Scores on Riley's stuttering severity instrument versions three and four for samples of different length and for different types of speech material

Helena Todd[1], Avin Mirawdeli[1], Sarah Costelloe[2], Penny Kavanagh[2], Stephen Davis[1] and Peter Howell[1]

[1]Division of Psychology and Language Sciences
University College London
[2] University Campus Suffolk

Address correspondence to: Peter Howell, Division of Psychology and Language Sciences, University College London, Gower Street, London Wc1E 6BT.

Telephone: +44 207 679 7566 (direct line) +44 207 36 4276 (fax)

Email: p.howell@ucl.ac.uk

Abstract

Aims: Riley (1994; 2009) stated that 200-syllable long speech samples were sufficient to compute his severity estimates. This was checked as well as whether procedures he supplied for assessment of readers and non-readers produced equivalent scores.

Methods: Recordings of spontaneous speech samples from 23 young children (aged between 2;8 and 6;3) and 31 older children (aged between 10;0 and 14;7) were made. Riley's (1994; 2009) severity estimates were scored on extracts of different length. The older children provided spontaneous and read samples and these were scored for severity according to reader and non-reader procedures.

Results: Analysis of variance and correlation analyses supported the use of 200 syllable long samples for obtaining severity scores. Some effects were noted between age groups which suggested fatigue affected the younger group. There was no significant difference in SSI-3 scores for the older children when the reader and non-reader procedures were used.

Conclusions: Samples that are 200 syllables long are appropriate for obtaining Riley's severity scores. The procedural variants provide similar severity scores.

This is an Accepted Manuscript of an article published by Taylor & Francis in Clinical Linguistics and Phonetics on 18 June 2014 available online: https://www.tandfonline.com/doi/abs/10.3109/02699206.2014.926991?journalCode=iclp20 [10.3109/02699206.2014.926991]

Scores on Riley's stuttering severity instrument versions three and four for samples of different length and for different types of speech material

1. 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) described in Riley (1994) and Riley (2009) respectively. These instruments provide an estimate of stuttering severity in 200-syllable long speech recordings, and the instrument has been evaluated statistically. There are bespoke versions of these assessments for readers (standardized using spontaneous and read speech samples) and non-readers (standardized using spontaneous speech samples alone). No checks have been reported about whether 200 syllable long samples give stable estimates. Although there are separate forms for readers and non-readers, the relationship between the scores for the two forms does not appear to have been examined. These issues were assessed in the empirical work reported below.

All versions of SSI derive scores from measures made on a speech sample and observations of physical concomitants made at the time of the recording. Whilst assessment of speech is important, other features need to be examined for different purposes. Some other contemporary instruments that assess speech and other behavior that are widely-used in the stuttering field, are briefly reviewed. Then the details about SSI-3 are given, that show the role that SSI-3 and SSI-4 fulfill relative to these other instruments. The description of SSI-3 indicates why the features that were assessed in the current report needed examination.

1.1 Selected review of contemporary instruments used for assessing stuttering

Four recent tests, all of which have been evaluated statistically, are described. The tests are the Wright and Ayre's Stuttering Self-Rating Profile, WASSP (Wright & Ayre, 2000), Yaruss and Queasal's (2006), Overall Assessment of the Speaker's Experience of Stuttering which has adult and child forms (OASES and ACES respectively), Vanryckeghem and Brutten's (2007) adult and child communication attitude tests, (CAT and KiddyCAT) and Gillam, Logan and Pearson's (2009) test of childhood stuttering (TOCS). The tests collect different types of information that are used for different purposes and with different age groups (as indicated). In each case, the instrument is described, any important design features are noted, brief information about statistical evaluation is given, the main applications of the instrument are indicated and, in some cases, details about test evaluation are presented.

WASSP is a test for adults who stutter who are in speech and language therapy. WASSP records how a person who stutters perceives his or her stuttering. The test has five subscales and scores are provided for each. The subscales are: behaviors, thoughts, feelings about stuttering, avoidance, and disadvantage. Wright and Ayre (2000) reported that the test has been assessed for reliability and validity. The test is usually administered at the start and end of a block of therapy to establish any changes that occur. WASSP applies, then, mainly to clinical assessment and focuses on behavioral factors other than speech.

OASES evaluates the experience of stuttering from the perspective of the person who stutters (Yaruss & Queasal, 2006). Yaruss, Coleman and Quesal (2004) have developed a version of OASES for children who stutter, the *Assessment of the Child's Experience with Stuttering* (ACES). The ACES has been translated into other languages, for example German (Metten, Zückner & Rosenberger (2007). The ACES is a 100-item self-report test that measures the psychosocial effects of stuttering on everyday life. The aspects measured are: 1) general perspectives about stuttering; 2) affective, behavioral, and cognitive reactions to stuttering; 3) functional communication difficulties; and 4) impact of stuttering on the speaker's quality of life (Yaruss & Quesal, 2006). The adult version (OASES) has been validated for English and provides an outcome variable in the form of a severity index.

Yaruss indicated that the tables for adult severity can also be used for the children's version (Metten, 2005). Franic and Bothe (2008) criticized OASES because it includes external or environmental factors that are beyond the direct control of health care providers (e.g. marital status and potential earnings). There are also problems in the design of the OASES/ACES questionnaires: In ASES for example, answers are given on a scale of 1 to 5 with the more positive answers always on the low numbers. The low numbers always appear on the left side of the response form. In all tests, some respondents have response biases (e.g. may always choose the left-most response). Consequently, it cannot be established whether positive OASES scores are due to true positive feelings or such a response bias.

Vanryckeghem and Brutten (2007) provided instruments for the assessment of communication attitude in adults and children who stutter. The Communication Attitude Test (CAT) is used with older children (Brutten, 1984) and the more recent KiddyCAT (Vanryckeghem & Brutten, 2007) is for use with children aged between three and six years. The tests are similar in conception. The main difference between the KiddyCAT and CAT is that the former has instructions and task demands that are linguistically and cognitively appropriate for the three to six years age group. The KiddyCAT test has 12 questions to which the child responds "yes" or "no" according to "what they think about their talking". The questions are balanced for positive or negative attitudes, with six questions framed for a positive, and six for a negative attitude or experience. This allows the issue of response bias to be addressed (advice how to avoid this is also given). Norms are based on data from 63 children who did not stutter and 43 children who stuttered, all aged between three and six years. Reliability was high and Vanryckeghem and Brutten (2007) cited several studies that supported the validity of the test. A child has to score two or more standard deviations above the mean of the non-stuttering group to be designated as having the speech-associated beliefs of a child who stutters. Scores of more than two standard deviations occurred between groups of children who stutter and children who did not stutter for the normative data. Inspection of the normalization data show that negative attitudes about stuttering did not occur for all the participants who stuttered: Specifically, approximately 30% of children who stuttered had scores of two or fewer negative attitudes (out of a maximum of 12) on KiddyCAT which is a score that is less than two standard deviations away from those of the non-stuttering children. The low-scoring children must have shown speech symptoms indicative of stuttering, otherwise they would not have been classified as stuttering (they did not have a negative reaction when they were tested). These observations suggest that speech assessments are also needed as children can present with these symptoms, but no negative attitude. An important feature of KiddyCAT is that it gives precise advice about how to use the scored information when a child has negative attitudes because the 12 test items address different needs and strengths. Consequently, a child's individual answer profile can be used to identify what needs attending to during intervention.

TOCS provides clinicians and researchers with an assessment of fluency skills and stuttering-related behaviours in children aged between four and 12 years. Four speech fluency tasks are used to identify children who stutter and to rate the severity of their stuttering. These are: 1) Rapid Picture Naming, that is used to determine how fluently children produce single words when they are under time pressure; 2) Modelled Sentences, that addresses children's ability to speak fluently when they have to formulate and produce sentences with a given level of syntactic complexity; 3) Structured Conversation, that evaluates children's ability to speak fluently when in dialogue with the person conducting the test; 4) Narration, that assesses how fluent the children are when producing a monologue. TOCS has been standardized for children who speak English. As well as its use in research on stuttering, the authors state it can be employed to: 1) identify children who stutter; 2) determine the severity of a child's stuttering; and 3) document changes in a child's fluency functioning over time.

Gillam et al.'s (2009) TOCS is the most recent of the tests considered and, so it has been used less than the others. Consequently, independent reports that have evaluated it are not available at present. The TOCS is closest in its format and intended application to Riley's SSI tests insofar as it measures speech in a variety of circumstances to provide a measure of severity.

This brief and selective survey illustrates that clinical instruments play different roles in the assessment of stuttering, Instruments like WASSP, OASES/ACES and KiddyCAT/CAT focus on assessment of features other than speech. WASSP and OASES are for applications with older children and adults whilst ACES, KiddyCAT and TOCS are used with young children. Whilst all the instruments claim a role in clinical assessment, some are more focused on this issue (WASSP, OASES) than others (TOCS).

The current study examined the SSI-3 and SSI-4 instrument that mainly assesses speech symptoms, a feature it shares with TOCS. SSI-3 has been used to report details of stuttering participants in more than 350 publications and it has also been translated into other languages (e.g. Bakhtiar, Seifpanahi, Ansari, Ghanadzade & Packman, 2010). Whilst it is frequently emphasized that there is more to stuttering than symptoms in speech, this does not mean that speech measurements are unimportant. The observation that around a third of three to seven year olds present with no negative attitude to their speech (see discussion of KiddyCAT above) shows that assessments of other features are essential too. SSI-3/SSI-4 and TOCS are speech-based instruments that can complement attitude measures. Next the main features of SSI-3 and SSI-4 are reviewed.

1.2 Riley's SSI-3 and SSI-4 instruments

1.2.1 Important design features

Estimates of percent syllables stuttered (%SS), the average duration of the three longest stuttering events, and physical concomitants are required to produce an overall SSI-3 or SSI-4 score. Counts of stuttered syllables are needed to calculate %SS. The symptoms that are considered as stutters are identical in SSI-3 and SSI-4 and are described precisely in the SSI-3 manual (Riley, 1994) where the following events count 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). Among the symptoms not counted as stutters are whole-word repetitions. Other authors include whole word repetitions when assessing stuttering, particularly in work with young children (Yairi & Ambrose, 2005). Although there are pros (Yairi, Watkins, Ambrose, & Paden, 2001) and cons (Brocklehurst, in press) concerning whether whole-word repetitions should be included as stuttering symptoms, Riley's (1994, 2009) prescription has to be adhered to, otherwise the SSI-3 and SSI-4 scores are not correct and the norms do not apply.

Riley (1994) permitted some flexibility in the procedures that can be used to obtain speech samples. This flexibility was allowed so that assessments could be made in a wide variety of environments (e.g. clinics or research laboratories). Two-hundred syllable long recordings are used to compute SSI-3 and SSI-4 scores whatever procedure is used. Readers provide read and spontaneous samples whilst non-readers just provide a spontaneous sample. Reader and non-reader forms are scored using different tables. The availability of reader and non-reader forms allows SSI-3 and SSI-4 to be administered to children of all ages.

1.2.2 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. 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 reliability were deemed by Riley to be satisfactory.

Checks were made for criterion (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 Yaruss and Conture's (1992) 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 reached acceptable standards of validity.

Lewis's (1995) statistical evaluation of SSI-3 was less favorably than that of Riley: She concluded that a particular SSI-3 score could reflect a considerable range of stuttering behaviors and, in the light of this, suggested that the SSI-3 does not represent a reliable or valid measure.

1.2.3 The main applications of the instrument

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

1.2.4 Evaluation

Whilst it has been noted earlier that Riley (1994) designed 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 when this instrument is employed. If the suggested changes in procedure in SSI-4 are implemented, either the test needs restandardizing or it has to be shown empirically that the changes do not affect the scores obtained.

Even things that seem advisable on *a priori* grounds should not be changed when using the instrument unless the test is restandardized, For instance, whilst Riley (2009) suggests that video recordings may be substituted for audio recordings, which would allow objective assessment of physical concomitants, only audio recordings were used when the standards were established (Riley 1994 does not mention that video recordings were used when the normative data were analyzed). A video recording might affect %SS and duration estimates as well as physical concomitants, making comparison with other reports where the scores were obtained on audio records impossible. Another example concerns physical concomitants where questions have been raised about whether they should be included because 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 assessments cannot be added that were not employed when the test was normed either. Whilst 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 just with these types of speech, 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).

Riley (2009) introduced a major change in how stutters are counted in SSI-4 over that used in SSI-3. This was to use a software counter that automatically indexes syllables, stutters and their durations. 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 which provides an estimate of stutter durations. Using the counter is a difficult task that affects these measures and the SSI-4 scores that ensue relative to those obtained with the SSI-3 procedures (Jani, Huckvale & Howell, 2013). Again SSI-4 would need to be restandardized and its performance checked statistically if this procedural change was made.

Another point which ensures that the standards apply is that there is no flexibility in interpreting the guidelines about symptoms to count in the SSI-3 handbook. In research, different authors do and do not include whole word repetitions in the counts of %SS. However, whatever the pros and cons of the argument about whether whole-word repetitions are or are not stutters, Riley's (1994) procedure of ordinarily not counting whole-word repetitions as stutters has to be adopted when obtaining SSI-3 scores for the standards to apply.

1.3 Summary and research questions

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

Nevertheless, it is not apparent why the length of the sample was set at 200 syllables and whether this is suitable for obtaining a stable SSI-3 score. Authors who have been taking speech samples for other purposes have considered that longer samples are required (Yairi & Ambrose, 2005; Sawyer & Yairi, 2006). Of course, it does not necessarily follow if they are correct that a sample of speech 200 syllables in length is too short to obtain an acceptable SSI-3 score estimate, nor whether longer samples lead to similar estimates to those at 200 syllables. Consequently, one thing tested in the experiment reported below is whether a 200 syllable long sample provides a stable SSI-3 score. This involved estimating SSI-3 from extracts from the same sample which differed in length (longer and shorter than 200 syllables). Further checks were made about whether SSI-3 scores correlated across sample lengths and whether 50-syllable extracts drawn from different positions in a recording affected the estimates that were obtained. Tests were made on pre-school and older children. The two age groups use different forms of Riley's (1994) test and the effect of length needs to be checked for both.

Whilst the reader and non-reader forms should produce equivalent results if they measure the same underlying behavior, this does not appear to have been checked. A test was made as follows: The recording of the older children had SSI-3s scored two ways. First, the spontaneous and read samples were used to calculate an SSI-3 score in the normal way for such participants. In addition, the spontaneous samples alone were analyzed using the non-reader procedure. The two SSI-3 scores were used to test whether use of the non-reader procedure on spontaneous samples gave different results to the reader procedure on spontaneous and read samples Whether SSI-3 scores change when one or two types of material are used, also indirectly addresses whether more samples would give a better estimate of SSI-3. Further analyses were performed to determine whether scores obtained by reader and non-reader forms correlated. This test can only be performed with older children who can read as they can provide samples of spontaneous and read speech (nothing in the manuals precludes a non-reader form from being used with a child who can read).

2. Methods

2.1. Participants and recordings

There were 23 children in the younger age group. These were recruited from the caseload of a speech-language therapist based in Suffolk UK and had been diagnosed as stuttering. There were 18 males and five females and their ages ranged from 2;8 to 6;3 (mean age: 4;7 SD: 1;0). A spontaneous recording was available for these children. Physical concomitants were rated according to Riley (1994) at the time of the recording.

There were 31 children in the older age group. They attended a specialist clinic dealing with stuttering in London. There were 22 males and nine females and the age range was 10;0 to 14;7 (mean age: 13;0 SD: 1;1). Spontaneous and read speech samples were available for these children. Physical concomitants were obtained in a similar manner to those of the younger children. Appropriate reading material was used for 8-9 year olds, 10-11 year olds, 12-13 year olds and for speakers older than 13 years (Riley, 1994).

2.2 Pre-processing of speech samples and reliability assessment

All audio recordings were uploaded and coded using Speech Filing System (SFS) software (Huckvale, 2013) downloaded 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. Two hundred and fifty syllables were transcribed for each recording and, in the case of the older children, for both material types (200 syllables is the required length prescribed per recording in the SSI-3 manual). Reliability and validity reports for the transcription techniques are reported in Howell, Davis and Williams (2008).

2.2. 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 the samples were divided to provide shorter samples for the length analyses are detailed.

2.2.1 Administration of the SSI-3 to readers.

Procedure for obtaining frequency scores

Separate counts were made of all syllables spoken and those syllables that were stuttered

Procedures for counting total syllables

Syllable counts were obtained directly from the transcriptions in the SFS files. Non-word fillers such as “erm” were counted as words, and so were included in the total syllable count (consistent with examples given in the SSI-3 manual, Riley, 1994).

Procedures for counting stuttered syllables

The SSI-3 manual defines stutters as “repetitions or prolongations of sounds or syllables (including silent prolongations)” (Riley, 1994). This definition was followed here. Each stuttering episode was counted as one stutter. So, in the following example there is one stutter out of a total of five syllables:

a) A a a a a and I want that one
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.

b) And and and and I want that one
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 manual.

Duration Score

Duration of each stutter was obtained using the SFS display and replay facilities. SFS has a traveling cursor and calibrated timeline. The duration of a stuttering episode was 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 three longest durations (irrespective of sample) were selected, 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). Physical concomitants were coded as distracting sounds, facial grimaces, head movements and movements of the extremities. For each of these aspects, the rater gave a score from 0 (none) to 5 (severe and painful looking). The five ratings were then summed which allowed a maximum possible score of 20 for this component. This number gives the task score directly.

Total Overall Score and differences when non-reader scores were obtained

Only the raw SSI-3 scores were used throughout in this study. These were obtained for all sample lengths and positions (selected 50-syllable sub-sections within the sample as described below) 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 the three component elements obtained as described above (frequency, duration, and 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 observations on the spontaneous sample alone.

For the readers, SSI-3 scores were also calculated ignoring the read sample and estimating the scores as if these older children were non-readers. 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 transcribed section in each recording (for both younger and older children) was divided into sub-sections that included the first 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 sections 50 syllables in length. 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.

3. Results

3.1 Analyses

IBM SPSS Statistic 21 was used to conduct all analyses. Data form the two age groups were analyzed separately because the scoring procedures and factors in the analyses differed.

3.2 Younger children

The SSI-3 scores for different sample lengths were compared across sample lengths by ANOVA. The five SSI-3 estimates for the 50-syllable extracts from the overall transcribed section were also compared by ANOVA. Whether or not the scores differ across subsections of different length and across positions, they may or may not correlate (assessed using Pearson's *r*).

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 (250,200, 150, 100 or 50 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

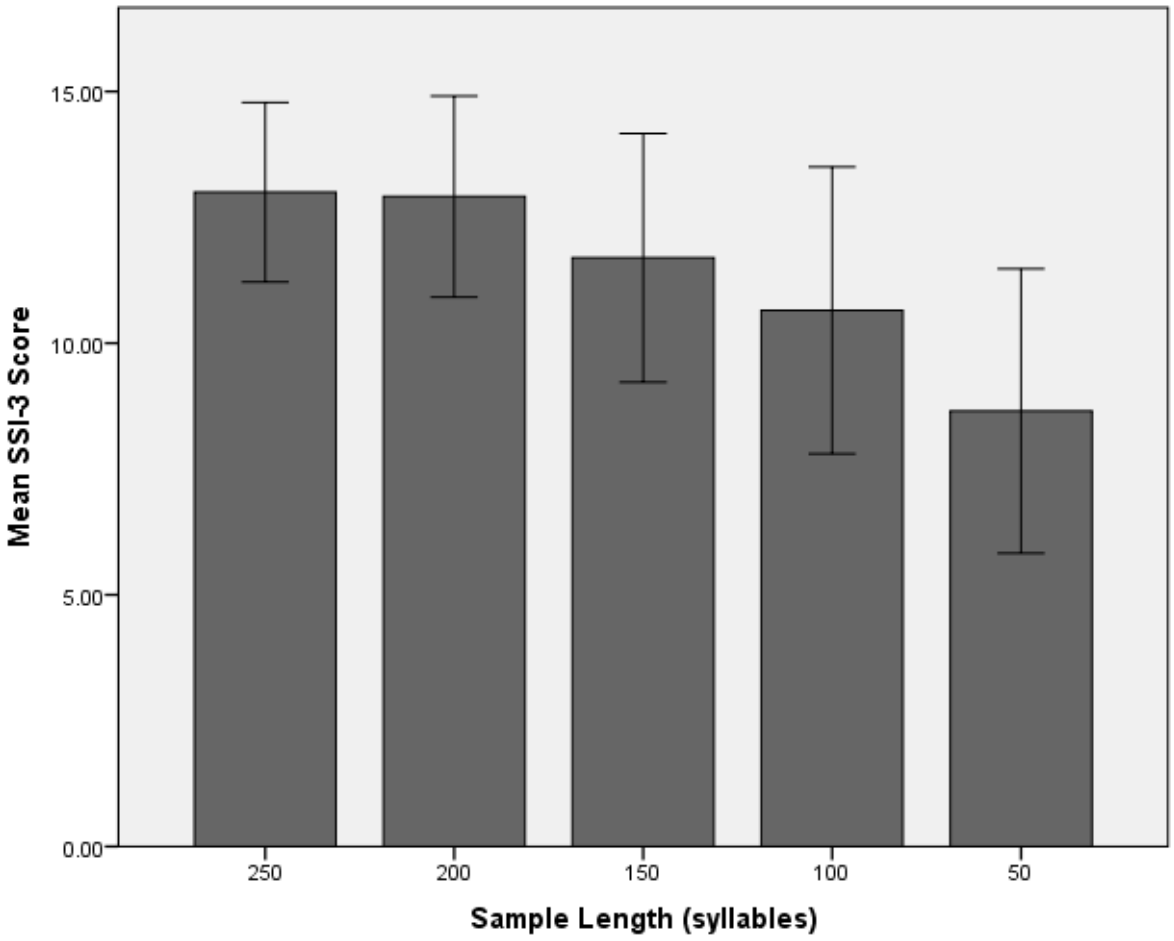

Figure .

Figure 1, and Tables 1 and 2 about here

Figure 1 shows that SSI-3 scores decreased as size of the sample decreased with the exception that the 250-syllable sample sections 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 .01 was used).

Pearson product moment correlations were computed on the SSI-3 scores for all pairwise comparisons of the five sample lengths. The coefficients are given in Table 2 and all were significant $p < 0.005$. The dropoff in the value of the correlation coefficients was most marked when 50 syllable samples were one member of the pair, although these were still significant.

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 a monotonic increase in mean SSI-3 scores as sample positions were later is apparent, though not significant.

Figure 2 and Table 3 about here

Pearson's *r* coefficients were computed for SSI-3 scores over different sample positions in a similar manner to those for sample size (here for all combinations of sample position for each 50 syllable sample). As seen in Table 3, some of the coefficients were not significant. There was a tendency for the SSI-3 scores to correlate with their near-neighbors (first with second, second with third, second with fourth, although not for third with fourth or fourth with fifth). The cross-time correlation would be consistent with the young participants getting progressively fatigued across the length of the sample.

3.2 Older children

Sample length and procedure type

An ANOVA with two within-groups factors was carried out using SSI-3 scores as the dependent variable. The factors were sample length (250, 200, 150, 100, and 50 syllables) and: procedure (scored according to reader procedure using 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). The means and 95% confidence intervals are shown for both procedures for all sample lengths in Figure 3.

Figure 3 and Tables 4 and 5 about here

Figure 3 shows that, again, there was a trend for SSI-3 scores to decrease as 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 4. The differences between the SSI-3 scores of 250 syllable-samples and 200-syllable samples were not significant for either the reader or the nonreader procedure. There was a significant difference between the SSI-3 scores for 200-syllable samples and 150-syllable samples (and lower) for the reader, but not the non-reader, procedure. In the nonreader procedure, the 200syllable SSI-3 scores differed from the 50 syllable SSI-3 scores.

The correlations across sample lengths were all significant for both reader and non-reader procedures and are summarized in Table 5).The value of the correlation coefficients were lower when shorter samples were compared (e.g. 50 and 100 syllables) when the non-reader procedures was used. As it is inadvisable to use these sample lengths anyway (they deviate from Riley's recommended 200 syllable samples), they would not be used in practice.

Sample position and procedure type

The ANOVA that compared SSI-3 scores across sample positions had the additional factor of procedure. Procedure (F 1, 30 = 0.026, p = 0.873, and position (F 3.046, 91.393 = 0.874, p = 0.459) were not 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.

The correlations between SSI-3 scores for all sample position pairs were significant for all comparisons and for both reader and non-reader procedures (summarized in Table 6). The fatigue pattern suggested for the younger children would not apply to the same extent with the older children which would explain this pattern of correlations.

Table 6 about here

Correlation of SSI-3 scores between reader and non-reader procedures at different sample lengths

If the reader and non-reader procedures measure the same thing they should correlate providing the sample size is adequate. To test this, the SSI-3 scores were correlated between samples of the same length obtained according to the two procedures. The coefficients are shown in Table 7. The coefficients are all impressively high (all above 0.82) which suggests that they all measure severity. The difference between the procedures is whether or not a reading was included when SSI-3 scores were calculated (only included in the reader procedure). Again it is necessary to be cautious, but including the read sample did not change the scores (otherwise the coefficients would have reduced dramatically).

Table 7 about here

---

4. Discussion

There were several main findings. Considering the younger children, first, 200-syllable long samples gave the same SSI-3 scores as samples that were 250 and 150 syllables in length. This suggests that SSI-3 scores are stable for 150-250 long samples. Riley (1994) recommended 200-syllable samples and this appears to be appropriate for obtaining SSI-3 estimates. Samples less than 150 syllables in length are not adequate as SSI-3 scores decreased significantly for sample of these lengths. Second, the correlations between SSI-3 scores for samples of different lengths were all significant. However, a drop off in coefficient values was noted for short samples (50 syllables in particular). The fact that the correlations were all significant showed that SSI-3 scores on different length segments were assessing related aspects of performance even though the shortest samples did not have statistically equivalent SSI-3 absolute values (ANOVA analysis) to those made on 200 syllable samples. Third, there was no significant difference between 50 syllable samples across utterance positions (five positions were examined). However, there was a monotonic, non significant, trend for SSI-3scores to increase from early to later positions. This did not occur with the older speakers. Finally, not all the correlation coefficients of the SSI-3 scores for 50 syllable extracts drawn from different positions were significant. A tendency was noted for the SSI-3 scores to correlate with their near neighbors (first with second, second with third, second with fourth). The correlation pattern across sample positions would be consistent with progressive fatigue across the length of the sample. This has implications for recommendations about what sample size is appropriate for use with young children that are discussed later. The restrictions on which samples correlated across sample positions did not apply to the older children.

The main findings with the older children had the extra factor, procedures in the ANOVAs. The first finding, based on the ANOVA on SSI-3 scores over samples of different lengths, was that 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-3 scores. Looking at the main effect of sample length (collapsed across the two procedures), 250 syllable and 200

syllable samples had the same SSI-3 scores, but all other lengths had significantly lower scores. This showed that 200-syllable samples provided stable SSI-3 scores.SSI-3 scores increased as sample length increased. This may be due to longer samples having a greater chance of a long stutter than for the shorter samples. Second, the correlations of SSI-3 scores across samples of different lengths were always significant for both procedures. The coefficients were lower for shorter sample lengths. The fact all correlations were significant suggests that the SSI-3 scores at all sample lengths were related, and the reduction in value of the coefficients showed that the longer samples may be more stable. Third, there were no differences in SSI-3 scores with respect to position of the sample or procedure for the 50 syllable extracts. Fourth, all the correlations across sample positions were significant (unlike with younger children) and this applied to both procedures. This indirectly supports the fatigue explanation offered with the younger children as these older children would not be expected to be susceptible in this way. Finally, SSI-3 scores obtained with reader and non-reader procedures showed very high correlations (above 0,82 in all cases). The difference between procedures is whether a read sample was included or not. This supports the validity of the non-reader procedure for SSI-3 scores (it gives scores that are not statistically distinguishable from the reader procedure made on the same speakers). An extreme implication of this result, not one we would draw, is that the non-reader procedure may be sufficient for older speakers.

This study was intended to establish whether sample length and reader/non-reader procedures affected SSI-3 scores. The implications the current results have for these issues is considered next. When sample length is discussed, some differences that were observed across age groups are considered.

4.1 Length of Speech Sample and age effects

The main questions about sample length were whether 200 syllable samples gave stable SSI-3 scores when compared with longer and shorter samples. SSI-3 scores were stable across sample of length 250-150 syllables with the younger children) and 250-200 syllables with the older children and, for the latter group, there were no differences when reader or non-reader procedures were used. When sample sizes were shorter than those indicated, SSI-3 scores were lower, which showed these short samples do not produce stable SSI-3 scores (i.e. sample size was not sufficient for obtaining the score). Overall, these findings support the use of 200 syllable long samples for obtaining an SSI-3 score as Riley (1994) advocated and this is an appropriate sample length for both age groups.

The conclusion that 200 syllable long sample is adequate for making an SSI-3 score does not necessarily mean 200 syllable long samples are appropriate for other purposes. Sample size is an issue that Yairi and Ambrose (2005) and Sawyer and Yairi (2006) have considered to identify and follow the course of stuttering in children. They advocate using longer samples for this purpose. Also, they critique authors who have used short samples in clinical studies, “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 2,044 words, whereas Schwartz and Conture (1988) used 85 to 650 words. Many studies in the past two decades were based on samples of 300 to 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.” In this they 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.

Findings about differences between the age groups have potential relevance for the question of what sample size is appropriate when younger children are examined. The

younger age group showed more variable SSI-3 scores across sample positions and the pattern of correlations tended to be significant only for near-neighbor samples (all correlations were significant for the older group of children). The significant correlations suggest a successive change over the timecourse of the sample. The change could be progressive fatigue. The fact that this is specific to the younger group would also offer incidental support for this interpretation. This interpretation needs further examination, but it suggests some caution should be expressed about advocating use of longer samples until some consensus is reached.

An SSI-3 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-3 are not the only feature of stuttering that needs measuring. A further point, often overlooked, is that SSI-3 is not simply a %SS measure (it includes duration and physical concomitant scores as well).

Samples of different length may be needed for different purposes, It is not known at present whether this is the case or not since studies using SSI-3 use different procedures to Yairi's group (Sawyer & Yairi, 2006) who advocate using longer samples. Whilst we have emphasized the need to appreciate the different roles SSI-3 has from those intended by Yairi and Ambrose (2005), there is, nevertheless certain points of agreement. The main ones are use of objective speech-based measures to characterize stuttering patterns and inclusion of a measure of duration of stutters. Further research is called for that compare different ways of making stuttering symptom counts (Howell & Lu, 2013; Roberts, 2011).

4.2 Procedure

The non-reader assessments have to be made on spontaneous samples alone, whereas the reader assessments are made on spontaneous and read samples. The non-reader provision is usually used with young children, but there is no prohibition on its use with non-reading older people who stutter. This allowed the validation of SSI-3 with older children who stutter who could read who were scored according to the reader and non-reader forms with the requisite materials. There was no significant difference between SSI-3 scores made the two ways. Also, the correlations of SSI-3 scores obtained with reader versus non-reader procedures were very high (always greater than 0.82) which suggests they are measuring the same underlying dimension of performance. This suggests that the two procedures for making SSI-3 scores are inter-translatable. The results may also suggest that a spontaneous sample scored according to the non-reader form would be sufficient to obtain an SSI-3 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. Apart from these situations, the reader form with both types of material should be used.

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). Also, the reading sample may be useful since avoidance may be present more in the spontaneous sample, resulting in a reduced severity score. Ward (2013) argued that the issue of avoidance may be overlooked in young children as they cannot provide a read sample. The comparison between SSI-3 scores obtained with reader and non-reader procedures does not support the view that materials provide different information. Once again the additional samples may be helpful for other purposes, but a read and spontaneous sample (older children) or spontaneous sample alone (younger children) are sufficient for obtaining an SSI-3 score. A further point supporting use of read and/or spontaneous samples is that these were used to provide the standards in SSI-3 and SSI-4 (Howell, 2013).

4.3 Caveats

Severity scores differed markedly between age groups with the younger children having lower scores. The younger children showed high rates of whole-word repetitions which were not included here in SSI-3 scores. They did influence the therapist's diagnosis of the child as a person who stutters. This raises the important question whether whole-word repetitions should or should not be considered as stutters (this was not addressed in the current study).

SSI-3 is more or less limited to speech symptoms (%SS and duration) with a measure of physical concomitants. The current study adds to the documentation of important properties of the SSI-3 instrument. However, it is recognized that wider assessments are needed for clinical purposes (see the discussion of other instruments in the introduction). We would emphasize however, that symptom measures as provided by SSI-3 and TOCS, will always be needed as the point made earlier that almost a third of children tested with Kiddycat do not have negative attitudes shows. SSI-3 could be improved. In particular, examination of physical concomitants is needed and, allied to this some improvement in procedures for measuring them is required. Before this is attempted, a better consensus is needed about whole-word repetitions and their role. As these changes are major and likely to require restandardization, many of the changes discussed that cannot be used in procedures for obtaining SSI-3 and SSI-4 scores should be evaluated and included. This includes shift to video, wider range of samples etc**.**

4.4 Conclusions

This study showed that Riley's (1994, 2009) recommendation of sample length for making SSI-3 and SSI-4 scores was correct (200 syllable long samples are sufficient). 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.

Acknowledgement

Parts of this work were supported by grants from the Dominic Barker Trust to Howell and Kavanagh, Ipswich UK and their support is gratefully acknowledged. We thank the participants and Dr Susanne Cook for their help.

References

Bakhtiar, M., Seifpanahi, S., Ansari, H., Ghanadzade, M., & Packman, A. (2010). Investigation of the reliability of the SSI-3 for preschool Persian-speaking children who stutter. *Journal of Fluency Disorders*, *35*(2), 87–91.

Brocklehurst, P.H., (2013) Stuttering prevalence, incidence and recovery rates *Journal of Fluency Disorders*, http://dx.doi.org/10.1016/j.jfludis.2013.01.002.

Brutten, G. (1984). *The Communication Attitude Test.* Unpublished manuscript, Southern Illinois University, Carbondale.

Conture, E. G., & Kelly, E. M. (1991). Young Stutterers' Nonspeech Behaviors During Stuttering. *Journal of Speech and Hearing Research*, *34*, 1041.

Franic, D. M., & Bothe, A. K. (2008). Psychometric evaluation of condition-specific instruments used to assess health-related, quality of life, attitudes, and related constructs in stuttering. *American Journal of Speech-Language Pathology*, *17*, 60–80.

Gillam, R., Logan, K., & Pearson, N. (2009). *TOCS: Test of Childhood Stuttering*. Austin, TX: PRO-ED.

Howell, P. (2013). Screening school-aged children for risk of stuttering. *Journal of Fluency Disorders*, *38*, 102–123.

Howell, P., Davis, S., & Williams, R. (2008). Late childhood stuttering. *Journal of Speech, Language and Hearing Research*, *51*, 669-687.

Howell, P., & Lu, C. (2013). Assessing risk for stuttering in children. *Journal of Fluency Disorders*, *38*, 63–65.

Huckvale, M. (2013) *Speech Filing System Windows version 1.75* [computer software]. University College London

Jani, L., Huckvale, M., & Howell, P. (2013). Procedures used for assessment of stuttering frequency and stuttering duration. *Clinical Linguistics & Phonetics*, *27*, 853–861.

Johnson, W., Boehmler, R., Dahlstrom, W., Darley, F., Goodstein, L., Kools, J., et al. (1959). *The Onset of Stuttering: Research Findings and Implications*. Minneapolis: University of Minnesota Press.

Lewis, K. E. (1995). Do SSI-3 scores adequately reflect observations of stuttering behaviors? *American Journal of Speech-Language Pathology*, *4*, 46-59.

Metten, C. (2005) *Evaluation einer Stotterintensivtherapie*, Diplom unveröffentlichte Diplomarbeit, Rheinisch Westfälisch Technische Hochschule Aachen.

Metten, C., Zückner, H., & Rosenberger, S. (2007) Evaluation einer Stotterintensivtherapie mit Kindern und Jugendlichen (Evaluation of an Intensive Stuttering Treatment for Children and Adolescents). *Sprache, Stimme, Gehör*, *31*, 1-10.

Onslow, M., Costa, L., & Rue, S. (1990). Direct Early Intervention with Stuttering - Some Preliminary Data. *Journal of Speech and Hearing Disorders*, *55*, 405–416.

Riley, G. (1994). *The Stuttering Severity Instrument for Adults and Children (SSI-3)* (3rd ed.). Austin, TX: PRO-ED.

Riley, G. (2009). *The Stuttering Severity Instrument for Adults and Children (SSI-4)* (4th ed.). Austin, TX: PRO-ED.

Roberts, P. (2011). Methodology matters. In P. Howell & J. van Borsel (Eds.), *Multilingual Aspects of Fluency Disorders*. Bristol: Multilingual Matters.

Sawyer, J., & Yairi, E. (2006). The effect of sample size on the assessment of stuttering severity. *American Journal of Speech-Language Pathology*, *15*, 36–44.

Schwartz, H. D., & Conture, E. G. (1988). Subgrouping Young Stutterers: Preliminary Behavioral Observations. *Journal of Speech and Hearing Research*, *31*, 62-71.

Vanryckeghem, M., & Brutten, G. (2007). *Communication attitude test for preschool and kindergarten children who stutter (KiddyCAT)*. San Diego, CA: Plural Publishing.

Ward, D. (2013). Risk factors and stuttering: Evaluating the evidence for clinicians. *Journal of Fluency Disorders*, *38*, 134–140.

Wright, L., & Ayre, A. (2000). *WASSP: The Wright and Ayre Stuttering Self-Rating Profile*. Bicester: Winslow.

Yairi, E., & Ambrose, N. (2005). *Early childhood stuttering : for clinicians by clinicians*. Austin, TX: PRO-ED.

Yairi, E., Ambrose, N., Watkins, R., & Paden, E. (2001). What is stuttering? *Journal of Speech, Language, and Hearing Research*, *44*, 585-592.

Yaruss, J. S. (1997). Clinical implications of situational variability in preschool children who stutter. *Journal of Fluency Disorders*, *22*, 187–203.

Yaruss, J. S., Coleman, C. E. & Quesal, R. W. (2006). Assessment of the Child's Experience of Stuttering (ACES), Miami, FL: Poster presented at the annual convention of the American Speech-Language-Hearing Association.

Yaruss, J. S., & Conture, E. G. (1992). Relationship between mother-child speaking rates in adjacent fluent utterances. Poster presented at the annual convention of the American Speech-Language-Hearing Association.

Yaruss, J. S., & Quesal, R. W. (2006). Overall Assessment of the Speaker's Experience of Stuttering (OASES): Documenting multiple outcomes in stuttering treatment. *Journal of Fluency Disorders*, *31*, 90–115.

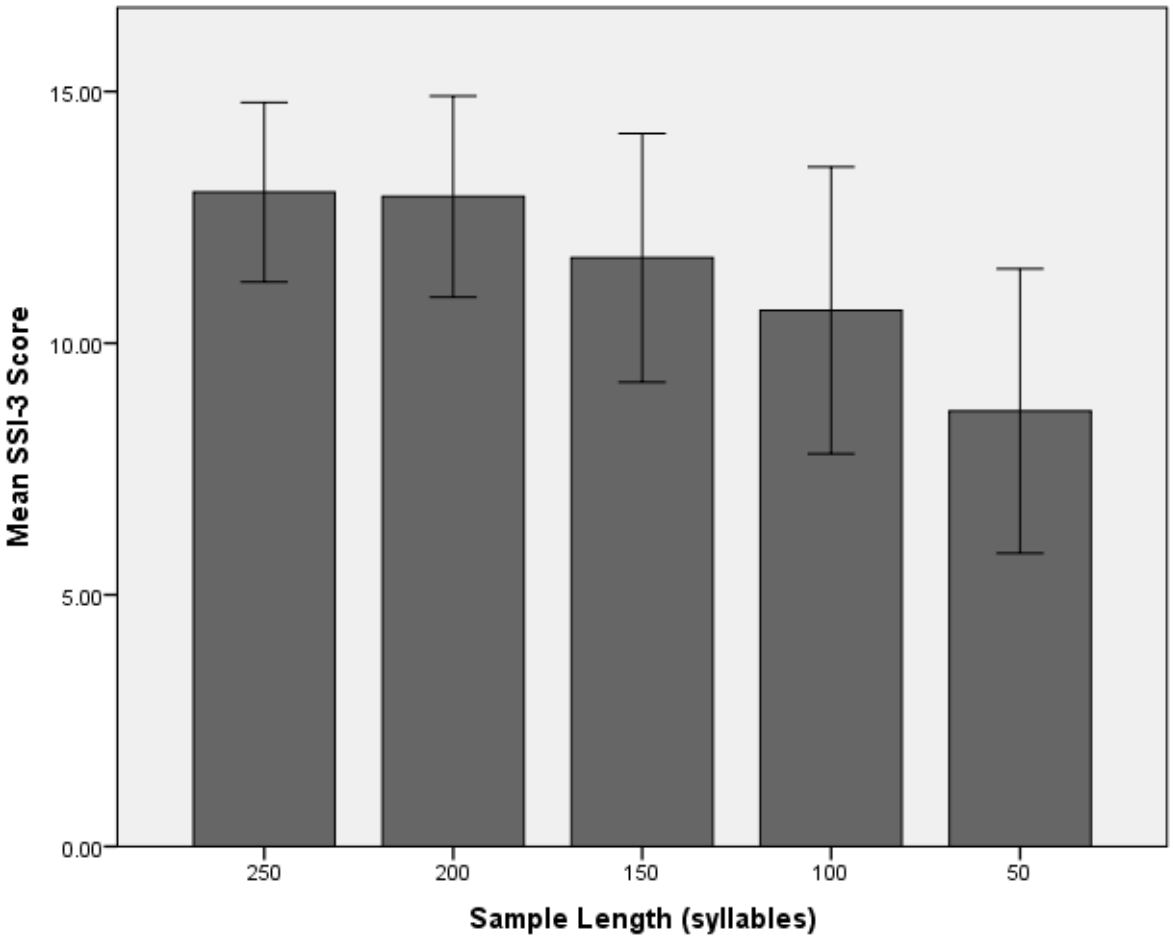

Figure 1: Younger group: means of SSI-3 scores across sample lengths

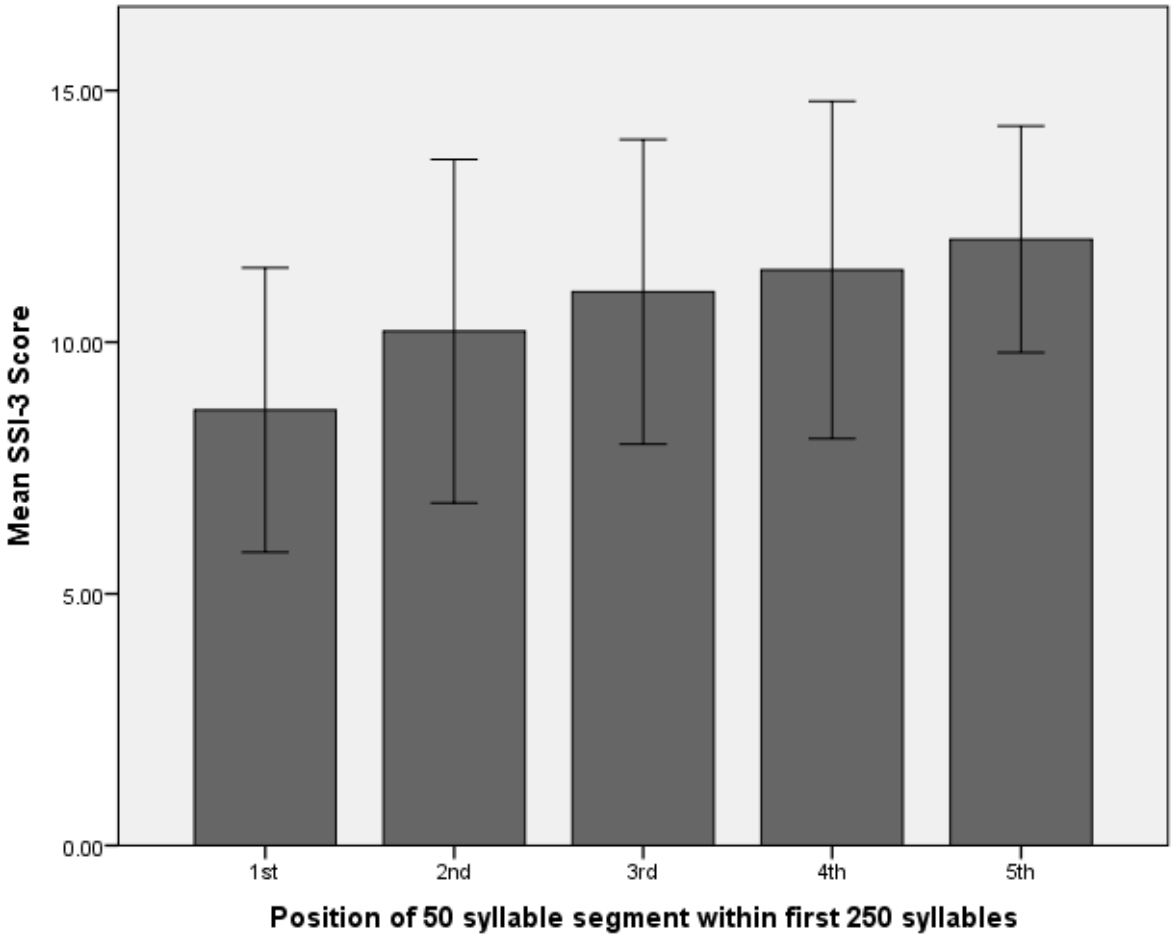

Figure 2. Younger group: means of SSI-3 scores for 50 syllable segments from different locations

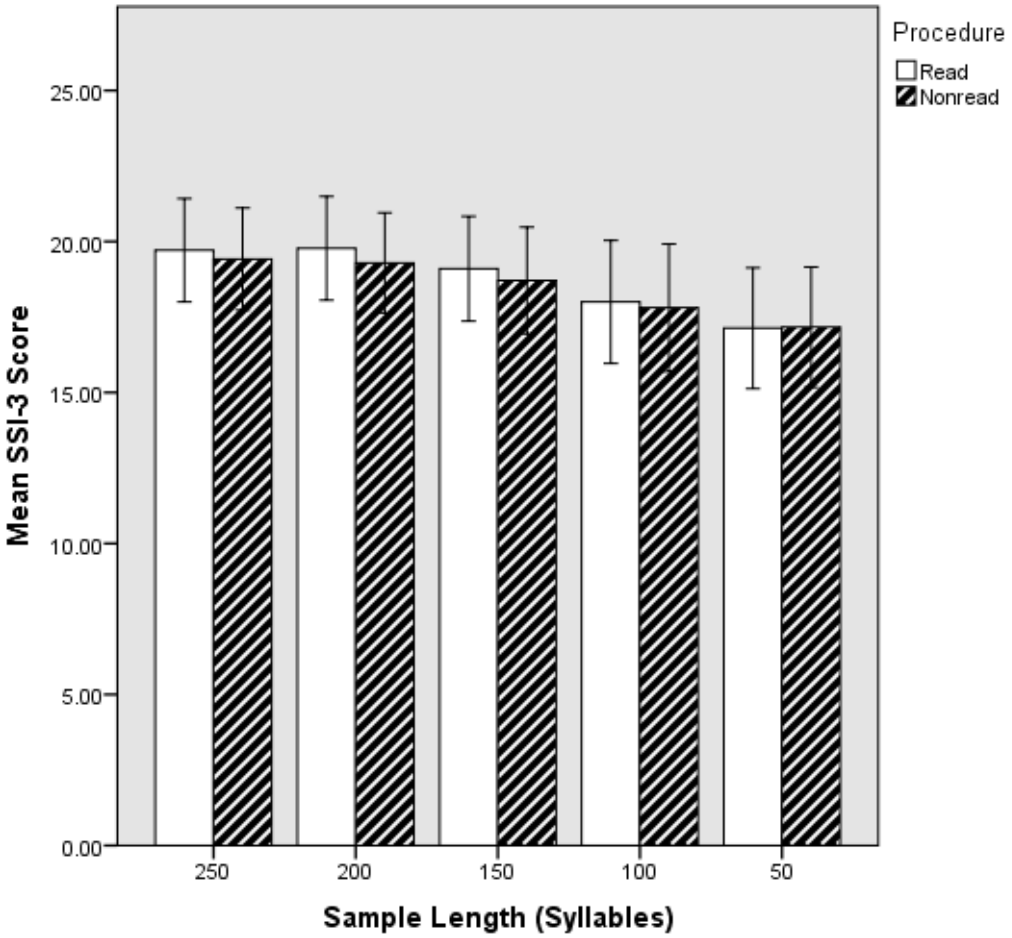

Figure 3: Older group: means of SSI-3 scores across sample lengths Reader and non-reader

Table 1 *t*-tests for comparisons of the younger children's SSI-3 scores for the 200 syllable long sample with other sample lengths. Degrees of freedom were 22 for all tests. * indicates significant at $p < 0.01$.

Comparison	t	Sig
250 and 200	0.20	0.847
200 and 150	1.95	0.064
200 and 100	2.80	0.010*
200 and 50	3.65	0.001*

Table 2 Pearson's *r* of the younger children's SSI-3 scores when samples of different length were compared. Degrees of freedom were 21 in all cases. All were significant $p < 0.005$

	250	**200**	**150**	**100**	**50**
250		.887	.841	.880	.584
200			.852	.817	.541
150				.895	.591
100					.704
50					

Table 3 Pearson *r* for the younger children's SSI-3 scores for 50 syllable segments from differing locations in the sample. Degrees of freedom were 21 in all cases. * indicates significant at $p < 0.05$ and ** significant at $p < 0.005$

	1st	**2nd**	**3rd**	**4th**	**5th**
1st		.451*	.288	.215	.046
2nd			.456*	.442*	.160
3rd				.265	.022
4th					.128
5th					

Table 4 *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 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. * indicates significant at $p < 0.01$

Reader procedure

Comparison	**t**	**Sig**
250 and 200	-0.34	0.738
200 and 150	3.09	.004*
200 and 100	3.79	0.001*
200 and 50	4.88	p<0.001*

Non-reader procedure

Comparison	**t=30**	**Sig**
250 and 200	0.57	0.572
200 and 150	1.43	0.163
200 and 100	2.50	0.018
200 and 50	3.04	0.005*

Table 5 Pearson's *r* of the older children's SSI-3 scores when samples of different length were compared. 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. All were significant $p < 0.001$.

Reader procedure

	250	**200**	**150**	**100**	**50**
250		.974	.943	.842	.797
200			.966	.885	.834
150				.897	.817
100					.935
50					

Non-reader procedure

	250	**200**	**150**	**100**	**50**
250		.962	.847	.787	.668
200			.884	.820	.708
150				.863	.723
100					.547
50					

Table 6 Pearson *r* for the older children's SSI-3 scores for 50 syllable segments from differing locations in the sample. 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. * indicates significant at $p < 0.05$ and ** significant at $p < 0.005$

	1st	**2nd**	**3rd**	**4th**	**5th**
1st		.712	.595	.657	.629
2nd			.645	.694	.515*
3rd				.664	.708
4th					.658
5th					

:

	1st	**2nd**	**3rd**	**4th**	**5th**
1st		.528**	.406*	.244	.694**
2nd			.582**	.201	.315*
3rd				.509**	.554**
4th					.421*
5th					

Table 7 Pearson *r* for the older children's SSI-3 scores for comparison of reader and non-reader procedures on the same sized samples. Degrees of freedom were 29 in all cases. All coefficients were significant $p < 0.001$

	Readers					
Non readers		250	200	150	100	50
	250	0.866				
	200		0.853			
	150			0.840		
	100				0.887	
	50					0.821