Journal of Fluency Disorders 38 (2013) 102–123

Contents lists available at SciVerse ScienceDirect

Journal of Fluency Disorders

Screening school-aged children for risk of stuttering☆

Peter Howell*

Division of Psychology and Language Science, University College London, Gower St., London WC1E 6BT, England, United Kingdom

ARTICLE INFO

Article history:
Available online 24 October 2012

Keywords:
Recovered developmental stuttering
Persistent developmental stuttering
Stuttering severity
Screening for fluency
Prognosis of stuttering

ABSTRACT

Objectives: Howell and Davis's (2011) model that predicts whether stuttering in eight-year old children will persist or recover by teenage was adapted for screening school-aged children for risk of stuttering. Stuttering-severity scores were used to predict whether children belonged to fluent or stuttering groups. Predicted group assignments were compared for models in which severity measures were made with whole-word repetitions excluded or included. The best model for distinguishing children who stutter (CWS) from fluent children was validated across a wide range of ages.
Design: Stuttering-severity scores from CWS (222 for development, and 272 for validation, of the models) and fluent children (103 for development, and 25 for validation, of the models) were employed. Models were developed that predicted prognosis and screened CWS and fluent children. All these analyses were conducted both with whole-word repetitions excluded and included in the stuttering-severity scores. The model that screened fluent children from all CWS which excluded whole-word repetitions was validated for children across a range of ages.
Results: All models achieved around 80% specificity and sensitivity. Models in which whole-word repetitions were excluded were always better than those which included them. Validation of the screening for fluency with whole-word repetitions excluded classified 84.4% of fluent children, and 88.0% of CWS, correctly. Some of these children differed in age from those used to develop the model.
Conclusion: Howell and Davis's risk factor model for predicting persistence/recovery can be extended to screen school-aged children for fluency.

Educational objectives: After reading this article, participants will be able to: (1) describe the difference between finding group differences and risk factor modeling in stuttering research; (2) summarize the strengths and weaknesses of stuttering severity instrument version three; (3) discuss how validation of diagnostic and screening models for fluency can be performed; (4) see how risk models have potential applications for screening for communication disorders in general.

1. Introduction

The first contact a speech language pathologist usually has with a child who stutters (CWS) is when the child attends a clinic for confirmation of diagnosis of the disorder and to decide on a course of treatment. There has been a period prior to the child's appearance at clinic where the child and his or her family had little or no professional advice about

☆ This work was supported by grant 072639 from the Wellcome Trust to Peter Howell. Thanks are due to Glyn Riley for comments on Appendix A.
* Correspondence address. Tel.: +44 020 7679 7566; fax: +44 020 7436 4276.
E-mail address: p.howell@ucl.ac.uk

0094-730X/$ – see front matter © 2012 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jfludis.2012.09.002

stuttering. Early clinical intervention is constrained when there is such a delay between when the disorder started and consultation at the clinic. This is potentially a problem because early intervention is usually considered to be more effective than later intervention (Yairi & Ambrose, 2005). The delay would be reduced if there were convenient methods for screening large unselected groups of children in order to identify stuttering at key stages in development (e.g. at school entry). However, no such screening instrument is currently available. Research, that may aid development of a screening instrument, has shown that children diagnosed as stuttering differ in many ways from fluent children (Yairi & Ambrose, 2005). Any of the factors that show significant differences are potentially useful for screening children for stuttering. The success with which such a factor correctly classifies the two groups of children as stuttering or fluent can be established by a statistical procedure such as logistic regression (Reed & Wu, in press). A clinical cohort does not have a supply of fluent children, and this is one reason why screening procedures have not been developed to date.

A related question, that has been the focus of much research, is whether the prognosis of children in a clinical cohort (identification of CWS who will go on to recover or persist) can be predicted at their initial examination (Howell, 2010; Yairi & Ambrose, 2005). Children who subsequently recover differ from the CWS who persist on the majority of the same measures found to differ between fluent control children and CWS (Howell, 2010). To establish whether any of these factors plays a role in long-term prognosis for stuttering, first measures on factors at the initial examination where prognosis is not known need to be obtained. Then the status of stuttering at an age at which stuttering has resolved into its recovered or persistent form has to be established. Finally, measures on factors obtained at the initial examination need to be correlated with persistence or recovery established at the later age at which prognosis was determined (Howell & Davis, 2011). The risk factors that predict persistence of the disorder may provide valuable information that help clinicians target resources on those who are most susceptible to long-term fluency problems (Reilly et al., 2009, p. 271; Yairi, Ambrose, Paden, & Throneburg, 1996, p. 74).

There are two important implications of this discussion: first, just showing that two groups (fluent children versus CWS, or CWS who will persist versus CWS who will go on to recover) differ when some factor is measured, does not establish that what was measured is a risk factor for either stuttering in general or for its persistent form. In risk factor analysis a measure is taken (independent variable) and it is established how well it predicts group membership (dependent variable). When groups are tested to see whether they differ on some measure, the groups are selected (independent variable) and the measure is examined to see whether it differs between the groups (dependent variable). As Reed and Wu (in press) pointed out, relative to risk factor analysis, studies that look for differences between groups reverse "the relationship between the outcome and the predictor variables, making the outcome of interest into the independent variable, and the predictors into the dependent variable"; Second, measures that increase the risk of starting to stutter are not necessarily the same as those that increase the risk of persistence (Howell, 2010). Similarly, measures that predict onset or prognosis of stuttering may or may not apply to screening.

The current study developed and assessed models for screening for stuttering that can be administered to unselected cohorts of children at selected ages, such as when they start school. The screening can be done in different ways, all of which are suitable for different clinical purposes. First, the screen might require children to be classified as fluent, likely to recover, or likely to persist (*screen for stuttering types*). This would be used if there is graded health care provision (e.g. parents of fluent children do not need to do anything about their child's fluency, parents monitor their child if he or she is considered likely to recover, whereas children likely to persist attend clinic). Second, the screen at the time of the first examination may need to separate the CWS who will go on to persist from fluent children and CWS who will later recover (*screen for persistence*). This may be appropriate if health services want to focus attention on children likely to have long-term fluency problems (i.e. the CWS who will persist). Third, the screen may be required to separate the children who are fluent from both those CWS who will go on to recover and those who will go on to persist (*screen for fluency*). This would be appropriate if health services want to examine all CWS irrespective of the expected path their stuttering will take. All three types of screen are addressed in this study.

The next question is what factor or factors to measure (the independent variables). As mentioned above, the risk factors that are successful when examining one topic (e.g. prognosis that has been worked on) may or may not be useful for other topics (here screening that has not been addressed previously). The factor that was examined as potentially useful for screening was that used by Howell and Davis (2011) in their investigation into prognosis in a heterogeneous sample of CWS who were followed up longitudinally between the ages of eight years and teenage. The essential detail about Howell and Davis's study needed at this point (full details are given in Section 1.3) is that although they examined a wide range of risk factors that were obtained on CWS around the age of eight years, only one of them predicted whether the CWS would recover or persist at teenage (teenage is the age at which most childhood stuttering has resolved into recovered or persistent form). This was stuttering severity measured according to version three of Riley's (1994) instrument, SSI-3 (see Appendix A for a description and an appraisal of SSI-3 and Section 1.1 for details about SSI-3 that are particularly pertinent for the current work). Consequently, SSI-3 was examined as the potential predictor factor for each form of screen in the study reported below. SSI-3 may be useful for screening for stuttering because it incorporates a measure of the symptoms of stuttering and these would be expected to be rarer in fluent children.

SSI-3 has a precise way of measuring severity. Probably the most notable aspect of SSI-3 is that it does not consider whole-word repetitions (WWR), as in "my, my, my friend", to be symptoms of stuttering when percentage of syllables stuttered (%SS) are calculated. There is debate about whether WWR should or should not be included in %SS counts. In the work below,

SSI-3 was calculated both with WWR excluded and included and the performance of the prognosis and screening models that resulted was compared.

The remainder of the introduction starts by giving background information about why SSI-3 may be successful in prognosis and screening (Section 1.1). Evidence about the general role of WWR with respect to stuttering and the potential roles of WWR in screening and prognosis of stuttering are considered in Section 1.2. Section 1.3 compares the proposed modeling approach with approaches adopted in other risk factor studies on stuttering (prospective work on an unselected cohort, work that reports retrospectively whether adults had stuttered, and Howell and Davis's work on prognosis).

1.1. Properties of SSI-3 that are potentially relevant for prognosis and screening

As indicated in Appendix A, the severity score supplied by SSI-3 is based on three components obtained on one or more speech samples: (1) the percentage of syllables stuttered (%SS) where stutters are defined in a precise way; (2) the duration of the three longest stutters in the samples; and (3) a score based on observed physical features extraneous to speech shown by the speaker (called physical concomitants). The conversion tables to produce values for each of the component measures so that they can be summed to give the total SSI-3 score are also given in Appendix A. The following issues about use of SSI-3 scores are singled out for further discussion here because they have relevance to performance of SSI-3 as a risk factor for prognosis and screening in the current study.

(1) **Exclusion of WWR in counts of %SS**: The general debate about whether WWR are, or are not, a symptom of stuttering is an important one (see the discussion in Section 1.2). Here, two alternative ways of making symptom counts that have been used in SSI-3 are considered (with WWR excluded or included). Riley (1994) stated that WWR are usually not counted as stutters (see Appendix A). He also specified when words in WWR may be counted as stutters, namely when they are "shortened, prolonged, staccato, tense, etc." However, when repeated words have the latter properties, they would be classified as part-word repetitions, prolongations or word breaks respectively. Because of this, they would be counted as stutters in any case, so this advice regarding when to count such WWR as stutters is not necessary (Howell, Soukup-Ascencao, Davis, & Rusbridge, 2011). Conversely, repeated words with these properties are not WWR. Other authors have a different point of view about WWR and whether they should be included in %SS for obtaining SSI-3 scores. For example, Anderson and Wagovich (2010) included *all* WWR in their counts of frequency of stuttered disfluencies (not just the short ones, etc.) that they then used to obtain SSI-3 scores.

    Instead of taking one position about whether or not WWR should be counted as stutters, in the study reported below empirical comparison was made between SSI-3 scores obtained with WWR excluded (Howell et al., 2011) and with them included (Anderson & Wagovich, 2010). For the stutter counts with WWR excluded, WWR were not counted as stutters unless they had properties that led them to be classified as other types of symptom allowed by Riley in counts of %SS. A further point to note is that leaving WWR out of SSI-3 calculations can affect duration scores and overall syllable counts as well as %SS (see Section 2). For the stutter counts with WWR included, all WWR were counted as stutters. The comparison of SSI-3 with WWR excluded and included should offer some indication about which way of counting %SS in SSI-3 is appropriate based on the impact this has on prognosis and screening.

(2) **Threshold cutoff value for fluent speakers**: A fluent threshold is required when SSI-3 is used in screening unselected samples of children. Thresholds are given in the SSI-3 manual for very mild, mild, moderate, severe and very severe stuttering. However, although Riley (1994) indicated that SSI-3 scores are useful for diagnosis, there is no cutoff below which a child would be called fluent and above which a child would be said to stutter. Howell and Davis (2011) estimated an approximate cutoff value of 8 (the lowest score of CWS) for fluent children (the current study revises and extends this). The precise value may be procedure-dependent (see point 4 below, which describes how SSI-3 scores vary across the procedures that are permitted in the manual). As discussed in Appendix A, the original standards were obtained from audio recordings, and spontaneous and read samples alone were used for readers. These conventions about format and speech sample type are adhered to in the current study so that Riley's conversion tables and standards can be applied.

(3) **Scaling of scores across the severity range**: Examination of the conversion tables for the %SS and duration components of SSI-3 show that the converted scores increase rapidly at the lower end of the scale (see Appendix A). Skewing components of the severity measure so their scores make proportionately more contribution for low-scoring individuals may be important for detecting children with low, but significant, levels of stuttering (further details about scaling of SSI-3 scores are given in Appendix A). The sensitivity of SSI-3 in situations where fluency problems are mild or absent may make the instrument particularly suited to separating fluent children from CWS.

(4) **Flexibility and constraints on how SSI-3 assessments are made**: There is some flexibility on how SSI-3 scores are obtained insofar as different procedures can be used. Riley (1994) offered this flexibility so that his instrument could serve diverse needs ranging from use in clinics to research laboratories, and this has obvious advantages. However, the way that the different procedures affect scores and the impact these have on prognostic and screening predictions need to be ascertained. This study makes a partial contribution by comparing SSI-3 with WWR excluded or included as risk factors for prognosis and screening. Other than this, the procedures employed were constrained so that all severity measurements were obtained in the same way because it is known that different procedures lead to differences in SSI-3 scores (Howell et al., 2011; Jani, Huckvale, & Howell, submitted for publication).

In summary, SSI-3 is potentially useful for screening because it excludes WWR and as it includes three different component measures (both of these may allow SSI-3 scores which may help distinguish fluent children from CWS). It was observed that SSI-3 has no fluency cutoff value and that SSI-3 scores are skewed so that mild fluency problems have more impact than severe problems in determining whether a child stutters. Although procedural flexibility is useful for ensuring that SSI-3 can be used in diverse situations, a single procedure has to be adhered to when data are pooled for analysis (Reed & Wu, in press).

1.2. Role of WWR in relation to stuttering

The current study provides some data as to whether WWR are symptoms of stuttering or not and their role in distinguishing fluent children from CWS in that the performance of the screening models with WWR excluded and included were compared. This section reviews other evidence about whether WWR should or should not be considered stutters.

Some studies have found differences in the nature and frequency of WWR between fluent and stuttering pre-schoolers that suggest WWR may have a different role in the two groups of children. For instance, pre-schoolers who stuttered produced monosyllabic WWR 3.5–4 times more frequently than non-stuttering pre-schoolers (Hubbard & Yairi, 1988; Yairi & Lewis, 1984), and fluent children tended to pause longer than CWS between the spoken segments of a WWR (Throneburg, Yairi, & Paden, 1994). Whilst these studies might suggest that WWRs differ between young CWS and fluent children, their role as a risk factor in screening for stuttering has not been established nor has their use in predicting risk for long-term stuttering (prognosis) been addressed.

In contrast, there is more evidence on older children and adults that suggests WWR are not symptoms of stuttering. There are several pieces of empirical evidence that indicate WWR have a different role to symptoms that are more typical of stuttering, such as prolongations, part-word repetitions and word breaks. Longitudinal work by Howell, Bailey, and Kothari (2010) showed that CWS who recovered had a higher rate of WWR relative to more typical symptoms than did persistent CWS when they were first examined (i.e. when both groups were stuttering), and these differences between the rates of these two types of symptoms across the groups of CWS increased up to teenage, at which age stuttering had resolved into recovered or persistent forms. This suggests that the decrease in proportion of WWR relative to more typical stuttering symptoms might be a risk factor for persistence of stuttering. One empirical study that supports this suggestion is that operant conditioning to increase the rate of WWR gave temporary improvement in fluency (Reed, Howell, Davis, & Osborne, 2007). A second study also supported the general implication that WWR are associated with recovery. Japanese is a highly inflected language. Consequently, since most WWR occur on monosyllabic words, Japanese speakers do not have many opportunities to produce WWR in their language. The more typical symptoms predominate in Japanese CWS (Ujihira, 2011). As a result of the low rate of WWR, the chance of recovering from stuttering in Japanese children is lower than that for English-speaking children (Ujihira, 2011). A final compelling piece of evidence that WWR have different roles to more typical stutters was from a scanning study (Jiang, Lu, Peng, Zhu, & Howell, 2012). First of all, it was established that more typical stutters have different brain activity patterns from other disfluencies that are commonly seen in fluent children's speech (WWR were not included in either group of symptoms at this stage). Subsequently, automatic classification procedures categorized the brain activity pattern of WWR as a member of the other disfluency class rather than the more typical class. Thus, the brain activity of WWR is different from more typical stutters and similar to that of the other disfluencies that are frequently seen in fluent speech.

In summary, opinion is divided about whether or not WWR should be considered as symptoms of stuttering. Rather than take a particular stance on this issue, the current study examined whether SSI-3 scores calculated with WWR excluded or included (as indicated, there are precedents for both of these in the literature) affected prognostic and screening classifications. If different classification patterns occur in prognosis and screening when SSI-3 is used with WWR excluded as opposed to included, then these would provide additional evidence about the status of WWR.

1.3. Comparison of the current approach with other approaches to risk factor modeling in stuttering

A recent study by Reilly et al. (2009) used a cohort of young English-speaking Australian children to examine whether a range of factors predicted which children would start to stutter. The risk factors examined were based upon studies that had established differences between CWS and fluent children. They then used logistic regression in a prospective study of a large sample of children and determined which children subsequently started to stutter. A few of the factors that were measured before stuttering started correlated significantly with stuttering outcome (indicating they may be risk factors for stuttering). However, the overall fit of the logistic regression model was poor. This arose because the sample consisted predominantly of fluent children, which restricted the ability of the logistic regression model to identify CWS as such (hits) and increased the chance of calling CWS fluent (misses) (Howell, 2009). This problem is common to most multivariate techniques when they are used to develop a model for data where one class dominates. In these situations, a satisfactory model is difficult to develop because there is a tendency to place all cases in the class with maximum members as this ensures an accuracy rate of at least that of the most frequent category (Howell & Davis, 2011; Reed & Wu, in press). To illustrate, if the chance of stuttering is 5%, five children in a sample of 100 will stutter. Automatic classifiers using data from one or a number of measures obtained from the 100 children would achieve 95% correct performance by calling all children fluent. This performance looks reasonable, but no CWS is correctly classified. The problem does not apply when the number of cases in each class is balanced. Other

aspects to note about Reilly et al.'s (2009) study were that there was significant selective attrition (as more mothers without degrees withdrew their children from the study than mothers with degrees), and that they considered WWR as a symptom of stuttering. Mothers with degrees might have particular interest in their child's language development, so the attrition of mothers without degrees is of concern. As has been seen in Section 1.2, WWR is a questionable symptom whose role with respect to stuttering has been debated for many years amongst experts. The inclusion of WWR would lead to misdiagnosis of stuttering in very young fluent children starting to speak if they are not symptoms of stuttering. This would explain the exceptionally high rates of stuttering reported by Reilly et al. (2009).

Ajdacic-Gross et al. (2009), was a large-scale retrospective and opportunistic study. Swiss army conscripts self-reported stuttering (there was no attempt to distinguish recovered from persistent forms). This revealed some factors that corresponded with reports in other work, and other unexpected factors. The strengths of the paper were that a large number of participants was involved and they underwent detailed psychiatric examinations. As it was a retrospective study about factors that increased the chance of any type of stuttering, it suggested some risk factors for stuttering in general, but it did not address prognosis. The limitations in the study were due to the nature of retrospective studies (reported stuttering in the past cannot be verified) and, as the data were obtained as part of a general screening for conscripts (not specifically for stuttering), the authors did not include certain factors that are usually considered pertinent to stuttering.

Howell and Davis's (2011) model for prognosis offers the possibility of addressing screening (neither of the previous studies would allow this). Howell and Davis (2011) used children who were confirmed to stutter at age eight who were known to have either continued stuttering or not (persistent/recovered) at teenage. The number of CWS who subsequently recovered was roughly equal to the number of CWS who persisted so development of the model avoided the imbalance problem discussed earlier in this section. Seven risk factors were available at age eight. These were: (1) Head injury; (2) Age at onset of stuttering; (3) Family history; (4) Handedness; (5) Speaking two or more languages in the preschool years; (6) Gender; (7) Stuttering severity. Logistic regression can automatically determine which factors from a group like this are significant predictors of the outcome and which are not (Reed & Wu, in press). Howell and Davis (2011) used this procedure and found that the SSI-3 scores were the only factor that predicted whether the CWS at age eight would recover or persist by teenage and they did so with around 80% sensitivity (called persistent CWS, persistent) and specificity (did not call CWS who would later recover, persistent). These findings have been replicated by Cook, Howell, and Donlan (2012) for a sample of German CWS.

Howell and Davis (2011) pointed out that their model could be adapted for use with fluent children and that it could be extended to younger ages, which together potentially allow its use in screening unselected samples of children over a broad age range. The steps in the argument are: that SSI-3 can be obtained from fluent children as well as CWS; data are required on all participants that ensure that a child is fluent or stuttering when first examined and where there are longitudinal measures on CWS over the age range eight to teenage so that the form of stuttering at teenage (recovered/persistence) can be confirmed; that the models for different types of screen need to be validated and, if children from outside the age range that was used when the model was developed are included in this and performance is acceptable, it can be assumed that the models are applicable across these ages; similarly when the models are validated with samples from an unselected sample of children where there are imbalances between the numbers of fluent children and CWS and between the number of CWS who later recover and CWS who persist, it can be determined whether the performance of the models established on balanced samples is maintained. If so, the models developed on balanced data automatically adapt to any imbalances between fluent and CWS classes such as those that occur in the population of children at large. These points are each considered in turn in more detail.

As mentioned, SSI-3 is a measure that can be obtained for fluent children. This is not the case for some of the other factors Howell and Davis (2011) obtained, such as age of stuttering onset. This unavailability issue would not apply to all of the other factors Howell and Davis (2011) examined. However, measurements on certain of the remaining factors are biased in other ways depending on whether they are obtained from fluent children or CWS, making them unsuitable for screening to distinguish these groups. As one example, Reilly et al. (2009) noted that history of reports of stuttering in the family changed dramatically as a child changed from not stuttering to stuttering, so family history measures depend on whether or not stuttering has been diagnosed. Similar influences might apply to factors like head injury if parents seek an explanation for why their child started to stutter. There are no such problems with SSI-3, which can be measured in the same way on fluent children as on CWS. Consequently, fluent children can be included in screening models that use SSI-3 as the sole predictor (Howell & Davis, 2011).

One aspect of stuttering in the eight-teenage age range is that there are roughly equal numbers of CWS who will persist or recover at teenage (the current study adds approximately equal numbers of fluent children that are used in model development). This distribution allows optimal models to be developed for separating fluent and stuttering groups. Models for all the screens require confirmation that the children are either fluent or stutter when they are first seen. A further requirement is that longitudinal data are needed for screens for stuttering types and persistence, so that CWS can be classified as recovered or persistent at teenage (as mentioned earlier, teenage is the age at which most developmental stuttering has resolved into recovered or persistent form).

After the above steps have been attended to, a model can be developed that predicts class membership of different combinations of fluent, CWS who will later recover and CWS who persist (depending on the type of screen). Any such model can be used to see which of the specified classes newly-assessed individuals are assigned to. These results can be used to validate the model against the available clinical assessment classifications for these new individuals where classifications

depend on the particular screening type. Such validations can be done for children of any age (not just those in the age range that the model was developed on). As well as validating the selected model, the data on children outside the age range for which the model was developed show how well the model generalizes over age groups.

Finally, the models can be examined with unselected samples of children to see whether their performance scales up appropriately when samples with natural imbalance are examined. For instance, if the screen for fluency was used to assess a sample of school children (i.e. with a different distribution of fluent children versus CWS from that in the sample on which the model was developed), does it maintain the ability to correctly classify 80% of the fluent children and 80% of the CWS? In general, whether the screening models scale appropriately across data with class imbalances can be established empirically by determining whether the classification performance of each group involved in the screen is maintained in the samples used for validation (around 80% correct for each type of classification).

1.4. Summary and predictions

In summary, Howell and Davis (2011) found that the only risk factor needed to predict prognosis was SSI-3. An SSI-3 score can be obtained for fluent children, so SSI-3 is used as the factor in the risk factor models developed for prognosis and screening in this study. Performance of the models with SSI-3 calculated with and without WWR included as symptoms of stuttering are compared empirically to see whether WWR should be considered symptoms of stuttering. All three types of screen are examined (for persistence, for fluency, for types of stuttering). Satisfactory models for risk of stuttering cannot be developed for data that have imbalances between fluency classes (Reilly et al., 2009). Imbalances between numbers of CWS who will recover and CWS who persist were avoided here by using data from a similar age range to that used by Howell and Davis (2011) for model development. Roughly equal numbers of fluent children were added when models for screening were developed again to minimize the imbalance problem. The models can be validated outside the range they were developed for when data are available on children who have been independently classified as fluent or stuttering and, for some models, classification of CWS as will recover later or persist. The assumption that the models automatically scale when the distribution across classes changes (e.g. towards a high number of fluent children compared to CWS in unselected samples of school children) can be tested.

The current study addressed the following questions concerning childhood stuttering: (1) was SSI-3 successful in predicting prognosis in the Howell and Davis (2011) study because WWR were excluded in its severity counts? This was assessed by seeing how well a model that uses SSI-3 at age eight to predict persistent or recovered outcome at teenage performed when WWR were excluded and included in the SSI-3 scores; (2) are SSI-3 scores useful for screening for: (a) persistence; (b) fluency; (c) stuttering types? As was the case when prognosis was examined, assessments were made with and without WWR included in SSI-3 scores; (3) at present, cases that can be used for validation are only available that indicate whether a child stutters or is fluent when first examined (not, in the case of CWS, whether the child persisted or recovered at teenage). Consequently, validation within and beyond the age range used in model development can only be performed for screening for fluency that does not require prognostic information. If this validation is successful, the model can be used to extrapolate prediction to cases outside the age range for which it was developed and to situations where the CWS are in a minority (e.g. at school intake).

2. Method

2.1. Participants

None of the children had a history of speech or language problems other than stuttering for the CWS. They had no history of hearing disorders apart from non-acute otitis media, in some cases, which lasted no longer than two weeks. This information was obtained from reports by parents. All children could read an age-appropriate text supplied in Riley (1994).

2.1.1. Recovered CWS and persistent CWS participants used for model development

The CWS participants who were used to develop the models for examining prognosis and screening were mainly those used in Howell and Davis (2011). All Howell and Davis's participants were employed plus 16 additional CWS who were assessed in the same way (these additional participants were included to maximize sample size). In total, there were 222 children aged between eight and ten years when first assessed (182 boys and 40 girls). Inclusion criteria were as follows. The specialist clinic took children from different areas around London, encompassing a mixture of socio-economic status (SES) communities. All these children had been referred to this clinic that specialized in developmental stuttering and each child was confirmed as stuttering by a trained and qualified specialist speech pathologist. All CWS were then given treatment. The treatment was in the form of a one-week intensive course and it was ascertained that this was the only treatment received. All the CWS who were used for model development were reassessed at teenage to see whether they had persisted or recovered.

Three standardized and validated instruments for assessing persistence and recovery for CWS at teenage were employed (Howell, Davis, & Williams, 2009). These were as follows: (1) a structured report that obtained specific information from one of the parents of the CWS (Parent Report Form); (2) a similar report from the CWS (Child Report Form); (3) a set of ratings given by a trained researcher who interviewed the child and a parent for at least 40 min (Researcher Report Form). All three

assessments had to be in agreement for the participant to be classified as persistent CWS or recovered CWS; this was true of all 222 children used here (in Howell & Davis, 2011, 16 CWS failed this requirement). One hundred and twelve of the group of 222 participants (50.5%) were classified as recovered CWS.

2.1.2. Fluent participants used for model development

One hundred and three fluent children were recruited from schools in the same areas of London as the CWS to ensure an SES match. The children were also matched to the initial test ages of the children in Section 2.1.1 (in the age range 8–10 years). There were approximately equal numbers of males (46) and females (57). Although there were more males than females in the corresponding sample of CWS, gender was not a significant predictor in Howell and Davis (2011). All reports about the children were independently validated by the teachers.

2.1.3. CWS participants used for model validation

Two hundred and seventy two CWS (223 male, 49 female) aged between five and 19 years were used for validating the model (number of participants at each age are reported in Section 3.3). These children were assessed in the same way as in the initial assessment of the CWS as described in Section 2.1.1. They attended treatment, but some were not available for subsequent follow-up because they had moved outside the catchment area or contact was lost. Others were followed up but were not in the 8–10 age range when they first attended clinic. Speech samples were available from this first session (not necessarily in the 8–10 year age range).

2.1.4. Fluent participants used for model validation

Twenty five fluent children were recruited for validation from the same source as the fluent children in Section 2.1.2. There were 19 males and six females. Inclusion criteria were the same as for the children in Section 2.1.2 (again the number of participants at each age are reported in Section 3.3).

2.2. Procedure for SSI-3 assessments

All the speech recordings for estimating SSI-3 used here were obtained when the children were first seen (which was before treatment for CWS). In all cases, a spontaneous speech sample and a reading of an age-appropriate text from Riley (1994) was obtained. The recordings were made in a quiet room by a trained researcher or a qualified speech pathologist. Physical concomitants for the spontaneous and read material were observed and noted at the time of the recordings according to Riley (1994). SSI-3 scores were obtained using the child-reader table from Riley's (1994) manual (see Appendix A for details).

The following comments concern how severity assessment was performed when different procedures were allowed in the manual or where the advice was ambiguous. First, SSI-3 allows several data formats to be used for collecting material to estimate a child's severity (Howell et al., 2011). Audio recordings were used here. 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/).

Second, the procedures used to assess the data involved selecting and playing short extracts. The first 200 words in spontaneous monologs and all words in the SSI-3 read text were analyzed (the texts are each approximately 200 words in length) as indicated in the SSI-3 manual. All syllables (fluent and stuttered) were annotated on the files according to Riley's (1994) description. The stuttered syllables were marked on the files as indicated in the manual first of all with all WWR excluded. Each designated stutter was considered a single syllable as advised by Riley (1994). Since WWR were counted as fluent at this stage, all repetitions of each word in a WWR were included in the syllable counts. In other words, multiple iterations of a monosyllabic WWR were counted as separate syllables. The %SS was then calculated using the fluent and stuttered syllable counts. The three longest stutters in each sample were located and their durations were estimated (SFS has a time scale which makes this accurate). Next %SS was calculated with WWR included. Here, the entire WWR event counted as one stuttered syllable (as directed by Riley for all other types of stutters). As well as counting the WWR as a stutter, the number of repeats of the word in the WWR was deducted from the fluent syllable count. If one or more (up to a maximum of three) WWR were longer in duration than any of the stutters used for the original duration counts, they replaced the latter and the duration score was recalculated.

Once all three component scores were obtained (%SS, duration and physical concomitants), %SS and duration were converted using the tables in Riley's manual (see Appendix A). 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).

2.3. Intra- and inter-judge reliability of SSI-3

Intra- and inter-judge reliability were assessed separately for %SS and duration of the three longest stutters (two of the scores required for calculating SSI-3). Eight judges, with the same training as those used to assess the materials, judged additional samples (one spontaneous and one reading) from 10 CWS, similar to those used here. The samples of speech used in this check were not part of the current study. Data from the eight judges on these 10 sample-pairs were used for assessing

inter-judge reliability. Agreement between all pairs of judges for %SS with WWR excluded and included varied from 82% to 89% and these resulted in *k* coefficients that represented a high degree of agreement (Fleiss, 1971). All *k* coefficients were above .75 which Fleiss characterizes as "excellent". Two judges repeated the assessments on these 10 sample-pairs a second time and these were used for assessing intra-judge reliability. The agreement for both of these judges on the 10 samples for %SS was 87%, giving a *k* coefficient which represented excellent agreement. The three longest events selected were always the same within or across judges, as the speech files being used for making SSI-3 scores were displayed in SFS as an oscillogram on a computer screen that displayed durations. For similar reasons, durations were very similar within and between judges (<3% difference in durations).

A second group of eight judges, with the same training as those used above, assessed samples from ten fluent children (a spontaneous recording and a reading for each child). Agreement for %SS with WWR excluded and included varied between 79% and 85%. These agreements gave *k* coefficients that were excellent according to Fleiss (1971), although they were slightly below the agreement levels for the CWS reported above. There was perfect agreement about which were the three longest stutters and duration was again within 3%.

2.4. Statistical analysis

2.4.1. Variables measured and setup for the binary logistic regression analyses performed

The SSI-3 scores that were calculated with WWR excluded and included, measured as described above, served as the independent variable in different analyses. Different dependent variables were employed in the various analyses (e.g. the target outcomes persistent versus recovered indicated in the last paragraph of Section 2.1.1 for prognostic analyses). The dependent and independent variables were entered into the binary logistic regression model setup option in SPSS when the dependent variable had two outcomes (multinomial logistic regression was set up in a similar way when there were three outcomes, e.g. fluent, CWS who later recover or CWS who persist).

The steps in the analysis procedure are illustrated for Howell and Davis's (2011) prognostic model, which is replicated later, in which the outcome was persistent or recovered (binary). The description here is simplified by focusing on the one predictor that was significant as this was the only predictor variable employed in the below analyses (SSI-3) for reasons given earlier.

2.4.2. Steps in a logistic regression analysis with one predictor

The odds ratio for prognosis was the probability of persistence divided by the probability of recovery. The next step was to take the natural log of this ratio (log odds) to ensure the statistic varied linearly. The coefficient associated with the predictor in the logistic regression equation indicated how much a change in the predictor changed the log odds of the outcome. To give a more understandable quantity, the log odds ratio was exponentiated (exp(b)) so it offered an indication of how much a point increase in the independent variable changed the dependent variable. Here, for example, how much the odds of persistence changed (dependent variable) for a point increase in SSI-3 score. Standard error estimates can be obtained for exp(b) and, if desired, they can be used to obtain 95% confidence intervals in the conventional way.

A goodness of fit procedure was used to establish whether the independent variable had a significant impact on the predicted outcome. This involved setting up a benchmark model (the null model) in which all predictors were absent (just SSI-3 was left out here). In Howell and Davis's (2011) study, the null model designated all cases likely to recover (these made up 52.3% of the actual cases). Sensitivity of the model (identification of the percentage of the group that persisted in this case) was 0% and specificity (identification of the percentage of the group that would recover later) was 100%. The same calculations were then made with SSI-3 included. In order to establish whether inclusion of the predictor significantly improved the prediction, the two models were compared to see whether the model with SSI-3 as predictor performed better than the model that made predictions at random (i.e. the null model). The comparison gives a statistic that varies according to the χ^2 distribution and from this it can be determined whether including the predictor makes a significant difference using χ^2 tables in the usual way.

2.4.3. Comparison of models for significance using SSI-3 with WWR excluded or included

The procedure for establishing whether a model that used SSI-3 with WWR excluded versus one that used SSI-3 with WWR included was significant, used the difference between the two goodness of fits in Section 2.4.2. That is, the difference between the fit of the models that used SSI-3 over the null model for SSI-3 calculated with WWR excluded and that for SSI-3 calculated with WWR included. This was assessed for significance against the χ^2 distribution in the usual way again.

2.4.4. Other goodness of fit indexes

Other goodness of fit statistics apart from that in Section 2.4.3 can be calculated. The ones reported below are the R^2 statistics of Cox and Snell (1989) and Nagelkerke (1991).

2.4.5. Classification tables

Classification tables were obtained on completion of the logistic regression analyses. In the example, the score of each child was evaluated on the logistic regression function. With the binary outcomes in the illustrative analysis, a score less than .5 would conventionally be designated "recover later" and a score greater than .5 would be designated as "persistent".

Once obtained, counts of children falling into each cell were entered in 2 × 2 contingency tables (here predicted to recover or persist versus actually recovered or persisted). Misclassifications for predicted probabilities less than .5 were cases with low predicted probability of persisting who went on, nevertheless, to persist. Misclassifications for predicted probabilities greater than .5 were cases with high predicted probability of persisting who went on, nevertheless, to recover.

2.4.6. Validation

Models need to be validated on independent data. The SSI-3 data from the children in Section 2.1.3 (CWS) and those in Section 2.1.4 (fluent) were used for these purposes. These children had not been used in the analyses that set up the models. The SSI-3 score of each child was used to obtain the model score and classification from the logistic regression function in a similar way to that described in Section 2.4.5. Once this classification was obtained, the accuracy of classification was established (e.g. whether a CWS was classified as fluent or CWS).

2.4.7. Summary of statistical analyses

Four types of model were investigated (one for prognosis outcome and three for screening outcome). For each of these four types, models with WWR excluded and included were evaluated. Exp(b) and its associated confidence interval are reported for all models. For all models, the χ^2 statistic indicating significance of a model with SSI-3 as a predictor over the null model (Section 2.4.2) and its associated degrees of freedom are given (all models showed a significant effect of having the predictor $p < .001$), as well as Cox and Snell's R^2 and Nagelkerke's R^2 goodness of fit statistics. Statistical comparisons between equivalent models where WWR was excluded or included (Section 2.4.3) are given. Classification tables are given for all models. The model that screened for fluency with WWR excluded was validated by independent cases of CWS and fluent children across a wide range of ages.

2.5. Ethics

Ethical approval for the study was granted by the University's Research Ethics Committee. Participants gave their informed consent after the procedures were fully explained.

3. Results

Results on the three main issues raised in the introduction (prognosis, screening, and validation of one of the models across a range of ages) are reported. Models where WWR were and were not included as stuttering symptoms in the computation of SSI-3 scores are given for the first two topics (only the model that used SSI-3 scores that excluded WWR was used in the validation analyses).

3.1. Predicting prognosis with and without WWR

Howell and Davis (2011), in their work on prognosis, only employed SSI-3 estimated with WWR excluded. This work is replicated below and comparison is made between model performance when SSI-3 scores excluded and included WWR. The data from the 222 eight-year old CWS who either continued or stopped stuttering at teenage (persistent or recovered) were used. Binary logistic regression models using SSI-3 measured in one of the two ways (separate analyses) were computed. The null model provided the benchmark against which improved classification when SSI-3 scores were entered as predictors was used to assess significance. The null model put all 222 cases in the likely to recover class (50.5% correct). Statistics on the models that excluded and included WWR are given in the first and second rows of Table 1 (labeled in the left-most column). The second column of Table 1 gives the χ^2 statistics for improvement of models that included the SSI-3 factor over the null model. The significant χ^2s in column two of rows one and two show that SSI-3 was a significant predictor for each model. Other statistics that indicate both the models excluding and including WWR gave a good fit to the data are Cox and Snell's R^2 and Nagelkerke's R^2 statistics (given in columns 3 and 4). Comparison across rows one and two shows that both these statistics were higher (better fits) when SSI-3 scores were calculated with WWR excluded than when they were included. The estimated odds ratio, exp(b), and confidence intervals around this are given in columns 5 and 6.

The classification tables for the models with WWR excluded and included are given in Table 2 (top half with WWR excluded and bottom half with WWR included). The model that used SSI-3 with WWR excluded showed an overall percentage correct of 81.1%. Sensitivity in locating CWS who would recover was 81.3% and specificity (calling CWS who persisted persistent) was 80.9%. These replicate the results of Howell and Davis (2011). Overall, performance of the model using SSI-3 with WWR included was 5.9% lower than with them excluded (bottom of Table 2). Both sensitivity (8.1% lower at 73.2%) and specificity (3.6% lower at 77.3%) were adversely affected when WWR were included. More CWS who would later recover were misclassified as persistent CWS than vice versa for this model. This happened because the CWS who would later recover produced more WWR than did the persistent CWS. The all-round poorer classification performance when WWR were included in SSI-3 scores showed that the symptoms did not aid prognostic classification and should be excluded from severity estimates (as Riley recommended). The fit of the models that used SSI-3 with WWR excluded compared to that which used SSI-3 with WWR included showed that the model with WWR excluded was significantly better than that with WWR included ($\chi^2 = 22.912$, degrees of freedom = 1, $p < .001$).

Table 1
Summary of the logistic regression models (multinomial when three categories were involved and binomial when two categories were involved). The model considered is given in column one. The χ^2 for the improvement of the model with SSI-3 included over the null model ($\chi^2_{\text{nulll-w/SSI-3}}$) is given in column two. The Cox and Snell R^2 and Nagelkerke R^2 goodness of fit statistics are given in columns three and four, respectively. Exp(b) and the associated confidence interval (CI) are reported in columns five and six.

Model	$\chi^2_{\text{nulll-w/SSI-3}}$	Cox and Snell R^2	Nagelkerke R^2	Exp(b)	CI
R versus *P* with WWR excluded	110.704 df = 1	.393	.524	1.306	1.217–1.401
R versus *P* with WWR included	87.792 df = 1	.327	.436	1.263	1.185–1.346
F/*R*/*P* with WWR excluded	382.884 df = 2	.754	.848	*R* = 1.358 *P* = 1.777	1.258–1.466
F/*R*/*P* with WWR included	364.585 df = 2	.720	.810	*R* = 1.392 *P* = 1.760	1.279–1.514
(*F* and *R*) versus *P* with WWR excluded	215.524 df = 1	.485	.671	1.333	1.246–1.427
(*F* and *R*) versus *P* with WWR included	184.034 df = 1	.432	.599	1.297	1.220–1.378
F versus (*R* and *P*) with WWR excluded	269.168 df = 1	.563	.788	1.378	1.281–1.482
F versus (*R* and *P*) with WWR included	236.913 df = 1	.518	.724	1.311	1.237–1.389

Table 2
Classification into CWS who would recover later versus persistent CWS when WWR were excluded from SSI-3 counts (top half) and included from SSI-3 counts (bottom half). The cells in the table are counts of individuals falling into the predicted/observed category and the marginals on the *x* and *y* axes represent percentage (%) correct.

		Predicted		
		Recovered	Persistent	% correct
Excluded				
Observed	Recovered	91	21	81.3%
	Persistent	21	89	80.9%
Overall %age		81.3%	80.9%	81.1%
Included				
Observed	Recovered	82	30	73.2%
	Persistent	25	85	77.3%
Overall %age		76.7%	73.9%	75.2%

3.2. Screening

Screening requires classifications involving fluent children as well as CWS. The available fluent children and the two types of CWS participants (those who would later recover and those who persisted) allowed models to be assessed for the three types of screen outlined in the introduction, that is screening for: stuttering types using multinomial logistic regression; persistence using binary logistic regression; and fluency also using binary logistic regression.

The idea that there may be one threshold SSI-3 value that can divide fluent children from all CWS and another that can divide persistent CWS from CWS who would later recover and fluent children was outlined in the introduction. Such SSI-3 thresholds may be influenced depending on whether the scores were calculated with WWR excluded or included. Either of the ways of obtaining SSI-3 scores could then improve classification for some or all of the three types of screening. This was tested by comparing each type of screening classification when SSI-3 scores were used that excluded or included WWR. The results provide an indication as to whether WWR are a useful symptom to include when making each of these three types of classification.

3.2.1. Screening for stuttering type analyses

Multinomial logistic regression was used to separate the children into three classes (fluent, CWS who would later recover and persistent CWS) using each type of SSI-3 score as the predictor (both when WWR were excluded or included) in separate analyses. Models with SSI-3 scores included as a predictor were significantly better than the null model when SSI-3 excluded and included WWR (column two of rows three and four in Table 1). Once again, Cox and Snell's R^2 and Nagelkerke's R^2 goodness of fit statistics were higher when SSI-3 scores were used with WWR excluded than when they were included. Exp(b) and its confidence interval are given in columns 5 and 6.

3.2.1.1. Prognosis results in the screening for stuttering type analysis. Examination of classification performance of just the persistent CWS and the CWS who would later recover (selected from the data in Table 3) in these analyses that include fluent children showed that, relative to the equivalent data in Table 2 which only used persistent CWS and CWS who would later

Table 3
Classification into fluent, CWS who would recover later and persistent CWS when SSI-3 was calculated with WWR excluded (top half) and when SSI-3 was calculated with WWR included (bottom half). The cells in the table are counts of individuals falling into the predicted/observed category and the marginals on the *x* and *y* axes represent percentage (%) correct.

		Predicted			
		Fluent	Recovered	Persistent	% correct
Excluded					
	Fluent	91	12	0	88.3%
Observed	Recovered	13	79	20	70.5%
	Persistent	1	20	89	88.3%
Overall %age		32.3%	34.2%	33.5%	79.7%
Included					
	Fluent	81	21	1	78.6%
Observed	Recovered	5	88	19	78.6%
	Persistent	2	24	84	78.6%
Overall %age		27.13%	40.9%	32.0%	77.8%

recover to develop the models, there was negligible change when SSI-3 was calculated with WWR excluded (.3% difference). Classification into prognostic categories improved slightly when the three stuttering types were screened compared to when just prognostic categories were classified (4.8%). The overall percentage correct for persistent CWS and CWS who recovered later was 80.8% in Table 3 compared to 81.1% in Table 2 when SSI-3 scores were obtained with WWR excluded, and 80% in Table 3 compared to 75.2% in Table 2 when SSI-3 scores were obtained with WWR included. There is no way to obtain a statistic to compare data from a three by three contingency table (Table 3) with data from a two by two contingency table (Table 2), so the data just discussed cannot be assessed statistically. However, based on the comparisons of the 2×2 and 3×3 classification tables, performance with persistent CWS and CWS who would later recover did not appear to deteriorate when classification of fluent children was also required when SSI-3 scores were obtained with WWR excluded or included.

A further fact of note is that overall performance for the persistent CWS and CWS who would later recover data in Table 3 is that correct classification rate was higher when SSI-3 scores were obtained with WWR excluded compared to those with WWR included (80.8% versus 80.0%) and the same applied for the two-class data in Table 2 (81.1% versus 75.2%). Thus, SSI-3 with WWR excluded produced superior overall performance in terms of classification accuracy over that using SSI-3 with WWR included when classifying persistent CWS and CWS who would later recover in situations where the classifications involved the two CWS groups and the fluent group or just the two CWS groups (Table 3).

3.2.1.2. Screening for stuttering type using SSI-3 measures obtained with WWR excluded and included. The preceding observations offer no indication concerning classification of fluent cases for the data in Table 3. (The fit statistics for the screening for stuttering type were indicated in Section 3.2.1.)

The classification data in Table 3 show that the percentage correct classification of fluent children was high, at 88.3%, when SSI-3 scores were obtained with WWR excluded, and performance here was 9.7% better than when SSI-3 scores were obtained with WWR included. The model that used SSI-3 with WWR excluded was significantly better than that with WWR included ($\chi^2 = 18.299$, degrees of freedom = 2, $p < .001$).

The confusions that occurred were mainly due to fluent children being classified as CWS who would later recover, which occurred whether SSI-3 scores excluded or included WWR. However, the number of confusions was higher when WWR were included. The confusions would arise if the speech of fluent children and CWS who would later recover classes overlapped (Wingate, 2001) and this was more marked when WWR were included in SSI-3 scores. Additionally, the CWS who would later recover were confusable with both fluent children and persistent CWS (particularly the latter) when SSI-3 was calculated with WWR excluded and included. The greater overlap in classifications of CWS who would later recover and persistent CWS, than CWS who would later recover and fluent children, suggests that all CWS should be placed in the same category. This contrasts with the grouping of fluent children and CWS who would later recover noted earlier. A further feature of note is that classification accuracy of CWS who would later recover was higher when WWR were included than when they were excluded (the opposite of what happened for the fluent children). Accuracy of classification performance for persistent CWS was better when WWR were excluded than when they were included. This would arise if WWR occur at different rates in persistent CWS than fluent children and CWS who would later recover. Although overlap was noted earlier between fluent children and CWS who would later recover and here between CWS who would later recover and persistent CWS, overall classification performance was good and, generally speaking, somewhat superior when SSI-3 scores were calculated with WWR excluded.

3.2.2. Screening for persistence using SSI-3 measures obtained with WWR excluded and included

Next SSI-3 scores with WWR excluded or included were examined as predictors when fluent children and CWS who would later recover were combined into one class and persistent CWS were the other class. The null model put all cases into the combined fluent children and CWS who would later recover class (66.2% correct), which is the benchmark for estimating whether SSI-3 scores improved the prediction. The fits were significantly better than the null model for both the model that

Table 4
Classification into fluent speakers and CWS who would recover later versus persistent CWS when WWR were excluded from SSI-3 counts (top part) and when WWR were included in SSI-3 counts (bottom part).

		Predicted		
		Fluent + Recovered	Persistent	% correct
Excluded				
Observed	Fluent + Recovered	194	21	90.2%
	Persistent	21	89	80.9%
Overall %age		90.2%	80.9%	87.1%
Included				
Observed	Fluent + Recovered	189	26	87.9%
	Persistent	32	78	70.9%
Overall %age		86.5%	75.0%	82.2%

used SSI-3 with WWR excluded and that which used SSI-3 with WWR included (column two of rows five and six of Table 1). The Cox and Snell's R^2 and Nagelkerke's R^2 fit statistics were better when SSI-3 scores were calculated with WWR excluded than when they were included. Exp(b) and its confidence interval for the models being discussed are given in columns five and six of rows five and six of Table 1.

The classifications made by the two models are shown in Table 4. Overall classification accuracy was 87.1% and 82.2% when SSI-3 was used with WWR excluded and included, respectively. Correct classification of the fluent children plus CWS who would later recover was 90.2% and classification of persistent CWS was 80.9% when SSI-3 was used with WWR excluded, whereas classifications of fluent children and CWS who would later recover was 87.9%, and of persistent CWS was 70.9% when SSI-3 was used with WWR included. Classification of persistent CWS, in particular, deteriorated when SSI-3 was calculated with WWR included (from 80.9% to 70.9%). This would have arisen if CWS who would later recover became more similar in severity level to persistent CWS when SSI-3 was calculated with WWR included.

Overall performance at distinguishing persistent CWS from combined fluent children and CWS who would later recover was better (by nearly 5%) when SSI-3 scores were calculated with WWR excluded. This suggests that fluency symptoms other than WWR are important for separating fluent children and CWS who would later recover from persistent CWS. This is consistent with the Howell and Davis (2011) view that the symptoms included in the published form of SSI-3 predict persistence of stuttering. The model that used SSI-3 with WWR excluded was significantly better than that with WWR included ($\chi^2 = 31.49$, df = 1, $p < .001$).

The misclassifications in Table 4 appear to be symmetric; there is a similar amount of confusion involving fluent children and CWS who would later recover being called persistent CWS as there is involving persistent CWS being wrongly categorized as fluent children or CWS who would later recover. Although the level of confusion was greater for SSI-3 calculated with WWR included, the pattern just discussed also occurred with SSI-3 calculated with WWR excluded.

3.2.3. Screening for fluency using SSI-3 severity measures obtained with WWR excluded and included

Next the SSI-3 scores with WWR excluded or included were examined as predictors when fluent children were compared with combined CWS who would later recover and persistent CWS. The null model put all cases into the combined CWS who would later recover and persistent CWS class (68.0% correct), which is the benchmark for estimating whether SSI-3 scores improved the prediction. The models that incorporated a form of SSI-3 (excluded and included in rows seven and eight, respectively) were both significantly better than the null model. As in previous analyses, Cox and Snell's R^2 and Nagelkerke's R^2 fit statistics were better when SSI-3 scores were calculated with WWR excluded than when they were included. Exp(b) and the confidence interval around it are given in columns five and six of rows seven and eight in Table 1.

The classifications are shown in Table 5. Overall classification accuracy was 89.2% when SSI-3 was calculated with WWR excluded and 88.3% with WWR included. There were more fluent children misclassified into the persistent CWS and CWS

Table 5
Classification into fluent versus CWS who would recover later and persistent CWS when WWR were excluded from SSI-3 counts (top part) and when WWR were included in SSI-3 counts.

		Predicted		
		Fluent	Persistent + Recovered	% correct
Excluded				
Observed	Fluent	84	20	80.8%
	Persistent + Recovered	15	206	93.2%
Overall %age		84.82%	91.5%	89.2%
Included				
Observed	Fluent	80	24	76.9%
	Persistent + Recovered	14	207	93.7%
Overall %age		85.1%	89.6%	88.3%

Table 6
Age group analysis as CWS (top) or fluent speakers (bottom).

	Number classed as CWS	Number classed as fluent	%age correctly classified
Age (in years)			
5–6	29	1	96.7
8 to <10	23	7	76.7
10 to <12	54	6	90
12 to <14	96	19	83.5
14 to <16	35	11	76.1
16 to <18	18	3	85.7
Overall	255	47	84.4
Fluent			
5 to <9	2	10	83.3
9 to <11	1	12	92.3
Overall	3	22	88.0

who would later recover group than vice versa when WWR were excluded and included in the SSI-3 scores. There were more of the confusions between the fluent children group and the combined persistent CWS and CWS who would later recover group when WWR were included than when they were excluded.

Correct classification of fluent children was 80.8% and for the combined CWS who would later recover and persistent CWS was 93.2% when SSI-3 scores were calculated with WWR excluded. Correct classification of fluent children was 76.9% and for the combined CWS who would later recover and persistent CWS was 93.7% when SSI-3 scores were calculated with WWR included. The model with SSI-3 calculated with WWR excluded was significantly better than that which used SSI-3 calculated with WWR included ($\chi^2 = 32.255$, df = 1, $p < .001$).

Comparison of model fits for SSI-3 calculated in equivalent ways across Tables 4 and 5 showed that classifications from Table 5 (screen for fluency) was significantly better than the classifications in Table 4 (screen for persistence). The χ^2 values were 53.644 for SSI-3 calculated with WWR excluded and 52.879 for SSI-3 calculated with WWR included (df = 1 and $p < .001$ in both cases).

To summarize, the high rates of correct classification for screening for fluency support adoption of this form of screening. Once again SSI-3 scores calculated with and without WWR, showed that performance was better when WWR were included and the screen for fluency was significantly better than that for persistence.

3.3. Validation of risk prediction model for screen for fluency with SSI-3 calculated with WWR excluded for CWS and fluent children across a range of ages

The screening for fluency model that used SSI-3 scores calculated with WWR excluded was validated. The screen for fluency was preferred over the screen for persistence for the reasons given in Section 3.2.3, and because persistent and recovered outcomes have not yet been confirmed for the CWS used in validation. The results from all previous analyses showed that models for prognosis and screening using SSI-3 without WWR gave better classification than those with WWR. This section addresses whether screening for fluency using SSI-3 scores with WWR excluded can be validated using independent data from CWS and fluent children. The SSI-3 score that separated fluent children from CWS was 13; fluent children can score up to 13 and be classified correctly.

The classification of each of the 272 CWS between the ages of 5 and 19 was examined using the SSI-3 cutoff value of 13 to separate fluent children from all CWS. The tallied data for different age ranges are given in Table 6. All 272 CWS were diagnosed as stuttering prior to this analysis, so these data were used to obtain % correct classification as CWS and % of misclassifications (Table 6). Correct classification as CWS was above 80% for all ages except 8–10 years (this was 76.7%) and 14 to <16 years (this was 76.1%). Similar analyses were performed for fluent children in the age range 5 to <9 years and 9–11 years. Correct classification of fluent children was above 80% for both age groups (83.3% and 92.3%, respectively).

4. Discussion

Analyses were reported for estimating prognosis, screening for: stuttering types (separate classification of fluent persistent CWS and CWS who would later recover), screening for persistence (classification of persistent CWS from the rest of the participants) and screening for fluency (classification of fluent participants from all CWS). There are five findings to highlight: first, the models for prognosis and screening only used SSI-3 as a predictor but, nevertheless, they were all successful at making their required classifications (sensitivity and specificity were almost always in excess of 80%); second, in all analyses, performance on models that used SSI-3 scores with WWR excluded was superior to that with them included; third, ability to predict prognosis was not affected when fluent classifications were also required; fourth, screening for stuttering types and fluency appeared promising, but screening for persistence was less successful; fifth, the validations showed that the screening model for fluency with WWR excluded from SSI-3 scores performed well with the CWS and fluent children in and out of the age range the model was developed for and offer some promise for screening school-aged children for fluency.

Prognosis (finding three), screening (finding four) and validation (finding five) are discussed further. The last part of the discussion addresses some general considerations that the findings have a bearing on, and some recommendations when the findings are applied in future work.

4.1. Prognosis

Howell and Davis's (2011) work showed that correct classification outcome as persistent CWS or CWS who would later recover of around 80% was obtained when SSI-3 was used as a predictor with WWR excluded. The current study confirmed that finding. It also extended the work by investigating whether including WWR in SSI-3 calculations, changed the accuracy of the predicted class membership (as persistent CWS or CWS who would later recover). The results of the first analysis where only persistent CWS and CWS who would later recover were used showed that classification performance was better when WWR were excluded from SSI-3 calculations than when they were included.

In the subsequent analyses three-way classification was performed (fluent children, persistent CWS and CWS who would later recover) and the results for the persistent CWS and the CWS who would later recover were selected and compared to the two-way classifications obtained previously (Section 3.2.1.1). These results showed that addition of the fluent children did not reduce classification performance for prognosis of persistent CWS and CWS who would later recover and performance with SSI-3 calculated with WWR excluded was again better than with WWR included. The latter finding adds to the growing body of evidence that show WWR differ from more typical stuttering symptoms that were reviewed in Section 1.2 and support Riley's (1994) and Wingate's (2001) view that WWR should not be considered as symptoms of stuttering.

4.2. Screening

Extension of the prognosis work to allow investigation of the earlier model's ability to screen children in different ways was made possible because Howell and Davis (2011) pointed out that their modeling approach could be extended to fluent children because it did not use variables that can only be measured on CWS. Howell and Davis also considered that starting with data from CWS in the age range eight-teenage was the preferred age group to develop models that distinguish forms of stuttering associated with persistent CWS and CWS who would later recover. The basic advantages in using this age range are that there are no imbalances between number of persistent CWS and CWS who would later recover and stuttering at teenage can be classified as persistent or recovered. The results from these models can be checked to see whether they maintain their performance at other ages and when they are required to classify samples that are representative of the actual population in terms of the frequencies of each of the speaker classes.

Three types of screening were attempted, each of which would serve a different health service requirement: These were to screen for stuttering types, for persistence and for fluency. All fits showed that those models that used SSI-3 scores as a predictor (both for those that excluded and included WWR) were significantly better than the null model and that they were satisfactory according to sensitivity and specificity criteria (performance on both these statistics was around or above 80%). In all cases, classification performance was better when WWR were excluded from SSI-3 counts than when they were included. Once again, as in the prognostic analyses, Riley's (1994, 2009) advice to exclude WWR from severity counts was supported.

Looking at this in more detail, across all the screening models (for stuttering types, persistence and fluency all with SSI-3 scores calculated with WWR excluded and included), overall correct classification performance ranged from 77.8% to 89.2%. Obviously, performance of the models needs to be reasonable accurate, but it may be useful to allow some leeway. Whilst it would be useful if rates toward the upper end could be achieved when screening instruments are applied in schools, 100% accuracy is probably not achievable and, arguably, is not desirable anyway. It is conceivable that the 10% of children misclassified as CWS may be at risk of stuttering or perhaps even some other communication disorder. It would be possible in principle to concentrate attention on this group of children. On the other hand, it should be noted that the levels of performance achieved here would result in a large number of fluent children called CWS because of the high incidence of fluent children in unselected samples. A question that needs to be debated is whether it is preferable to identify as many as10% or 20% of cases as CWS who do not currently stutter. A further question is whether convenient and practical methods can be established so screening assessments can be made in schools or clinics and whether they can achieve similar levels of performance. At the same time, it should be considered whether the symptom set used in SSI-3 ought to be modified so as to allow other communication disorders to be detected in the screen (see Section 4.5).

The next question considered is which screening model is preferred. Looking at the classification data first, the screening for fluency with WWR excluded showed 89.2% correct classification whereas that for persistence showed 87.1% correct classification. Performance of the corresponding models with WWR included was worse in both cases (88.3% and 82.2%, respectively), but the screen for fluency maintained its advantage. This suggests that the screen for fluency was best in terms of overall classifications. This result was backed up by comparisons of model fits across the two types of screen both for SSI-3 calculated with WWR excluded and included. Based on these statistics, the screen for fluency is favored specifically when SSI-3 scores are used with WWR excluded from %SS counts. This model was validated in Section 3.3.

Only the persistence and fluency screening models can be compared statistically (the screen for stuttering types uses multinomial logistic regression). However, in the analysis into three speaker groups itself, classification of fluent children using SSI-3 with WWR excluded, correctly screened almost 90% of the sample, and this dropped to less than 80% when SSI-3

scores were obtained with WWR included. Screening for fluency types can be attempted as more data on CWS who have known persistent and recovered outcomes become available.

A further issue to be considered is about the way classification of speakers changes between the different type of screen and how this may relate to the symptoms (WWR in particular) shown by the different speaker groups. The data in Table 3 show that the overlap between fluent children and CWS who would later recover was more marked when WWR were included than when they were excluded. This suggests that these two groups are less distinguishable when WWR were included. For the data of the CWS who would later recover in the same table, there was also greater overlap with the persistent CWS than with the fluent children whether WWR were excluded or included. This suggests that the CWS who would later recover are more similar to the persistent CWS than they are to the fluent children (the opposite of what was observed when fluent children were examined). Taking both these findings together, they suggest that the CWS who would later recover as a whole have affiliations with both the other groups (some CWS who would later recover are more similar to fluent children whilst others are more similar to persistent CWS at the time they were initially assessed). The CWS who would later recover group may merit more clinical and research attention.

The CWS who would later recover and persistent CWS were in different classes for the screening for persistence data in Table 4 (the former children were grouped with fluent children, whereas the latter was in a class of its own). When WWR were counted as stutters, overlap between these classes increased. This can be explained on the grounds that CWS who would later recover have a high number of WWR, but persistent CWS do not (Howell et al., 2011). This elevates the %SS of the CWS who would later recover that brings them closer to the persistent CWS that then increases their confusability. Looking at Tables 3 and 4 data together to see the impact of including WWR in SSI-3 counts, Table 3 data show that fluent children overlap with CWS who would later recover when WWR are included and the data for the screen for stuttering types in Table 4 show that CWS who would later recover overlap with persistent CWS in the screen for persistence. Including WWR in SSI-3 counts is deleterious in different ways in the two cases.

Comparison of the data in Table 4 (screen for persistence) and Table 5 (screen for fluency) show that the former has higher overall misclassification rates. Also, the misclassifications in Table 4 occur symmetrically (fluent children and CWS who would later recover are as confusable as vice versa) whereas Table 5 shows an asymmetry (fluent children are more confusable with all CWS than vice versa). These data suggest that grouping fluent children and CWS who would later recover together loses some power in classifying fluent children and is not advisable bearing in mind that there are many more fluent children than CWS. The screen for fluency is preferred. Also, whereas the screen for stuttering types (Table 3) looked promising, the screen for persistence that also uses prognostic outcomes appears unadvisable specifically because it groups fluent children with CWS who would later recover whereas they should be kept separate.

Table 5 data are also consistent with the conclusion that fluent children and CWS who would later recover should be kept separate. The overall 3.9% poorer classification of fluent children when SSI-3 scores were calculated with WWR included for the screen for fluency (Table 5). This suggests an influence of high rates of WWR in fluent children, which increased the confusability between them and the CWS who would later recover and, since the latter category is combined with persistent CWS in this analysis, would affect classification of all CWS.

4.3. Validation of performance of the model for screening for fluency with WWR excluded from SSI-3 counts in the target age range and beyond

The model that screened for fluency with WWR excluded was selected for validation as it gave better performance than the equivalent screen for persistence model. Another advantage of this model is that it does not require prognostic outcome data (persistent versus recovered) which are not currently available at present (i.e. the screen for persistence and that for stuttering types cannot be evaluated at present). For this model, an SSI-3 score of 13, obtained according to the current procedures, separated fluent children from CWS. The screen for fluency was validated and performance for the majority of age groups and for fluent children as well as CWS was almost always above 80% (a conventional level for acceptance of a model). A feature of particular note concerning the CWS is that performance held up when CWS outside the range for which the model was developed were examined. In particular, performance was good down to age five which is nearer the age for onset of stuttering than the age range used for setting up the model (the same applied to children older than the upper age used in model development). The good performance down to age five supports the use of this form of screening with children at the age at which they start school. It would be useful to examine CWS at younger ages in future work in the same way. Validation out of the age range for which the model was developed is not only a strong test of a model, but also, more importantly, if validations are successful, would allow the model to be used to screen fluent children so as to identify any potential CWS at any age (particularly under-fives). More limited validations on fluent children showed that the performance in different age groups held up and gave correct classifications of 83.3% and 92.3% for children aged 5 to <9 years and 9–11 years. Thus, the data on fluent children show that the level of performance with CWS was not attained at the expense of misclassifying fluent children as CWS.

The other forms of screening may merit examination when there are more validation data where persistent and recovered outcomes are known. Additional work is necessary to see whether prognosis can be estimated out of the age range of those models as well. The results so far show that a form of SSI-3 might be useful for screening general samples of children for stuttering down to school-entry age. Some provisos about how this would need to be performed that mainly concern the specific assessment methods employed here are considered under recommendations.

4.4. General implications

The findings have implications for other work that shows factors not examined here differed between persistent CWS and CWS who would later recover or between fluent children and CWS in general. Although the success of severity as measured by SSI-3 as a predictor of prognosis and screening has been emphasized, the model did not have 100% sensitivity and specificity performance, so there is scope for improvement. A screening instrument that misclassifies between 10 and 20% of fluent children as CWS is going to cause parents unnecessary concern in many cases where there is an indication that their child stutters unless they have access to clinical services. The opposite point of view was also discussed in Section 4.2. That is, some latitude might be useful when classifying stuttering in children (i.e. that 100% accuracy in identifying fluent children may not be desirable so that children at risk who have not started to stutter can potentially be identified). If this is the case, it may be important to place more emphasis on ensuring that no CWS is called fluent (as achieved in the screen for fluency) than on calling a fluent child as one who stutters. Other severity thresholds than those used here could be explored that trade off hits (correctly identification of CWS) and false alarms (i.e. call a fluent child a CWS) that change performance in any agreed way that is desired. To achieve this, some optimization of these thresholds would need to be performed.

Another general issue concerns the view that there is overlap between young fluent children and CWS at an early age causing CWS to be assigned to the fluent category at the earlier ages (Bloodstein & Bernstein-Ratner, 2007). Whilst confusability between fluent children and CWS who would later recover support this when WWR are included in SSI-3 scores, reasons for separating fluent children and CWS who would later recover were given in Section 4.2. Further work is necessary to resolve this question.

4.5. Recommendations

Screening school-age children is achievable at reasonable rates of sensitivity and specificity for all three types of screen. The results could be used in clinics if the current procedures are adopted. However, these procedures are probably unrealistic to use in general screens of, for example, at school intake. Further work needs to be performed to establish related procedures based on these ideas, which could be used with large samples of children, possibly including some with diverse forms of communication disorder. Whatever procedure is adopted, agreement is needed about what the goals are in terms of identifying either large or small numbers of CWS and ways of optimizing thresholds in order to maximize performance on these classes need to be explored. Also, a decision needs to be made as to whether screening into three or two groups (and if the latter is the case whether one of the groups should be fluent children or one group should be persistent CWS). The performance of the models with respect to Riley's flexible assessments that he allowed in SSI-3 needs to be explored so that workable and cost-effective protocols for screening children can be obtained.

The model may or may not work for children very close to onset (Yairi & Ambrose, 1999). One argument that suggests it would not work with such children was given by Yairi and Ambrose (1999, p. 1106). They stated: "It is clear, however, that the initial level of disfluency does not distinguish between the two groups [CWS who would later recover and persistent CWS]. In fact, the children who later recovered exhibited slightly more disfluencies on initial evaluation. Significantly, although in many cases complete recovery occurs a year or two later, the sharp reduction in SLD [stuttering-like disfluencies] for the recovered group, as opposed to the relative stability of SLD for the persistent group, creates a large gap as soon as 12 months post-onset. The departure of the curves at that point may serve as an important prognostic feature" (Yairi & Ambrose, 1999, p. 1106). Their data suggest that one predictor of persistence is the failure of a child to reduce sharply his/her stuttering-like disfluencies (what are called here stutters plus WWR) over the first year after onset, not his/her initial severity near onset. The point is repeated in Section 4: "There is no indication whatsoever that children who recovered had initially milder stuttering. To the contrary, the data show a slight tendency toward more severe stuttering initially in the recovered group." (Yairi & Ambrose, 1999, p. 1108). A reason that the current model may work with such data is that part of the reason for the high severity levels near onset may be because children with high rates of WWR would be considered to stutter according to Yairi and Ambrose's procedure. The evidence reviewed in the introduction show that high rates of WWR are an indication of recovery. This suggests that the link between recovered/persistent outcome and SSI-3 in very young children should be re-examined with %SS calculated with WWR excluded. A specific prediction based on what has been stated here is that the CWS who would later recover will have high rates of WWR in comparison to CWS who persist around the age of onset of stuttering.

It was assumed in the current study that a model with no imbalances between numbers of fluent, recovered CWS and persistent CWS would scale up in proportion to incidence of fluency in the general population to estimate the chance of stuttering in a random sample. Whilst the validation results confirm this assumption, it needs to be checked, as using balanced groups is the most convenient way of developing a model. One possibility would be to process Reilly et al.'s. (2009) or other population-cohort data through the current model using the present procedures. This would also allow comparison of outcomes that use their assessments of stuttering and the current ones.

As well as deciding what a screening instrument should be used for, which determines what percentage of cases are located after screening, research need to address the approximate 10% of cases of fluent children who are called CWS. It is necessary to establish whether or not these are high risk cases in follow-ups. If this is so, allowing some level of misclassification might be a useful thing as it would provide clinicians with an objective way of identifying children at risk of subsequent stuttering.

A related question concerns the nature of the children deemed to stutter who do not. Do these children have other communication problems and, if so, how well does SSI-3 separate all such disorders? On the face of it, it would seem unlikely that the symptom set in SSI-3 is appropriate for all communication problems (e.g. those seen in children with apraxia, etc.) and the symptom set could be improved so that it performs a general communication screen.

5. Conclusion

The findings show that Howell and Davis's (2011) model for predicting risk of stuttering can be successfully adapted to classify fluent children. This offers the possibility of it acting as a screening instrument for stuttering with children at the age of school intake. The model is successful at classifying CWS outside the age range for which it was originally developed.

CONTINUING EDUCATION

Screening school-aged children for risk of stuttering

QUESTIONS

1. Diagnostic screening models are used to:
   (a) Separate children to be classified as fluent, likely to recover, likely to persist
   (b) Determine which CWS will persist in a sample that includes fluent children and CWS who recover
   (c) To estimate which children in a similar sample are fluent
   (d) All of the above
   (e) None of the above
2. Risk factor modeling is used in this study for:
   (a) Prognosis
   (b) Diagnosis
   (c) Screening
   (d) None of the above
   (e) All of the above
3. Whole-word repetitions improve the following assessments:
   (a) None
   (b) Diagnosis
   (c) Prognosis
   (d) Diagnosis and prognosis
   (e) Screening
4. The screening model separates fluent children from those who stutter in the following age ranges:
   (a) 8–10 years
   (b) 8-teenage
   (c) Ages 5–18 years
   (d) Above teenage
   (e) At no age
5. The main statistical method employed in this study was:
   (a) Analysis of variance
   (b) Logistic regression
   (c) Factor analysis
   (d) Analysis of covariance
   (e) Mann–Whitney *U* test

Appendix A. Description and assessment of Riley's stuttering severity instrument

A.1. General scope and applications

The stuttering severity instrument (SSI) was developed by Riley over the course of several years. It is currently in its fourth revision (Riley, 2009). The main version discussed here is the third one (SSI-3) as described in Riley (1994). The relationship of this to the fourth version (SSI-4) and the reasons SSI-3 was selected for discussion are given later in this appendix. The ways SSI-3 was assessed statistically are described below. SSI-3 has been used to report details of stuttering participants in more than 350 publications and it has also been translated into several other languages.

As well as its use for characterizing participants in experiments, 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. As an example of the latter, Howell et al. (2009) used SSI-3 to validate the child,

parent and researcher assessments that were subsequently used for classifying participants aged between eight years and teenage as persistent or recovered (referred to in Section 2.1.1).

A.2. Administration of SSI-3

A.2.1. Overview

SSI-3 can be administered to all ages. For assessments of any individual, scores for frequency and duration of stuttering events, and a measure of physical concomitants have to be obtained and these are combined to produce an overall SSI-3 score. The ways these three measurements are collected from non-readers and readers are described.

Assessments for children who cannot read are based on elicited speech samples alone. This is an example of the flexibility SSI-3 has (referred to in point four in Section 1.1), although this flexibility also has drawbacks. Another example of its flexibility is that assessments can be based on audio or video recordings. An illustration of a problem that arises in this case is that the norms for SSI-3 and SSI-4 were developed based on audio recordings. Consequently, the norms are not directly applicable to video recordings without restandardization even though video records are preferable in some ways. Video recordings have the advantage that they provide a permanent recording of body movements associated with stuttering, for assessment of some of the physical concomitants. When audio recordings alone are made, the examiner has to make notes about any physical concomitants at the time they are collected.

Riley advises on the collection of speech data. With children who cannot read, for example, stimulus pictures where something appears wrong or out of place or where an accident is about to occur are used to elicit speech. Leading statements about the image are made to encourage the child to engage in conversation. Questions, interruptions, and mild disagreements that simulate features that occur in normal conversation are allowed. Each speech sample has to be at least 200 syllables long. Riley recommends that several speech samples should be obtained. This would include read and conversational speech in the case of readers; the fourth revision recommends obtaining additional forms of speech samples. This raises the issue of whether the norms were developed for all the types of material, which was not the case (see the end of this appendix for further discussion of this point). When the norms were obtained with particular types of speech, they only apply to those specific forms. Consequently, forms of speech not used in standardization should not be used (in contrast to what Riley recommends). The audio tracks from video or audio tapes are used to determine the frequency and duration measures of stuttering. Guidelines are given on how to translate these into scale scores.

A.2.2. Frequency score

A.2.2.1. Frequency score for non-readers. The percentage of stuttered syllables (%SS) is obtained first. Separate counts of all syllables spoken and those syllables that are stuttered are obtained. Riley notes that it is difficult to listen for syllables and count them together in real time. In SSI-3, he advocated that assessors should listen and make a dot on a piece of paper for each syllable first and count them up afterwards. SSI-4 includes software that permits both syllable and stutter counts to be obtained simultaneously and the duration of stutters to be obtained. The method suggested in SSI-4 simultaneously counts syllables on one key of a mouse and stuttered syllables and their durations on the other key (duration is measured by holding the key down for each stutter). This is a complicated procedure. Even minor changes in procedure affect syllable and stutter counts (Howell et al., 2011). Since the dual syllable and stutter counting system has not been assessed for reliability and the norms were obtained with the simpler procedure, SSI-3 norms would not apply if the dual-counter was used. More generally, the same procedure ought to be used when results are compared as there is no guarantee that the flexible procedures employed give corresponding results. In the research reported in the article, recordings were transferred to a PC and the counts were made with the assistance of a wave editor. The editor allows short extracts to be played, so the syllables in them can easily be counted and checked if necessary. It is probably impractical to use this procedure when assessments are made in clinics.

The events counted as stutters are defined as "repetitions or prolongations of sounds or syllables (including silent prolongations)" (Riley, 1994, p. 4). The SSI-3 also notes 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).

Stuttered repetitions of a syllable count as a single stuttering event. For example, "I don't wa, wa, wa, wa, want any pepper," has seven syllables and one stuttering event. The stuttering events are expressed as a percentage of all syllables which gives %SS. When there are multiple samples of the same speech type, the %SS is computed for each speaking sample, and then averaged and the latter is used as the raw %SS. As indicated, Riley (1994) is precise in his definition of how to count stutters and syllables. These may be important features behind its success as a prognostic (Howell & Davis, 2011) and screening instrument (current article). The frequency score is converted to a task score using the left-hand side of Table A.1 (appropriate for non-readers).

A.2.2.2. Frequency score for readers. The procedure for participants who can read is similar except that there is an additional reading task. Reading material appropriate for 8–9 year olds, 10–11 year olds, 12–13 year olds and adults are supplied for English participants (each text is approximately 200 syllables). Percentage of stuttered syllables (%SS) is computed in the

Table A.1
The %SS score is used to obtain the frequency task score. The scores for children who cannot read (non-reader) are given at left (frequency score based on elicited speech samples), and for people who can read at the right (frequency score based on elicited speech samples and readings of supplied text).

Non-readers		Readers			
Speaking		Speaking		Reading	
%SS	Task score	%SS	Task score	%SS	Task score
1	4	1	2	1	2
2	6	2	3		
3	8	3	4	2	4
4–5	10	4–5	5	3–4	5
6–7	12	6–7	6	5–7	6
8–11	14	8–11	7	8–12	7
12–21	16	12–21	8	13–20	8
22 & up	18	22 & up	9	21 & up	9

manner described earlier. The difference in procedure is that separate frequency scores are obtained for elicited and read materials (each of these is obtained from the appropriate columns in the right hand section of Table A.1). Elicited and read results are *added* to give the overall frequency score. Table A.1 shows that summing speaking and reading scores for readers is equal to speaking scores for non-readers. This indicates that these two forms of speech alone were used to obtain task scores for these two groups of speakers. Other points to note about the task scores obtained from %SS are that more weight is given when speakers have low %SS and that there is no specified way of incorporating three or more forms of speech (e.g. reading, spontaneous monolog and recordings made on the telephone) either in SSI-3 or SSI-4.

A.2.3. Duration score

The duration score is the time, in seconds, of the three longest stuttering events in the sample. This can be accurately measured from audio files that are transferred to computer. A display with a traveling cursor and calibrated timeline is then used to measure duration (Riley says that a stopwatch can be used when computers are not available). These three durations are averaged. Once the average duration has been computed, it is converted to a scale score using Table A.2. Durations of less than 1 second are difficult to measure accurately by stopwatch (though not by computer). They are designated either "Fleeting" or "One-half second" and receive a duration scores of 2 or 4 points, respectively

Using the frequently employed frequency score of 3% SS (or SLD) as a criterion for stuttering/not stuttering, this corresponds to a task score of 8. In SSI-3, the duration score can make an additional contribution varying from +25% (2 points) to +225% (+18 points). These extra points may be important in classifying children in different ways for the various screens examined in the paper. The duration scores are based on speech and reading samples for children who can read.

A.2.4. Physical concomitants

The physical concomitants score is based on observations of all of the speaking samples that are scored (recorded at the time, or obtained from videotapes afterwards). Four aspects of physical concomitants are assessed; Riley's descriptions of each of these are as follows (Riley, 1994, p.11).

"**Distracting Sounds**: This category includes any non-speech sounds that accompany the stuttering. For example, the child may continually clear his or her throat or may swallow. Other common sounds include noisy breathing, whistling noises, sniffing, blowing and clicking sounds. The evaluator must determine the extent to which these sounds are distracting to a listener.

Facial grimaces: Any abnormal movement or tension about the face counts in this category. Examples of abnormal facial behaviors are pressing the lips together tightly, pursing the lips, tensing jaw muscles, blinking or partially closing the eyes, protruding the tongue, and uncoordinated jaw movements.

Head Movements: These movements generally consist of turning the head away from the listener to avoid eye contact, looking down at the feet, scanning the room and looking up at the ceiling.

Table A.2
The average of the three longest stutters is obtained and scored on the following scale to give the duration score.

Fleeting	.1–.4 s	2 points
One-half second	.5–.9 s	4 points
One full second	1.0–1.9 s	6 points
2 s	2.0–2.9 s	8 points
3–5 s	3.0–4.9 s	10 points
5–9 s	5.0–9.9 s	12 points
10–29 s	10.0–29.9 s	14 points
30–59 s	30.0–59.9 s	16 points
60 s or more	60.0 and up	18 points

Table A.3
Auditory and visual physical concomitants associated with stuttering are each scored on the following scale to give the physical concomitants score.

0 = none
1 = not noticeable unless looking for it
2 = barely noticeable to casual observer
3 = distracting
4 = very distracting
5 = severe and painful-looking

Movements of the extremities: Any general body movement such as shifting around in the chair counts in this category. Other common movements include specific movement of a limb such as foot tapping, excessive movement of the hands about the face, fidgeting with something in the hand, or swinging an arm."

A separate judgment is made for each anatomical area (face, head, extremity), and for distracting sounds. Each of the four judgments is scored between 0 and 5 using the scale in Table A.3. The four scores are added to give an overall physical concomitant score between 0 and 20.

It is not clear how the weighting between %SS, duration and physical concomitant scores was derived. Physical concomitants can make an additional contribution relative to a frequency score of 3%SS of between 0 and 250%. In terms of absolute SSI-3 scores, the physical concomitant scores alone can give a score of up to 20 points which would allow a child who cannot read who shows no overt stutters but who has all physical concomitants scoring at maximum to be designated "moderate" severity (Table A.4).

Physical concomitants are the most problematic aspect of SSI-3. They are least reliable (Bakhtiar, Seifpanahi, Ansari, Ghanadzade, & Packman, 2010; Lewis, 1995). Riley's (1994) reason for retaining them was that the inter-correlations between the three main sub-components alone were lower than when all were used. This indicated none of the parameters used alone would produce the same indications about severity as the combined SSI-3 score. Procedures for assessing physical concomitants need to be improved to see whether training and methodical application of the advice can improve the results (the procedures are probably not comparable across labs at present). It is also necessary to establish the impact these scores make to diagnosis, prognosis, screening and treatment outcome. Physical concomitants scores are based on speech and reading samples for readers and just speech samples for non-readers.

A.2.5. Total overall score

The total overall score is obtained by adding together the scores for the three component elements (frequency, duration, and physical concomitants). Severity of stuttering, as measured by these parameters, can be ascertained by comparing this score to the age-appropriate normative data presented in Table A.4 (left for non-reading children and right for children who can read). For both parts of the table, the score for a person being tested can be converted into an indication of five categories that describe the level of stuttering severity as very mild, mild, moderate, severe or very severe. Thus, using the left part of Table A.4, a child who cannot read who scores 14 can be characterized as mild severity (further tables are available for adults who stutter but these are not given here). There is no category for not stuttering, which is required when fluent cases are included in classifications. In fact a child who cannot read who has no stutters or physical concomitants (task score of 0) can only be classed, at minimum, as "very mild".

A.3. Riley's statistical assessment of the SSI-3

With regard to questionnaires in general, construct reliability refers to the extent to which the instrument is an effective measure, whilst validity is the extent to which the questionnaire is measuring what it purports to measure. If a questionnaire is unreliable it is impossible for it to be valid; on the other hand it is possible for a questionnaire to be reliable but invalid.

A.3.1. Intra-judge reliability of the SSI-3

A satisfactory test instrument will give similar results when a person uses it on the same data on different occasions. There are different ways of quantifying how close the scores are when a test is repeated. Riley (1994) reported intra-judge

Table A.4
Percentile and severity equivalents of SSI-3 total overall scores for non-reading and reading children based on Riley (1994). Results were obtained from a group of CWS who could not read ($N = 72$) and a group of CWS who could read ($N = 139$).

Non-readers		Readers	
Total overall SSI-3 score	Severity	Total overall SSI-3 score	Severity
0–10	Very mild	6–10	Very mild
11–16	Mild	11–20	Mild
17–26	Moderate	21–27	Moderate
27–31	Severe	28–35	Severe
32 and above	Very severe	36 and above	Very severe

reliability of the SSI-3. Intra-judge reliability was estimated for frequency and duration by five judges who scored 17 samples twice each. Self-agreement ranged from 75.0% to 100% across the two scores. All mean agreements were above 80%.

A.3.2. Inter-judge reliability of the SSI-3

Different judges should give similar results when assessing the same samples. Once again there are various ways of assessing this. Riley assessed inter-judge reliability for frequency, duration, physical concomitants and overall SSI-3 score using 15 students from a graduate-level class in stuttering. Agreement for frequency ranged from 94.6% to 96.8%, but was somewhat poorer for duration (58.1–87.2%), physical concomitants (59.8–97.5%) and overall SSI-3 score (71–100%). However, all mean agreements were again above 80%.

A.3.3. Criterion validity of the SSI-3

Criterion validity is a measure of how well a set of variables predicts an outcome based on information that relates to other measures known to relate to the behavior in question. Riley showed that total SSI-3 scores correlated with the %SS frequency scores (the latter has often been used as a severity measure in the past). Correlations for all age groups were significant. This would have been expected, as SSI-3 itself includes a %SS measure. It has also been reported that SSI-3 scores correlated with scores from the Stuttering Prediction Instrument (Yaruss & Conture, 1992). Both these findings were considered to indicate adequate criterion validity.

A.3.4. Construct validity of the SSI-3

It is also advisable to check the internal consistency of the components in a questionnaire (construct validity). One way in which Riley assessed construct validity was by looking at the correlations between total SSI-3 scores and its separate components (frequency, duration and physical concomitants). All the correlations between overall scores and the components were significant, indicating good construct reliability.

A.4. Final comments on SSI-3 and SSI-4

Riley's own statistical assessments have been presented. Lewis (1995) has evaluated SSI-3 statistically less favorably. As mentioned, SSI-3 and SSI-4 have been translated into other languages (Bakhtiar et al., 2010). The procedures that are necessary for this are demanding (see Karimi, Nilipour, Shafiei, & Howell, 2011).

SSI-4 differs from SSI-3 in three respects: (1) the fourth revision includes the computer program that automates the assessment of stuttering severity that was discussed above (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 stuttering severity recommended in SSI-3. The latter methods were used to obtain the norms and new norms would be required if the computer program produced 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 such as 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 and these are also used in SSI-4. The norms would not apply if the additional material was included (Riley, 2009). 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, thus making the assessments equivalent to SSI-3. To emphasise this, the assessments are referred to as SSI-3 although assessments made in this way are permitted in SSI-4.

The data reported in the current article show that, despite several problems, SSI-3 performs well with respect to screening, Howell and Davis (2011) show it performs well with prognosis and Cook et al. (2012) show how it can be used to follow treatment outcome.

References

Ajdacic-Gross, V., Vetter, S., Muller, M., Kawohl, W., Frey, F., Lupi, G., et al. (2009). Risk factors for stuttering: A secondary analysis of a large data base. *European Archives of Psychiatry and Clinical Neuroscience*, *260*(4), 279–286.

Anderson, J. D., & Wagovich, S. A. (2010). Relationships among linguistic processing speed, phonological working memory, and attention in children who stutter. *Journal of Fluency Disorders*, *35*, 216–234.

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*, 87–91.

Bloodstein, O., & Bernstein-Ratner, N. (2007). *A handbook on stuttering* (6th ed.). Clifton Park, NY: Thomson Delmar.

Cook, S. P., Howell, P., & Donlan, C. (2012). Stuttering severity, psychosocial impact and language abilities in relation to treatment outcome in stuttering. *Journal of Fluency Disorders*, http://dx.doi.org/10.1016/j.jfludis.2012.08.001

Cox, D. R., & Snell, E. J. (1989). *77ie analysis of binary data* (2nd ed.). London: Chapman and Hall.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Howell, P. (January 17, 2009). Predicting stuttering onset/[E-letter], pediatrics. http://pediatrics.aappublications.org/cgi/eletters/123/1/270 Accessed 17.01.09.

Howell, P. (2010). *Recovery from Stuttering*. New York: Psychology Press. ISBN-13: 978-1-84872-916-2

Howell, P., Bailey, E., & Kothari, N. (2010). Changes in the pattern of stuttering over development for children who recover or persist. *Clinical Linguistics and Phonetics*, *24*, 556–575.

Howell, P., & Davis, S. (2011). Predicting persistence of and recovery from stuttering at teenage based on information gathered at age eight. *Journal of Developmental and Behavioral Pediatrics*, *32*, 196–205.

Howell, P., Davis, S., & Williams, R. (2009). The effects of bilingualism on speakers who stutter during late childhood. *Archives of Disease in Childhood*, *94*, 42–46.
Howell, P., Soukup-Ascencao, T., Davis, S., & Rusbridge, S. (2011). Comparison of alternative methods for obtaining severity scores of the speech of people who stutter. *Clinical Linguistics and Phonetics*, *25*, 368–378.
Hubbard, C., & Yairi, E. (1988). Clustering of disfluencies in the speech of stuttering and normal speaking preschool children. *Journal of Speech and Hearing Research*, *31*, 228–233.
Jani, L., Huckvale, M., & Howell, P. Measurement of stuttering frequency and their duration depends on the procedures used to assess them. *Journal of Fluency Disorders*, submitted for publication.
Jiang, J., Lu, C., Peng, D., Zhu, C., & Howell, P. (2012). Classification of types of stuttering symptoms based on brain activity. *PLoS One*, *7*(6), e39747. http://dx.doi.org/10.1371/journal.pone.0039747
Karimi, H., Nilipour, R., Shafiei, B., & Howell, P. (2011). Translation, assessment and deployment of stuttering instruments into different languages: Comments arising from Bakhtiar, et al., Investigation of the reliability of the SSI-3 for preschool Persian-speaking children who stutter [J. Fluency Disorders 35 (2010) 87–91]. *Journal of Fluency Disorders*, *36*, 246–248.
Lewis, K. E. (1995). Do SSI-3 scores adequately reflect observations of stuttering behaviors? *American Journal of Speech-Language Pathology*, *4*, 46–59.
Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.
Reed, P., Howell, P., Davis, S., & Osborne, L. (2007). Development of an operant procedure for content word characteristics in persistent stuttering children: Initial experimental data. *Journal of Stuttering Therapy, Advocacy and Research*, *2*, 1–13.
Reed, P., & Wu, Y. Logistic regression in stuttering research. *Journal of Fluency Disorder*, http://dx.doi.org/10.1016/j.jfludis.2012.09.003, in press.
Reilly, S., Onslow, M., Packman, A., Wake, M., Bavin, E. L., Prior, M., et al. (2009). Predicting stuttering onset by the age of 3 years: A prospective, community cohort study. *Pediatrics*, *123*, 270–277.
Riley, G. D. (1994). *Stuttering severity instrument for children and adults (SSI-3)* (3rd ed.). Austin, TX: Pro Ed.
Riley, G. D. (2009). *Stuttering severity instrument for children and adults (SSI-4)* (4th ed.). Austin, TX: Pro-Ed., Inc.
Throneburg, R. N., Yairi, E., & Paden, E. P. (1994). Relation between phonological difficulty and the occurrence of disfluencies in the early stage of stuttering. *Journal of Speech and Hearing Research*, *37*, 504–509.
Ujihira, A. (2011). Stuttering in Japanese. In P. Howell, & J. van Borsel (Eds.), *Fluency disorders and language diversity. Communication disorders across languages* (pp. 145–174). Bristol, England: Multilingual Matters.
Wingate, M. E. (2001). SLD is not stuttering. *Journal of Speech Language, and Hearing Research*, *44*, 381–383.
Yairi, E., & Ambrose, N. G. (1999). Early childhood stuttering. I: Persistency and recovery rates. *Journal of Speech and Hearing Research*, *42*, 1097–1112.
Yairi, E., & Ambrose, N. G. (2005). *Early childhood stuttering*. Austin, TX: Pro-Ed.
Yairi, E., Ambrose, N. G., Paden, E. P., & Throneburg, R. N. (1996). Predictive factors of persistence and recovery: Pathways of childhood stuttering. *Journal of Communication Disorders*, *29*, 51–77.
Yairi, E., & Lewis, B. (1984). Disfluencies at the onset of stuttering. *Journal of Speech and Hearing Research*, *27*, 154–159.
Yaruss, J.S., & Conture, E.G. (1992). Relationship between mother-child speaking rates in adjacent fluent utterances. (Abstract). *Asha*, *34*, 210. (Technical paper presented at the annual convention of the American Speech-Language-Hearing Association, San Antonio, TX).

Peter Howell is Professor of Experimental Psychology at University College London. He became interested in stuttering in the 1980s when he was working on feedback processes. In 1987 he published the first results with frequency altered feedback at the inaugural conference of the Nijmegen series. He recently published a well-received book entitled “Recovery from Stuttering”.