’s (1994) recommendation of minimum sample length for obtaining SSI-3 and SSI-4 scores was correct (i.e. 200-syllable-long samples give a stable score compared with shorter samples which do not). The change to recommending a sample of a minimum of 150 syllables in Riley (2009) appears to be a misguided as no further work on standardization was conducted; therefore, the earlier recommendation of 200-syllable long samples should be adhered to. The different procedures used when dealing with readers versus non-readers gave equivalent results when computed on the same group of speakers. This supports the equivalence of the two procedures used to obtain SSI-3 and SSI-4 scores.

Acknowledgements

The support by the Dominic Barker Trust, Ipswich, UK, is gratefully acknowledged. We thank the participants for their help.

Declaration of interest

The authors report no conflicts of interest. Parts of this work were supported by grants from the Dominic Barker Trust, Ipswich, UK, to Howell and Cavenagh.
References


