Published in final edited form as:

Clin Linguist Phon. 2011 May ; 25(5): 368–378. doi:10.3109/02699206.2010.538955.

Comparison of alternative methods for obtaining severity scores of the speech of people who stutter

Peter Howell, Tajana Soukup-Ascencao, Stephen Davis, and Sarah Rusbridge

Division of Psychology and Language Sciences and Centre for Human Communications, University College London

Abstract

Purpose—Riley’s Stuttering Severity Instrument (SSI) is widely-used. The manual allows SSI assessments to be made in different ways (e.g. from digital recordings or whilst listening to speech live). Digital recordings allow segments to be selected and listened to while the entire recording has to be judged when listened to live. Comparison was made between expert judges when they used these digital and live procedures to establish whether one method was more sensitive and reliable than the other.

Method—Five expert judges assessed eight speakers four times each in two judgment conditions (digital versus live). The eight speakers were chosen so that they spanned a wide range of stuttering severity. SSI version 3 (SSI-3) estimates were obtained on all occasions.

Results—An ANOVA showed a three-way interaction between sessions, speakers and condition that indicated that digital and live judgments varied across speakers and across sessions.

Conclusion—The predictions that were upheld were: 1) SSI-3 scores made from digital segments are more sensitive than SSI-3 scores made on the entire live signal; 2) Digital and live judgments vary with respect to speaker’s stuttering severity and across test sessions.

Keywords

Developmental stuttering; severity estimates; stuttering severity instrument; persistent stuttering; recovered stuttering

Riley’s Stuttering Severity Instrument is used for the objective assessment of severity and in other examinations of stuttered speech. Current versions of the instrument measure three parameters: 1) stuttering frequency; 2) duration of selected stutters; 3) observed physical concomitants. Riley’s instrument is considered to provide a more complete analysis than does percentage of stuttered syllables alone which is a measure of stuttering frequency alone (Miller & Guitar, 2010). The third revision, SSI-3 has been used extensively to date (see Riley, 1994 for references to work using this version of the instrument). Recently a fourth revision was published, SSI-4, that has not been used as extensively as SSI-3 because of the short time it has been available (Riley, 2009).

The differences between SSI-3 and SSI-4 are: 1) The fourth revision includes a computer program that automates the assessment of stuttering severity (this was not available in SSI-3). However, to date the program has not been assessed for reliability and validity. Also, results with the program have not been compared against the methods for obtaining

Address Correspondence to: Peter Howell, Division of Psychology and Language Sciences, University College London, Gower Street, London WC1E 6BT, England. Telephone Number: +44 (0) 20 7679 7566, Fax number +44 (0) 20 7436 4276, p.howell@ucl.ac.uk.

Declaration of interest. The authors report no conflicts of interest.

stuttering severity recommended in SSI-3. The latter methods were used to obtain the norms and new norms would be required if the program produces different results; 2) SSI-4 includes a self-report instrument. This is not used for calculating the SSI-4 scores (it provides ancillary information); 3) The SSI-4 manual advocates obtaining beyond-clinic speaking samples and samples obtained over the telephone. The norms that were obtained for SSI-3 were derived from a reading of a set passage and spontaneous recordings. Obtaining a range of assessments in different conditions is a commendable practice. However, as the SSI-3 norms do not apply if the additional assessments are included (Riley, 2009, p.12), they are best left out when SSI-4 scores are obtained so that the norms can be used. Overall, the additional features in SSI-4 are either not tested for reliability and validity, not necessary for severity assessments or would preclude use of the published norms. For these reasons, the current study dispenses with these additional features that make the assessments equivalent to SSI-3. To emphasize this, the assessments are referred to as SSI-3 although assessments made in this way are permitted in SSI-4.

SSI-3 and all other versions of the instrument have several features that illustrate good practices when making severity estimates. These are, in addition to requiring recordings to be made in different speaking situations: 1) a clear and easy to use scoring protocol; 2) reliability and validity have been examined statistically; and 3) there are published standard scores for English for the restricted set of speaking samples.

Riley (1994) indicated that SSI-3 is suitable for a number of clinical purposes. The purposes that Riley (1994) mentioned and illustrations where SSI-3 has been used are: 1) SSI-3 has been used as part of diagnostic evaluation. Severity estimates may be particularly important in children when trying to distinguish mild cases of stuttering from cases where a child shows normal nonfluencies. One peer-reviewed articles have also reported SSI-3 estimates for participants who stutter and for controls (Davis, Shisca & Howell, 2007); 2) SSI-3 can assist in tracking changes in severity during and following treatments. Miller and Guitar (2009), for example, used a form of SSI-3 in a study into long-term outcomes of the Lidcombe program for early stuttering; 3) SSI-3 can be used to describe the severity distribution in experimental groups that include people who stutter. An example of this is Howell, Davis and Williams (2008) who used SSI-3 to characterize children who stuttered aged 8-teenage; 4) SSI-3 can be used to validate other stuttering measures. Howell, Davis and Williams (2009) used SSI-3 to validate child, parent and researcher assessments that were subsequently used for classifying participants aged between eight years and teenage as persistent or recovered.

All versions of SSI allow for some flexibility in how assessments are made in order to permit them to be used in a variety of environments ranging from the clinic to well-equipped research laboratories. For example, duration of stutters is permitted to be measured with a stopwatch or from speech files stored on computer, and physical concomitants can be based on direct observations noted down at the time of assessment or extracted subsequently from videos made when the assessment was done. Also, stuttering frequency measures can be made from audio (as in the study reported below) or video recordings. An advantage of using audio files for judgment is that they are easily archived on computer without excessive demands on storage that occur with video data. Whilst flexibility allows SSI assessments to be made in a wide range of situations, the optional assessment procedures this entails raises questions about which methods are preferable and whether or not the various methods permitted in the manual give similar results.

To date, no comparison has been made to establish whether different procedures give different severity scores. Such information would be useful for making recommendation about which procedures are most sensitive (identify most stuttering) and reliable (produce

similar results when judgments are made on different sessions) as it is possible that severity acts as a risk factor for predicting the onset and course of stuttering (Howell, 2010; Riley, 1994). The work of Davis et al., (2007), Howell et al. (2008) and Howell et al. (2009) made SSI-3 estimates on digital recordings of audio speech which were segmented and listened to repeatedly. This digital method was used as one of the procedures in the current study.

The below work compared this procedure (referred to as digital) with a form closer to the simpler ones allowed in the SSI-3 and SSI-4 manuals (nearer to what might be done in a clinic). These manuals allow speech to be assessed whilst it is listened to continuously without a break (referred to here as live). Live procedures require a recording as at least two passes need to be made on the speech (one to assess number of syllables and one to assess stutters and their duration). Audio recordings are sufficient for digital and live assessments of both these components of the severity indices and such recordings were employed here. Care was taken to ensure that the digital and live procedures involved the same exposure to the speech being judged. Keeping procedures comparable is part of good design, but it restricts the way judgments are made (as indicated earlier, a wide range of procedures is allowed by Riley, 1994). The research questions that are examined here are about differences between judgment procedures.

Live procedures for making severity assessments, might not be easy for everyone to perform. One reason for this is because Riley (1994) indicated that whole-word repetitions should be excluded in counts of %SS when obtaining severity estimates. Since some users are practiced in assessments that include whole-word repetitions as stutters, their results need to be thoroughly checked. This can be done when digital transcriptions are used as a permanent recording is available against which the symptoms identified can be checked. Several research laboratories already use digital recordings which they employ to make their SSI-3 assessments (Howell, Davis & Bartrip, 2009). If it is shown that digital methods provide better reliability and sensitivity than other methods, assessments using them would be the ones to choose for research work and it might be advisable to recommend them for more general clinical use in later revisions of the instrument. However, before that is done, the hypothesized differences between digital and live methods need documenting. The question addressed in the below study concerned whether SSI-3 scores obtained from digital recordings are more sensitive than SSI-3 scores made live and whether sensitivity varies across the speakers that are judged and across repeated judgment occasions (sessions). Higher sensitivity with respect to digital recordings would be evident if more stutters are detected with the digital method than the live method.

The following study compared difference in SSI-3 scores when the whole sample was listened to versus when segments were listened to. The procedure where the whole sample was listened to is referred to as “live” as decisions would need to be made like this when there is no permanent recording. The procedure where segments are listened to as “digital” as they presuppose some form of permanent record (in the current case, digital recordings). Attention was paid to ensure the digital and live assessments were obtained as similarly as possible (e.g. same length of exposure to the test material). The assessors were highly trained. Comparisons were made between the digital and live estimates to address: 1) whether more stuttering was detected during the digital SSI-3 assessments than the live ones; 2) whether SSI-3 scores varied across speakers with different severities (i.e. whether severity scores depended on how much speakers stuttered); 3) whether SSI-3 scores varied across occasions (sessions).

Method

Participants

Five judges performed the experiment. All were native speakers of British English who were trained listeners and each had at least 100 hours experience of making transcriptions.

Material from speakers who stutter

Speakers with a range of severities were included so that the way this dimension affected judgments could be examined. The speech material employed was selected from Howell et al. (2010). Howell et al.'s material was from children who stuttered who were recorded in three age ranges between eight and teenage (8-10, 10-12 and 12 plus-15 years 7 months). The samples for the 12 plus age range were employed as these have the highest range of SSI-3's. This is because some of the children had recovered and were expected to have low SSI-3 scores whilst others had persisted to this age and were expected to have high SSI-3 scores (the range of SSI-3 scores was lower at the younger test ages). The other advantage in using this material is that although some of the speakers were expected to have low SSI-3 scores, they were rigorously assessed at inclusion to ensure that they were stuttering (see Howell et al. 2010 for details).

Eight children were selected based on the SSI-3 scores at age 12 plus given in Howell et al (2010). Information about individual children and summary statistics are given in Table 1. The children selected covered a wide range of absolute severities and were roughly equally-spaced across the range of severities. Although not necessary for the current study, five were classed as persistent and three as recovered at teenage and all had a documented history of stuttering from age eight. Two of the speakers were female (P1 and R4).

Speech samples

The audio test recordings from the children at age 12 plus included a reading of an SSI-3 text appropriate for 12-13 year olds and a recording of a monolog for all speakers (16 files in total, the readings were complete renditions of the text and the monologs were at least two minutes and a minimum of 200 words). The minimum requirements were set to adhere to Riley's (1994) specifications as the study is about this instrument. In fact, all spontaneous speech samples were significantly longer than 200 syllables. Two hundred syllables is set as a minimum requirement in SSI-3 presumably because they count actual syllables and severe stutterers can have difficulty producing utterances with this number of intended syllables. The speech samples were recorded on a Sony DAT recorder using a Sennheiser K6 microphone. They were transferred to a PC and uploaded for analysis into Speech Filing System software (SFS, freeware available at <http://www.phon.ucl.ac.uk/resource/sfs/>). The first 200 words in spontaneous monologs and all words in the SSI-3 read text were examined (the texts are each approximately 200 words in length). Physical concomitant scores were obtained for the spontaneous and read material at the time the recordings were made and these were used to calculate all SSI-3 scores.

Components of the Stuttering Severity Instrument Version 3 (SSI-3) estimates

The SSI-3 scores obtained in this study do not: 1) use the computer instrument for assessing stuttering severity that needs testing; 2) employ the extraneous self-report instrument; and 3) use extra-clinical samples and telephone samples that prohibit use of the norms.

The SSI-3 assessments were based on: 1) percentage of stuttered syllables (frequency score); 2) average stutter duration of the three longest stutters in a reading (duration score); 3) a physical concomitants assessment (e.g. distracting sounds, facial grimaces, etc.). A comment about changes to the duration measure in different versions of the SSI is required: In the

original version of the SSI (Riley, 1972), Riley required the duration of the three longest *blocks* (silent prolongations of an articulatory posture) to be measured, but subsequently shifted to using the duration of the three longest *stutters* (Riley, 1994; 2009). The latter practice is followed here. However, it should be noted that moving from blocks to stutters may have different impacts on severity measures as blocks occurring within whole-word repetitions could have been included in the earlier formulation, whereas whole-word repetitions and silent blocks within these sequences are not included using the later formulation. The way the judges assessed the three components in both judgment conditions is described next.

Frequency of stuttering

Stuttering behaviors were defined as “repetitions or prolongations of sounds or syllables (including silent prolongations)” (Riley, 1994, p.4). Riley (1994) also noted which behaviors are not included within the definition of stuttering: “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 judges were instructed to follow these guidelines to determine what were considered as stutters in both judgment conditions.

Two quantities are needed to obtain the percentage of stuttered syllables (%SS) for both judgment procedures: 1) counts of all syllables spoken; and 2) counts of syllables that are stuttered. Riley advocated obtaining these separately and this was done for both judgment conditions.

Duration

The duration score is the time, in seconds, of the three longest stuttering events in the sample. For each judgment condition, once the average duration was obtained it was converted to a scale score using the reader’s Table in the computation form (Riley, 1994, p. 10).

Physical Concomitants

The same male made the physical concomitant scores for all of the speakers at the time the recordings were obtained. These were obtained for reading and spontaneous speech conditions. The male who made the judgments was not involved in the other judgments in this study. Each physical concomitant score was obtained in the way Riley (1994) recommended (summarized briefly below).

Four aspects of physical concomitants were assessed. Riley’s descriptions of each of these are as follows (Riley, 1994, p.11; Riley 2009, p.9).

Distracting Sounds—This category includes any non-speech sounds that accompany the stuttering. Riley gives as examples noisy breathing, whistling noises, sniffing, blowing and clicking sounds etc. These are evaluated in terms of whether they are distracting to a listener.

Facial grimaces—Abnormal facial behaviors are illustrated by pursing of the lips, tensing the jaw muscles, blinking the eyes and protruding the tongue.

Head Movements—These are apparent when the participant makes head movement to avoid eye contact (e.g. turns the head away from the listener, looks down at the feet or around the room).

Movements of the extremities—These are movements of other parts of the body such as shifting around in the chair, foot tapping or fidgeting.

Ultimately, the judgments are subjective. One issue not commented on by Riley, is the co-occurrence of physical concomitants from different regions of the body (e.g. pursing of the lips coincident with head movement). When these occurred, they were noted in both affected categories. A separate judgment was made for each anatomical area (face, head, extremity), and for distracting sounds. Each of the four judgments was scored between 0 and 5 depending how distracting the concomitants were judged to be. The scale values were: 0 = none; 1 = barely noticeable unless looking for it; 2 = barely noticeable to casual observer; 3 = distracting; 4 = very distracting; 5 = severe and painful looking. The four scores were added to give an overall physical concomitant score between 0 and 20. Table 2 summarized the physical concomitant scores for category and the overall scores for each of the eight participants.

Total Overall Score

For both judgment procedures, the total overall score was obtained by adding together the scores for the three component elements (frequency, duration, and physical concomitants). The SSI-3 score was then looked up in Riley's (1994, p.12) Table 3 (this is appropriate for children who can read). A percentile can be obtained from this table and the percentiles are also given as stanines and one of five adjectival categories that describe the level of stuttering severity as very mild, mild, moderate, severe or very severe. Only the percentiles are reported in this article.

Procedure for obtaining frequency scores

For the digital procedure, the judges first listened to a 10s stretch of speech. This was listened to twice. On one occasion syllables were counted and marked on paper and on the second, stuttered syllables were counted and marked on paper again following the guidelines in Riley (1994). Two checks were allowed: once for all syllables and once for stuttered syllables.

For the live recordings, the complete file was listened to twice (once for all syllables and once for stuttered syllables) and one further check each was allowed for each aspect judged. For both judgment procedures, the results from spontaneous and read records were used to obtain the overall frequency score.

Procedure for obtaining duration scores

For the digital records, the longest duration segments were located in each 10 s section. Their durations were measured and noted from the audio files on computer using a display with a traveling cursor which has a calibrated timeline. The list of long duration segments was examined subsequently to obtain the three longest stutters. Stutters that straddled sections were grouped together at this stage.

A stopwatch was used as Riley recommends when live judgments were made and judges attempted to measure the duration of each individual stutter and note them down. The list of long duration segments was examined as with the digital method to obtain the three longest stutters. For both measures, the three longest durations were averaged.

Counterbalancing of conditions

Three of the judges assessed the two recordings from the first four of the eight speakers according to the same judgment procedure (e.g. live) judging eight recordings (four speakers read and spontaneous samples) in total in this phase. They then switched to the other judgment procedure (in this case digital) and judged the two recordings from each of the remaining four speakers (another eight recordings). They then judged the two samples of the first four speakers they had judged in the other judgment procedure (in the example, since the first four speakers had been judged live initially, these were now judged digitally). Then they judged the second four speakers in the other judgment procedure. At this point they had judged all eight speakers two samples once each in each judgment procedure. Each judge repeated this entire set of judgments four times over. The other two judges assessed the speech in the same way except that they started with the digital judgments rather than the live ones. The order with which the material from the five speakers was judged was counterbalanced for both sub-groups of four judges so that two speakers heard the speakers in one order and the other two judged the speakers in the reverse order.

The judgments were self paced with participants typically doing an hour at a time and up to four sessions within a day. Each of the four complete sets of judgments (sessions) had to be completed over three days maximum. Minimum breaks of a day were taken before another session commenced (longer breaks up to a maximum of four days were allowed between sessions).

Results

The three research questions concerned whether judgments varied over speakers, procedures and sessions. All three attributes were included as factors in the same ANOVA. The ANOVA was mixed model and the factors included were judgment condition (digital versus live), speakers (eight levels) and session (the four sessions; these were treated as a repeated measure) and the dependent variable was SSI-3 score. There were main effects of procedure ($F_{1,64} = 6.56, p = .013$) and speakers ($F_{7,64} = 178.5, p < .001$). The main effect of procedure showed that SSI-3 scores were significantly higher for the digital than for the live condition. Thus, it appears that the digital procedure is more sensitive than the live procedure. The main effect of speakers showed that the speakers differed in severity score.

There was no main effect of sessions overall, but this factor interacted with speakers ($F_{7,64} = 2.5, p = .025$) and with judgment procedure and speakers ($F_{7,64} = 2.8, p = .014$). The interaction between sessions and speakers showed that SSI-3 scores varied across sessions at different rates for different speakers. The three-way interaction between sessions, speakers and condition showed that SSI-3 scores varied at different rates for different speakers and that this depended on whether judgments were made in digital or live conditions. Generally speaking, there was a bigger difference between judgment conditions for speakers with a more severe stutter. The differences between judgment conditions were limited to the more severe stutterers in the early sessions, whereas in the later sessions the differences extended to less severe stutterers. On repeated assessment sessions, judges are either refining the sensitivity in the digital assessment procedure or losing sensitivity in the live procedure.

Discussion

To summarize the findings, the ANOVA addressed the issue of whether SSI-3 scores obtained under different procedures that are permitted in the manual, varied with speaker severity and across sessions when the judgments were repeated. There were main effects of judgment procedure that showed the digital procedure led to more stuttering being detected (higher SSI-3 scores) than did the live procedure. More importantly, difference between

judgment conditions depended on speaker severity and varied across sessions. The effect of sessions was less than that of speaker (the main effect of the session factor was not significant but that of speaker was markedly significant with a p of less than .001). In practice this means that live procedures underestimate severity of stuttering particularly in acute cases. Clinicians need to know that they may be under-reporting severity, particularly in acute cases. A clinician might perform an SSI-3 live assessment and classify a client as moderate, whereas assessment of the same data with digital procedures would locate more stutters and the client would be classified as severe. There are clear clinical implications if the moderate classification (rather than the severe one) is used as a factor in deciding about treatment for the client. It is possible to work out correction factors to bring scores obtained with different procedures into closer agreement. However, a caution is that although scores from different procedures would fall on the same scale after correction, separate standardizations for each procedure would still need to be performed.

Both procedures used to obtain the SSI-3 scores reported here are also allowed in SSI-4 (reference is made to use of computers, although Riley also permitted non computer-based assessments in clinics). (Riley ensured this in order to make SSI-4 downward compatible with SSI-3.) Whilst SSI-4 has some additions relative to SSI-3, these were left out in the current study for reasons given in the introduction. In brief, recordings over the phone were not used because this recording condition was not included when the standard scores were obtained. Also, the computer counting software may offer saving in clinicians' time, but it has not been checked for reliability and validity. With the computer counting method, comparison of counts of syllables and stutters with benchmark data need to be performed to establish their accuracy. If there are discrepancies between the computer and benchmark assessments, rescaling might be applied to bring the scores into correspondence, but this does not mean that the procedures (computer and benchmark) meet the same data collection standards. This, and the point about corrections for different procedures to bring the scores into alignment, underlines the importance of being clear about what procedures were employed when data are reported whether for use in clinics or in research reports: SSI-3 scores obtained whilst listening to complete records and obtaining syllables and stutters counted with the software that Riley (2009) supplied are not directly comparable with digital SSI-3 scores. Full documentation of how assessments are made is the minimum required in clinical work. Whilst it may not currently be expedient for clinicians to collect, store and analyze data according to digital procedures, this should be a long-term goal particularly for clinical work intended for publication.

In principle, it will be possible to make syllable and stutter counts according to SSI-3 recommendations in one pass once the SSI-4 computer procedure has been assessed. This would not necessarily require a permanent recording. However, until such checks on the software have been made, severity scores have to be made on temporary or permanent storage media (audio, video or computer hard disks) so that two passes can be made for assessing: overall number of syllables in pass one and number of stutters in pass two, as these steps were performed separately when the SSI-3 norms were obtained.

The digital procedure has to work from a stored file whereas live procedures can work from such a file or, once the software program has been assessed, syllables and stutters could be assessed in one pass and a permanent record would not be essential (although still desirable). The current results show that live procedures based on assessing syllables in one pass and stutters in another are not as sensitive as the digital procedures. Thus, it would appear unlikely that making the two judgments live in a single pass would be as sensitive as the digital procedure. Consequently, digital procedures should be adopted wherever possible, so a stored version has to be made.

There are many additional advantages in having files stored on computer: machines are routinely backed up, so records are unlikely to be destroyed. This is not always the case with video recordings. For example, Miller and Guitar (2010) noted that two of three clinicians whose data they wanted to use had not retained the video tapes. A further reason for computer storage is to meet data protection requirements in countries like Australia that specify a minimum length of time that research data should be retained (not a maximum one as some people think). Digital storage is efficient and easily controlled and provides semi-permanent storage so that keeping materials for specified times can be adhered to. There are also reasons that recommend digital storage that are grounded in research. For example, having a permanent recording allows alternative methods of assessing severity to be investigated on the same files with little additional effort (Howell, 2010). Examples would be where percentage of stuttered syllables, %SS, are defined in alternative ways, to those specified by Riley to make his severity estimates. Also, as new methods are developed for counting stuttering symptoms (Howell, Bailey & Kothari, 2010), results can easily be recalculated. For reasons like these, Howell (2010) advocated that digital archiving ought to be a requirement for funded research and noted that this is consistent with data protection that requires long-term storage.

The earlier review indicated that assessment of computerized counting methods, like that included in SSI-4, are needed urgently. Other further work should examine the effects of the procedures that we have reported here in less expert judges. Work ought to look at different specification of symptoms where some variation across sessions might be expected if symptoms included/excluded vary in ease of detection. If this is the case, further questions that should be raised are how many sessions are necessary for judgments to stabilize and whether this occurs at the same rate for digital and live procedures. Addressing these questions would help establish good practice by providing bias-free estimates and confidence intervals about reported judgments as well as potentially providing a basis on which to scale judgments made by one set of judges with a given procedure so they can be compared with other judgments made using different procedures. Practically speaking, having more reliable procedures whilst at the same time retaining the flexible features of SSI-3 where possible and being able to refer to standardized scores is an obvious advantage in outcome research. More applied work is needed to develop SSI-3 so it can be used as a diagnostic instrument and for its use in assessing stuttering prognosis.

Acknowledgments

This work was supported by grant 072639 from the Wellcome Trust to Peter Howell. Address correspondence to: Peter Howell, Division of Psychology and Language Sciences and Centre for Human Communications, University College London, Gower Street, London WC1E 6BT, England. p.howell@ucl.ac.uk

References

- Davis S, Shisca D, Howell P. Anxiety in speakers who persist and recover from stuttering. *Journal of Communication Disorders*. 2007; 40:398–417. [PubMed: 17157866]
- Howell P. *Recovery from Stuttering*. Psychology Press. 2010
- Howell P, Bailey E, Kothari N. Changes in the pattern of stuttering over development for children who recover or persist. *Clinical Linguistics and Phonetics*. 2010; 24:556–575. [PubMed: 20462359]
- Howell P, Davis S, Williams R. Late childhood stuttering. *Journal of Speech, Language and Hearing Research*. 2008; 51:669–687.
- Howell P, Davis S, Bartrip J. The UCLASS archive of stuttered speech. *Journal of Speech, Language and Hearing Research*. 2009; 52:556–569.
- Howell P, Davis S, Williams R. The effects of bilingualism on speakers who stutter during late childhood. *Archives of Disease in Childhood*. 2009; 94:42–46. [PubMed: 18782846]

- Jones M, Onslow M, Packman A, Williams S, Ormond T, Schwarz I, Gebiski V. Randomised controlled trial of the Lidcombe programme of early stuttering intervention. *British Medical Journal*. 2005; 331:659. [PubMed: 16096286]
- Lewis KE. Do SSI-3 scores adequately reflect observations of stuttering behaviors? *American Journal of Speech-Language Pathology*. 1995; 4:46–59.
- Miller B, Guitar B. Long-term outcome of the Lidcombe Program for early stuttering intervention. *American Journal of Speech-Language Pathology*. 2009; 18:42–49. [PubMed: 18845699]
- Riley, GD. Stuttering severity instrument for children and adults (SSI-3). 3rd ed.. Pro-Ed, Inc; Austin, TX: 1994.
- Riley GD. A Stuttering Severity Instrument for Children and Adults. *Journal of Speech and Hearing Disorders*. 1972; 37:314–322. [PubMed: 5057250]
- Riley, GD. Stuttering severity instrument for children and adults (SSI-3). 3rd ed.. Pro Ed; Austin: TX: 1994.
- Riley, GD. Stuttering severity instrument for children and adults (SSI-4). 4th ed.. Pro-Ed, Inc; Austin, TX: 2009.
- Ryan, BP. Programmed therapy for stuttering children and adults. Charles C. Thomas; Springfield, IL: 1974.

Table 1

Information about individual participants and summary statistics. Children are identified in column one with an indication of speaker group (persistent or recovered, P or R) and a number. Participant numbers correspond with those used in Howell et al. (2010). Age (in months) and SSI-3 at the time the children were originally seen when they were in the age range 8 to 10 years is given in column two and the corresponding SSI-3 estimate on this occasion is given in column three. These are background details about the children as these recordings were not used in the experiment. Age (in months) and SSI-3 score from Howell et al. (2010) are given for the age range 10-12 years in columns four and five. These recordings were used in the current assessments. These SSI-3 assessments were made by a trained independent judge who was not one of the judges used in the current experiment (the judgments were made in a similar way to the digital judgments reported here).

Id	Age	SSI-3	Age	SSI-3
	8-10	8-10	12 +	12 +
P2	110	34	156	36
P14	96	34	156	31
P7	119	38	169	29
P9	118	31	183	29
P1	118	30	187	25
R4	116	24	158	21
R11	118	34	161	15
R2	119	30	196	11
Mean	114.25	31.87	170.75	24.62
SD	7.94	7.94	15.80	8.45

Table 2

Physical concomitant scores for the five speakers. The separate scores for each of the four categories (distracting sounds, facial grimaces, head movements and movements of the extremities) are given as well as overall scores.

Id	Distracting sounds	Facial grimaces	Head movements	Movements of the extremities	Total score
P2	0	0	2	2	4
P14	1	0	2	4	7
P7	0	0	1	3	4
P9	0	1	3	3	7
P1	1	0	2	1	4
R4	0	0	0	0	0
R11	0	0	0	2	2
R2	0	1	3	2	6