Review

Event- and interval-based measurement of stuttering: a review

Ana Rita S. Valente†‡, Luis M. T. Jesus†§, Andreia Hall†¶ and Margaret Leahy∥
†Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Aveiro, Portugal
‡Department of Education (DE), University of Aveiro, Aveiro, Portugal
§School of Health Sciences (ESSUA), University of Aveiro, Aveiro, Portugal
¶Department of Mathematics, University of Aveiro, Aveiro, Portugal
∥Department of Clinical Speech and Language Studies, Trinity College Dublin, Dublin, Ireland
(Received October 2013; accepted April 2014)

Abstract

Background: Event- and interval-based measurements are two different ways of computing frequency of stuttering. Interval-based methodology emerged as an alternative measure to overcome problems associated with reproducibility in the event-based methodology. No review has been made to study the effect of methodological factors in interval-based absolute reliability data or to compute the agreement between the two methodologies in terms of inter-judge, intra-judge and accuracy (i.e., correspondence between raters’ scores and an established criterion).

Aims: To provide a review related to reproducibility of event-based and time-interval measurement, and to verify the effect of methodological factors (training, experience, interval duration, sample presentation order and judgment conditions) on agreement of time-interval measurement; in addition, to determine if it is possible to quantify the agreement between the two methodologies

Methods & Procedures: The first two authors searched for articles on ERIC, MEDLINE, PubMed, B-on, CENTRAL and Dissertation Abstracts during January–February 2013 and retrieved 495 articles. Forty-eight articles were selected for review. Content tables were constructed with the main findings.

Main Contribution: Articles related to event-based measurements revealed values of inter- and intra-judge greater than 0.70 and agreement percentages beyond 80%. The articles related to time-interval measures revealed that, in general, judges with more experience with stuttering presented significantly higher levels of intra- and inter-judge agreement. Inter- and intra-judge values were beyond the references for high reproducibility values for both methodologies. Accuracy (regarding the closeness of raters’ judgements with an established criterion), intra- and inter-judge agreement were higher for trained groups when compared with non-trained groups. Sample presentation order and audio/video conditions did not result in differences in inter- or intra-judge results. A duration of 5 s for an interval appears to be an acceptable agreement. Explanation for high reproducibility values as well as parameter choice to report those data are discussed.

Conclusions & Implications: Both interval- and event-based methodologies used trained or experienced judges for inter- and intra-judge determination and data were beyond the references for good reproducibility values. Inter- and intra-judge values were reported in different metric scales among event- and interval-based methods studies, making it unfeasible to quantify the agreement between the two methods.

Keywords: stuttering, event-based measurement, interval-based measurement, inter-judge, intra-judge, accuracy.

What this paper adds?

A systematic review has been completed regarding event- and interval-based measurements of frequency of stuttering. Interval-based measurement was developed to overcome reproducibility problems from event-based methodology. This study was developed to verify if it is possible to quantify the agreement between the two methods to conclude what is the most reproducible or more accurate way to assess stuttering frequency and to analyse the effect of methodological factors on agreement of time-interval measurement. Both methodologies presented agreement values beyond the highest references for all metric scales. It was not possible to perform a comparison study between methods, so it would be useful in future research to quantify the agreement between the two assessment methods.
Event- and interval-based measurement of stuttering: a review

Introduction

Several authors have shown that therapy for adults who stutter (AWS) using behavioural treatment (e.g., fluency shaping and/or stuttering modification approaches) can be effective, resulting in a significant decrease in stuttering frequency (Andrews et al. 1980, Bothe et al. 2006, Herder et al. 2006). The frequency or percentage of stuttered syllables (%SS) or words (%SW) has been used to assess changes in fluency as a result of intervention. Stuttering behaviours that include part-word repetitions, prolongations and blocks (or other defined stutters) are considered as core behaviours of the disorder (Van Riper 1982) and these are determined and counted, and their occurrence reported as a frequency or percentage (Brundage et al. 2006). As an important part of the assessment researchers and clinicians suggest that reproducibility of frequency estimation should be discussed to develop ‘reliable and valid measurements [. . . ] for documenting treatment candidacy and outcomes, and for research and replications’ (Brundage et al. 2006: 272).

To assess the ‘degree to which repeated measurement in stable study objects, often persons, provide similar results’ reliability and agreement parameters were estimated (de Vet et al. 2006: 1033). The focus of the concepts were different: agreement (i.e., absolute reliability) refers to the closeness of repeated measurements (by different judges or by the same judge on various situations) concerning measurement error (Cordes 1994, Cordes and Ingham 1994b); reliability (i.e., relative reliability) explains how well subjects can be distinguished regardless of measurement error, and is related to dependability of obtained data or test scores (Cordes 1994). As an umbrella term, the word ‘reproducibility’ has been used to refer both concepts (de Vet et al. 2006). Percentage agreement, standard error of measurement (SEM) and limits of agreement are typical parameters for agreement, and intra-class correlation (ICC) is a common index for reliability reports (de Vet et al. 2006, Gisev et al. 2013).

Frequency counts in the assessment of stuttering are a popular outcome measure, with a widespread use. However, studies of listener agreement for the occurrence of stuttering events have reported unit-by-unit inter-judge agreement to be below 60% (Curlee 1981, Coyle and Mallard 1979, Emerick 1960, MacDonald and Martin 1973, Young 1975, Martin and Haroldson 1981, Martin et al. 1988). There is also evidence of inter-clinic disagreement for elements of frequency counts, including: total count of stuttering events (Kully and Boberg 1988, Ham 1989, Cordes et al. 1992); speech judged as stuttered based on perceptual definitions, i.e., threshold for stuttering is not fixed for any judge or for many judges; speech judged as stuttered based on different lists of disfluency types (Cordes 2000, Einarsdóttir and Ingham 2005, Young 1975). Such lack of agreement can lead to different diagnosis or severity description at different clinics, or even to different definitions of stuttering (Brundage et al. 2006). In addition, the relevance of estimating severity on the basis of frequency of stuttering events has been criticized by Leith et al. (1993) because of the lack of valid and reliable severity instruments, and because elements of stuttering (such as blocks accompanied by tension) may indicate the emotional involvement of people who stutter (PWS) when stuttering.

Other studies based on total counts of stuttering (Craig 1990, Gaines et al. 1991, Guitar et al. 1992, Lewis 1991, Onslow et al. 1990, Prins and Hubbard 1990, Zebrowski 1991) reported high estimates of inter- and intra-judge agreement (above 90%). Many problems have been demonstrated in frequency counts of the studies mentioned above, e.g., the high agreement for total counts of stuttering judgments produced high values, but many of these agreement estimates neither presented judgment agreement for total events counts, nor corroborated that the data and experimental analysis were unaffected by differences between and among judges. Such problems lead to the conclusion that differences between low agreement reported in stuttering measurement research and the high agreement reported in other stuttering research may be more apparent than real (Cordes and Ingham 1994b).

An alternative measurement system is based on behaviours occurring within a defined time-interval (Cordes et al. 1992). This methodological variation (time-interval analysis) was introduced by Schloss et al. (1987) and does not focus on individual stuttering events but on the presence of a stuttering event within a time interval (Alpermann et al. 2012). Alpermann et al. (2012) showed that the ease of judging intervals instead of individual events could make this type of measurement more feasible for clinicians.

In the time-interval methodology, video or audio samples of speech are divided into short intervals of time allowing judges to make a binary judgment (Brundage et al. 2006, Cordes et al. 1992). Intervals that contain at least one stuttering event judgment are denominated stuttered intervals and intervals with no stuttering event judgments are labelled as non-stuttered intervals (Cordes et al. 1992). In this context, stuttering events are defined as ‘whatever specified observers point to as stuttering’ (Bloodstein 1990: 392) based on a modified version of the perceptual definition of stuttering (Martin and Haroldson 1981, Bloodstein and Ratner 2008). Two additional supplements are added in order to control judgment idiosyncrasies: judges should only identify stuttered disfluencies; judges should consider that one stuttering event may encompass more than one syllable or more than one word (Cordes et al. 1992).
One important implication of time-interval analysis is that training based on this methodology increases agreement and accuracy (i.e., correspondence judgment of stuttered or non-stuttered interval with the criterion established by a group of authoritative judges (Cordes and Ingham 1999)) of inexperienced judges (Cordes and Ingham 1996, 1999, Einarsdóttir and Ingham 2005, Ingham et al. 1993b, Yaruss 1997, Cordes and Ingham 1994a), making judgments more replicable and precise (Cordes and Ingham 1994a). Ingham et al. (1998b) developed a software and hardware system based on time-interval methodology—Stuttering Measurement Assessment and Training (SMAAT)—that include judgment stimulus from PWS (adults) divided into 5 s intervals and as judged by 10 highly experienced authorities. The use of the SMAAT program improves inter-judge agreement, intra-judge agreement and accuracy in untrained judges (Cordes and Ingham 1999). Ingham et al. (1999) developed the Stuttering Measurement System (SMS), which is a standardized training programme to make real-time judgments who stutter in terms of speech rate, stuttering frequency and speech naturalness, in children and adults samples previously assessed by experts. The frequency data are computed in an event-based methodology, but the SMS program allows the option of calculating stuttering frequency using an interval-based methodology (the number of non-stuttered intervals can also be used to estimate the stutter-free speech rate).

A modified approach of time-interval analysis with three categories (fluent intervals, stuttered intervals and intervals with trained speech pattern), is currently being studied (Alpermann et al. 2010, 2012). The modified time-interval analysis was introduced by Natke and Alpermann (2010) to distinguish the use of trained speech patterns (acquired during stuttering therapy) from stuttered and fluent speech by means of time-interval analysis. Alpermann et al. (2010, 2012) concluded that modified time-interval analysis is a reliable and valid measurement technique.

Time-interval-based measurement is proposed as a focus of attention here, based on concerns regarding the agreement of counting instances of stuttering (Yaruss 1997). The comparison between a new measurement method (i.e., time-interval) and one previously proposed (i.e., event-based) is necessary to assess how this new methodology is likely to differ from the established one, by computing their agreement (Bland and Altman 2010, Bartlett and Frost 2008).

This paper presents a review of event-based measurement and time-interval analysis in terms of reproducibility data, as well as a discussion related to the comparison between the two methodologies. It also presents a review regarding the effect of different methodological factors (i.e., training, experience, interval duration, sample presentation order and judgment conditions) on inter- and intra-judge agreement related to the time-interval method.

**Aims**

This paper aims:

- To review and discuss reproducibility data of event-based analysis and time-interval analysis.
- To discuss the effect of different methodological factors (training, experience, interval duration, sample presentation order and judgment conditions) on inter- and intra-judge agreement of time-interval method.
- To discuss the comparison between the two methodologies in terms of which one represents the most reproducible or more accurate way to assess stuttering frequency (event-based or interval-based) based on selected studies.

**Methods**

**Review**

The first two authors of this paper conducted independent literature searches to locate as many relevant studies as possible. A systematic search of literature was conducted on electronic databases (ERIC, MEDLINE, PubMed, B-on and Cochrane Central register of Controlled Trials and Dissertation Abstracts) during January–February 2013. Reviewers looked for studies related to accuracy and reproducibility of time-interval measurement and with reproducibility data for event-based measurement, from 1992 (year after which the two measurement systems coexisted) to the present year. Keywords included: ‘stuttering’, ‘stammering’, ‘stuttering judgment’, ‘time-interval measurement’, ‘time-interval analysis’, ‘intra-judge time-interval’, ‘inter-judge time-interval’, ‘inter-judge reliability’, ‘intra-judge agreement’, ‘intra-judge reliability’, ‘percentage of syllable stuttered’, and ‘agreement of percentage of syllable stuttered’. The two reviewers also attempted to identify studies published in languages other than English.

Each citation abstract retrieved was reviewed independently by the two reviewers in order to assess the appropriateness of the study for inclusion in the systematic review. If it was indicated by the abstract that the study met the inclusion criteria (or if the abstract was unclear), the full text of the article was obtained and assessed for inclusion in the review, based on the inclusion criteria presented below.

When a complete copy of all potential studies was obtained, an independent evaluation of each study was
conducted by the two reviewers, in order to produce a final inclusion decision. In case of disagreement, reviewers discussed key issues until a consensus decision was attained.

**Inclusion and exclusion criteria**

For a study to be included in this review, the reviewers used the following inclusion and exclusion criteria:

- **Participants characteristics:** the participants were those who had been diagnosed as PWS (adolescents and adults) without other co-occurring disorders, using labels such as ‘stuttering’ or ‘stammering’.
- **Judges characteristics:** the judges were professionals with different backgrounds that have diverse experience levels related to stuttering.
- **Nature of the intervention:** studies that reported an intervention refer to the use of a training programme using judgment training based on event- and interval-based methodologies. The reviewers excluded studies based on other methodologies, such as self-judgment, time-modified intervals and transcription-based techniques.
- **Outcome variables:** the outcomes included were accuracy, number of stuttered intervals, %SS, ICC, Pearson product-moment coefficient correlation and Cohen's kappa. When the metric scale was not described, the studies were excluded. The outcome variables related to validity, graphical representation and disfluency-types were not included.
- **Methodological quality:** no restrictions were imposed regarding methodological quality.

**Coding**

Papers meeting the inclusion criteria were coded for judgment experience, outcome and methodological factor as well as statistical method used. The coding was conducted independently by the two reviewers, each using a complete copy of the retrieved article. Agreements between the reviewers occurred for 95% of the occasions and the disagreements were solved through a discussion.

**Biostatistical methods**

Both reviewers individually extracted data related to inter- and intra-judge values as well as accuracy from selected studies. Data from each study are summarized in tables A1, B1 and B2 in the appendices.

**Results**

**Overview of the literature search**

The reviewers identified a total of 495 articles via initial search, based on different word combinations. The articles were flagged based on title and abstract and repeated articles were excluded. Following evaluation of the abstract and methodology, the number of articles was reduced to 70 for further review. The articles were excluded due to repetitions observed on electronic databases and because of the absence of inter- and intra-judge values. The detailed review of the 70 studies based on the exclusion criteria (i.e., articles with reproducibility data related to self-judgement, modifier time-interval, transcription-based or disfluency types, as well as graphical representation and a logarithmic scale transformation of values were excluded) revealed that 48 articles presented inter-judge and/or intra-judge values that could be used in the review. Data are shown in tables A1, B1 and B2 in the appendices.

**Characteristics of selected studies and results summary**

The majority of studies related to event-based methodology reported indices of reliability. However, some selected studies used the concepts of agreement and reliability interchangeably, not being specific about the reproducibility concept reported. For that reason, the reviewers decided to use the term ‘reproducibility’ when referring to event-based methodology and agreement when mentioning time-interval methodology.

**Event-based methodology**

A total of 32 studies related to event-based methodology were selected (see table A1 in appendix A). The aims and objectives of the selected studies refer to a broad scope (not specific to the study of the impact of methodological factors on inter- and intra-judge agreement). However, reproducibility data of %SS was a common outcome measure between the studies. Those studies were mostly related to factors that included: the study of several stuttering intervention programmes (e.g., Camperdown, SpeechEasy, Stuttering Modification Programs or cognitive–behavioural therapy); the effect of delayed auditory feedback (DAF) and frequency altered feedback (FAF); the effect of personality traits and temperament on stuttering frequency.

Event-based methodology is used for frequency calculation in assessment instruments, such as the Stuttering Severity Instrument—4 (SSI-4) (Riley 2009) or the Test of Childhood Stuttering (TOCS) (Gillam et al. 2009). SSI-4 is a frequently used instrument for
determining aspects of stuttering severity in children and adults. This instrument uses %SS for frequency computing. TOCS is used to assess speech fluency skills and stuttering-related behaviours in children 4–12 years old (Gillam et al. 2009). For the assessment of stuttering frequency, TOCS uses the percentage of stuttered words. The use of event-based methodology in standardized tests, in research and in clinical decision-making process led to the widespread use of this methodology (Yaruss 1997).

Similar to some interval-based methodology studies (Finn 1997, Godinho et al. 2006, Ingham and Cordes 1997, Ingham et al. 1995), the majority of the selected event-based studies described judges as experienced (researchers or speech and language therapists—SLT), trained or enrolled in a training programme, as experienced and trained judges were, in general, more reliable than judges with no experience or with no training (Ingham et al. 1993a, 1994a, 1996, 1999, Einarsdóttir and Ingham 2008). This was observed across the studies, given that the inter- and intra-judge reproducibility exhibit high values for all indices (Hinkle et al. 2003, Landis and Koch 1977, Nunnally and Bernstein 1994, Baer et al. 1987, McHugh 2012, Cordes et al. 1992).

**Interval-based methodology**

A total of 16 studies related to time-interval analysis methodology were selected (see table B1 and B2 in appendix B). Ten studied the influence of different methodological factors (training, experience, interval duration, sample presentation order and judgment conditions) on inter- and intra-judge agreement.

All but one of the selected studies (Einarsdóttir and Ingham 2008) was conducted with speech samples of adolescent/adult population.

Five studies conducted training programmes with judges to discriminate between agreed intervals of stuttered and non-stuttered speech (Ingham et al. 1993b, Cordes and Ingham 1994a, 1996, 1999, Einarsdóttir and Ingham 2008). Two (Einarsdóttir and Ingham 2008; Ingham et al. 1998a) of these five studies used a software and hardware system known as Stuttering Measurement Assessment and Training (Ingham et al. 1998a).

Four studies conducted an investigation with judges who had different stuttering judgment experiences (Ingham et al. 1993a, Brundage et al. 2006, Cordes et al. 1992, Cordes and Ingham 1994c). These studies showed that, in general, judges with more experience with stuttering presented significantly higher levels of interval-by-interval intra- and inter-judge agreement than judges with less experience (Cordes et al. 1992).

Two studies investigated the effects of interval duration on observers’ judgments of stuttering (Cordes and Ingham 1994c, Cordes et al. 1992). There does not appear to be one obvious interval duration that satisfies all the factors that appear to be relevant in selecting interval duration. Nevertheless, some findings suggest the use of intervals with a duration between 3 and 5 s (Cordes and Ingham 1994c).

Two studies reported inter-clinic comparisons of stuttering judgments. One was related to graduate and undergraduate students (Ingham et al. 1993a); the other was about recognized authorities in stuttering research (Cordes and Ingham 1995). Inter-clinic comparisons revealed that judges from the same research centre tended to do similar stuttered judgments. However, there was a great amount of disagreement between authorities in stuttering research from different research centres and an uniformly high intra-judge agreement among judges (Cordes and Ingham 1995).

One of the selected studies explored differences in stuttering judgments related to sample presentation order (Ingham et al. 1993b) and one other study examined differences in audiovisual and audio-only judgment conditions (Ingham et al. 1993a). No significant differences were found across sample presentation order or for audiovisual/audio-only conditions (Ingham et al. 1993a, 1993b).

Six studies (Ingham et al. 1995, 1997, Fox et al. 2000, Finn 1997, Godinho et al. 2006, Ingham and Cordes 1997) had different aims and objectives, but reported inter- and intra-judge agreement of stuttering frequency using time-interval analysis technique. Three of these six studies (Ingham et al. 1995, Finn 1997, Ingham and Cordes 1997) describe judges as experienced researchers or SLT.

Accuracy, intra- and inter-judge agreement was higher for trained groups when compared with non-trained groups (Ingham et al. 1993b, Cordes and Ingham 1994a, 1996, 1999, Einarsdóttir and Ingham 2008, Godinho et al. 2006). Comparing stuttered intervals judgments and non-stuttered intervals judgments made by students and clinicians with judgments made by highly experienced judges, the accuracy levels were higher for non-stuttered intervals than for stuttered intervals (Brundage et al. 2006).

**Discussion**

**Inter- and intra-judge reproducibility of event-based methodology**

Reproducibility data related to %SS were reported as Pearson product-moment correlation, Cohen’s kappa, ICC and agreement percentage (see table A1 in appendix A). Pearson product-moment correlation values for inter-judge ranged between 0.88 and 0.99 and intra-judge data ranged between 0.86 and 1.0. The
values represent a high correlation (Hinkle et al. 2003).

Cohen’s Kappa inter-judge values ranged from 0.73 to 0.98 and intra-judge values varied between 0.76 and 0.93, which can be interpreted as a substantial to almost perfect agreement correlation (Landis and Koch 1977).

ICC inter-judge values ranged was between 0.83 and 1.00; intra-judge ranged between 0.89 and 0.99. The reported values mean that more than 89% of the observed score variance is due to true score variance and that the balance of the variance (1 minus the ICC value) is attributable to error (Weir 2005). The reported outcomes were greater than 0.70, which is classified as good (Nunnally and Bernstein 1994).

Studies that report agreement percentage present inter-judge values that vary between 88% and 98% and intra-judge varying from 95% to 99%. These values were above the minimum acceptable agreement (i.e., > 80%) (McHugh 2012, Baer et al. 1987).

The low number of samples/syllables as well as the small number of judges used to assess them can help to explain the high levels of intra- and inter-judge values. Also the use of experienced and trained judges, could also have contributed to the high values reported, as stated in the SSI-4 manual ‘with practice most clinicians can achieve self-agreement of 85% or higher; […] more stringent training procedures should achieve 90%’ (Riley 2009: 17). Frequency estimation through event-based methodology presents some other problems related to the reproducibility indices chosen. ICC and Person product-moment correlation (the more usual indices) are not a report of absolute reliability but relative reliability. Absolute reliability (i.e., agreement) is ‘preferable in all situations in which the instrument will be used for evaluation purposes’ and is a characteristic of the instrument instead of a performance with a specific sample (relative reliability) (de Vet et al. 2006: 1038). Relative reliability is ‘highly dependent on the variation in the population sample’ and is ‘only generalisable to samples with a similar variation’ (de Vet et al. 2006: 1038).

In the studies (Ingham et al. 1993a, Cordes and Ingham 1994c, 1995, Brundage et al. 2006, Cordes et al. 1992) in which training was not applied showed that intra-judge agreement (reported as agreement percentage) of time-interval measurement ranged between a mean of 76% and 98% for more experienced judges, and between 60% and 96%, for inexperienced judges. Values of inter-judge agreement (reported as agreement percentage) for experienced judges ranged between 57% and 87%. Inexperienced inter-judge agreement values ranged between 74% and 89%. Generally, studies concluded that inter- and intra-judge agreement were higher in more experienced judges (Ingham et al. 1993a).

Nevertheless, different studies presented distinct results concerning the experience with stuttering, mainly due to ill-defined judges’ experience with stuttering (Ingham et al. 1993a). In this sense, more recent studies that do not analyse the influence of methodological factors (Ingham et al. 1995, Finn 1997, Godinho et al. 2006, Ingham and Cordes 1997) used experienced judges for agreement purposes based on assumption that, generally, more experienced judges reached better agreement values.

The studies in which training was applied to inexperienced judges led to a higher increment in accuracy, inter- and intra-judge agreement in experimental groups, when compared with control groups. Ingham et al.’s (1993b) study of inter-judge agreement results were 71% in pre-training and increased to 86% after training. The accuracy decreased from 5.0 errors (i.e., a non-correspondence between judges’ scores and the judgment defined by a group of authorities in stuttering assessment) on pre-training to 1.6 errors on post-training. Cordes and Ingham’s (1999) study, training lead to a greater accuracy, whether judges needed to learn to match more stuttering events or learn not to over identify non-stuttered intervals. Accuracy and inter-judge agreement also increased for samples who had not been used for training, which means that training had effect for familiar (i.e., samples used during training sessions) and unfamiliar samples (i.e., samples not used during training sessions). Related to training with highly agreed intervals (i.e., 100% agreement), it has been shown that it is more effective for intra-judge agreement (increase from 82% in pre-training to 88% in post-training) and inter-judge (increase from 79–80% to 84–88% in post-training) than training or exposure to poorly agreed (intra-judge increase from 82% in pre-training to 86% in post-training; inter-judge decrease from 81% in pre-training to 78% in post-training) or randomly selected intervals (Ingham et al. 1993b, Cordes and Ingham 1994a). Software for judges’ training (SMAAT) with samples from AWS (Ingham et al. 1998a) or preschool children who stutter (Einarsdóttir and Ingham 2008) used highly agreed intervals to improve the correspondence between judges’ score and the criterion judgement established by a group of authoritative judges in

**Inter- and intra-judge agreement of interval-based methodology**

Inter- and intra-judge agreement data related to interval-based technique were reported as agreement percentage between or within judges. Agreement will be discussed in terms of the impact of methodological factors.

Duration is another methodological factor explored in two of the selected studies (Cordes and Ingham 1994c, Cordes et al. 1992). Cordes et al. (1992) found that inter-judge agreement was maximized at greater than chance levels for intervals between 3 and 4 s. Two further studies (Ingham et al. 1993a, 1993b) used 4 s interval-recordings. However, the 4 s interval duration was chosen based on interval analysis of event-recorded data, which means that this interval could not be the best time interval for interval judgments (Cordes and Ingham 1994c). Cordes and Ingham (1994c) made direct comparisons between different interval durations, within an interval-recording judgment system and proposed the use of intervals with a duration between 3 and 5 s with acceptable agreement. After Cordes and Ingham's (1994c) study, all the selected studies for the systematic review used 5 s interval duration.

The appropriateness of using interval-based methodology has been discussed in the literature. It was stated that the subdivision of speech into intervals presented randomly to the judge affects the assessment of co-articulatory gestures, insofar as the ‘speech gestures can be influenced by the preceding and following gestures [. . .]’ and its duration can be longer than the interval duration (Guntupalli et al. 2006, p. 10). The same authors also refer to under- or over-count of stuttering events that can occur in time-interval based methodology. Due to issues that were discussed by Guntupalli et al. (2006), the increase in agreement produces a decrease in validity. The duration of time intervals has also been discussed. Howell et al. (1998) and Howell (2005) suggest that long intervals lead to a ceiling effect with all intervals being judged as stuttered, which impairs the detection of changes during intervention. Howell (2005) also found that the effect of interval duration is dependent on the severity of stuttering (i.e., the more severe the stuttering, the more likely intervals were to be judged consistently as stuttered).

Yaruss (1997) also discusses some concerns related to time-interval methodology. The author argues that this methodology cannot be easily transferred to clinical settings and nor can it be employed during the clinical decision making process with clinical utility (i.e., time-interval does not present a frequency value with clinical significance).

Despite the disadvantages noted in the literature related to the use of interval based methodology in the assessment of stuttering, time interval methods have often been used to estimate behaviours due to the impracticality of focusing on multiple categories of behaviour/multiple subjects (Meany-Daboul et al. 2007).
Is it possible to quantify which of the two methodologies represents the most reproducible or more accurate way to assess stuttering frequency based on selected studies?

To analyse which of the two methodologies presents better inter-judge, intra-judge or accuracy related to assessment of the frequency, a method comparison study would be necessary in order to inspect the differences between measurements made by the two methods. The two methods should be in the same metric or scale (i.e., equal agreement index estimated from both methodologies during a pair of measurements from each subject) to quantify their agreement (Bartlett and Frost 2008).

The stuttering frequency of the same sample of subjects needs to be measured (and the calculation of inter-judge, intra-judge or accuracy needs to be made) with both measurement methodologies (Bartlett and Frost 2008) to quantify the agreement between methods and to assess which method would give the least inter-rater and inter-method variation (Bowers et al. 2009). Researchers can visually assess the agreement/disagreement between measurement techniques based on graphical techniques such as the one proposed by Bland and Altman (1986). Data from the subject’s measurements made by the two methods could be plot (i.e., the difference between measurements against the mean) and limits of agreement could be calculated. Such limits of agreement would represent the range within which it is expected that most of the differences (95%) lie.

Considering reproducibility data of event-based methodology and interval-based methodology it is important to mention the use of the two methodologies in the same speech samples in two studies, those of Bothe (2008) and Cordes and Ingham (1999). Bothe (2008) assessed the similarities between judgments made in the same speech sample with interval-based methodology and disfluency types present (a categorical event-based approach) in each 5 s interval from 20 children who stutter. The author concluded that the agreement for disfluency types was very low and that between 54% and 55% of the intervals agreed by the highly experienced judges (to be stuttered or non-stuttered) could be categorized similarly on the basis of disfluency types data (if a disfluency type was accepted as present based on 80% of judgments). In spite of the comparison of binary stuttered/non-stuttered judgment and disfluency types, it is not possible to quantify the agreement/disagreement because variables are different and with obviously distinct metrics. The other study (Cordes and Ingham 1999) used a previously developed software and hardware system for training (SMAAT) to investigate if a training programme based on time-interval methodology could improve the judgement of stuttering events. After training, inter-judge, intra-judge and accuracy values increased (agreement was only calculated for time-interval methodology and accuracy for both methodologies). Even though the accuracy data were calculated for both methods on the same speech sample, quantitative specific data does not exist for event-based methodology, making it unfeasible to compare the two methods.

Based on the data of selected studies, presented in tables A1, B1 and B2 in the appendices, related to inter-judge, intra-judge or accuracy for the two methodologies, it can be concluded that it was not possible to quantify which of the methodologies represent the most reproducible and more accurate method to assess stuttering frequency because the majority of event-based methodology studies reported relative reliability indices and interval-based methodology studies reported indices of absolute reliability (i.e., percentage of agreement). Further, in the majority of studies, the two measurement methods were not used in the same speech sample (Bartlett and Frost 2008). It was not possible to assess the consistency of study outcomes due to the lack of data (i.e., confidence intervals reported or the necessary data to compute).

Although a method comparison study has not been implemented, many studies reported that time-interval-based measure may lead to a ‘markedly’ (Ingham and Cordes 1997) or ‘demonstrably’ (Brundage et al. 2006) better inter- or intra-judge agreement than the traditional obtained with procedures required to identify events (%SS). Some studies (Curlee 1981, Coyle and Mallard 1979, Emerick 1960, MacDonald and Martin 1973, Young 1975, Martin and Haroldson 1981, Martin et al. 1988) about listeners’ agreement of stuttering events (which reported unit-by-unit inter-judge agreement below 60%, as reported by Cordes et al. 1992) used an agreement index developed by Young (1975) based on the number of observers (n), the total number of words marked as stuttered (T) and the total number of different words marked as stuttered (Td):

\[
\text{Agreement index (Young 1975)} = \left[ \frac{1}{n-1} \right] \left[ (\frac{T}{Td}) - 1 \right]
\]

The interval-based methodology estimates the percentage of judges’ agreement for the occurrence of stuttering based on the positive agreement index (PA index), in which the number of positive intervals (i.e., NP, an interval in which 80% or more of the judges recognized at least one stuttering event) and the number of ambiguous intervals (NA) were considered as factors:

\[
\text{PA index} = \frac{NP}{(NP + NA)}
\]

Despite the same type of metric scale (percentage of agreement, which is an index of absolute reliability), the two measurement techniques are still not comparable due to estimation differences inherent in the formulas used for calculation.


Study limitations

The main limitation of the present study was related to the statistical methods used in the selected studies for the review. Specifically, the majority of selected studies that used event-based methodology present indices of relative reliability (e.g., Pearson product-moment correlation and ICC), which are less stable over different population than agreement indices. The number of studies with reproducibility values reported is reduced \( n = 48 \), compared with the number of studies retrieved from the database initial search \( n = 495 \). The lack of studies reporting agreement indices leads to uncertainty related to the closeness of the scores stemming from the event-based methodology during repeated measurements (de Vet et al. 2006). It should also be noted that although the large number of studies \( n = 495 \) identified in initial search, only 48 present sufficient agreement data suitable for review. The lack of agreement data makes it impossible to generalize experimental effects to other populations, as well make conclusions about their consistency or stability in replications (Breakwell et al. 2012, de Vet et al. 2006).

Conclusions

‘The frequency of fluency breaks is often one of the most obvious aspects of the problem and to some degree impacts the perception of severity, particularly by the listener’ (Manning 2010: 163). Researchers and clinicians concur that objective definitions of stuttering, and reliable and valid measurements procedures are both desirable and a core necessity for treatment outcome research and replication (Brundage et al. 2006).

Both interval- and event-based methodologies use experienced and/or trained judges for reproducibility purposes. Trained judges as well as the small number of samples used for reassessment in event-based methodology studies could contribute to agreement values beyond the references for good reproducibility values in all indices (Hinkle et al. 2003, Landis and Koch 1977, Nunnally and Bernstein 1994, Baer et al. 1987, McHugh 2012). It can also be concluded that it was impracticable to quantify the agreement between inter-judge, intra-judge or accuracy measurements made by the two methods in the selected studies (Bartlett and Frost 2008).

Frequency is an aspect that should be relatively easy to calculate, but it is difficult to obtain reliable tabulations of event counts (Curlee 1981, Coyle and Mallard 1979, Emerick 1960, MacDonald and Martin 1973, Young 1975, Martin and Haroldson 1981, Martin et al. 1988, Ham 1989, Kully and Boberg 1988, Cordes et al. 1992). Researchers attempted to minimize this problem by developing better descriptions of stuttering (Packman and Onslow 1998), developing new methods for the detailed written analysis of stuttering (Yaruss et al. 1998), developing training methods (Cordes and Ingham 1999, Ingham and Cordes 1997) or using consensus judgment procedures (Cordes 2000). Stuttering, as a complex behaviour, can be measured with different methods, based on the desired parameters and the aims of the assessment. With the ideas reported in the present paper, clinicians/researchers should reflect upon the influence of methodological factors and choose the most appropriate frequency assessment method.

A suggestion for future research is to implement a method comparison study between event- and interval-based methodologies, both in the same metric scale of agreement values. The aim of such a (method comparison) study would be to verify whether the measurements made by the two methods are ‘sufficiently close’ or whether the methods can coexist due to good agreement in terms of reproducibility values (Bartlett and Frost 2008).

Acknowledgments

This study was developed as part of the Ph.D. Thesis of the first author at the University of Aveiro, Portugal. This work was partially funded by National Funds through FCT—Foundation for Science and Technology, in the context of project PEst-OE/EEI/UI0127/2014. This research was also partly supported by a doctoral grant (SFRH/BD/78311/2011) from FCT to Ana Rita S. Valente. Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References


Event- and interval-based measurement of stuttering: a review


GILLAM, R. B., LOGAN, K. J. and PEARSON, N. A., 2009, Test of Childhood Stuttering (Austin, TX: PRO-ED).


Max, L. and Baldwin, C. J., 2010, The role of motor learning in stuttering adaptation: repeated versus novel utterances in a


## Table A1. Results summary of selected studies related to reproducibility data of %SS

<table>
<thead>
<tr>
<th>Study</th>
<th>Judges</th>
<th>Number of samples/syllables assessed</th>
<th>Inter-judge</th>
<th>Intra-judge</th>
<th>Statistical method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Langevin and Boberg (1993)</td>
<td>Two: research assistant, the rater, was trained to criterion level at the Institute by the clinical director and the clinical director</td>
<td>Five samples from 10 subjects (chosen from a pool of at least, 600 samples)</td>
<td>0.99</td>
<td>0.999</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Boberg and Kully (1994)</td>
<td>Two independent raters (with 6 h of training)</td>
<td>26 samples from six subjects (chosen from a total of 49 PWS)</td>
<td>0.99</td>
<td>Information not available</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Packman et al. (1994)</td>
<td>Two: the first investigator and an independent judge</td>
<td>24 speaking sessions chosen from a totality of 88 speaking sessions</td>
<td>0.99</td>
<td>0.99</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Kalinowski et al. (1996)</td>
<td>Two judges</td>
<td>One-third of the total sample (2400 syllables)</td>
<td>0.88</td>
<td>Not calculated</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Sparks et al. (2002)</td>
<td>Two: one rater enrolled in an intensive graduate fluency course and one experienced speech–language pathologist with stuttering disorders</td>
<td>150 syllables</td>
<td>Not calculated</td>
<td>0.92, 0.87, 1.00, 1.00</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>O’Brian et al. (2003)</td>
<td>Two clinicians</td>
<td>10% of the total sample (88 samples)</td>
<td>0.99</td>
<td>1.00</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Menzes et al. (2008)</td>
<td>Two experienced speech and language therapists (SLT)</td>
<td>10% of the total sample (200 samples)</td>
<td>0.96</td>
<td>0.99</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>O’Brian et al. (2008)</td>
<td>Two SLT</td>
<td>10% of the total sample (100 samples)</td>
<td>0.94</td>
<td>0.98</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Lincoln et al. (2010)</td>
<td>Two SLT</td>
<td>12.5% of the total sample (240 samples)</td>
<td>0.96</td>
<td>0.94</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Cream et al. (2010)</td>
<td>Two SLT</td>
<td>10% of the total sample (705 samples)</td>
<td>0.8</td>
<td>0.95</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Manning and Beck (2013)</td>
<td>Two raters: the first author and a trained doctoral student</td>
<td>6000 (40%) words chosen from a total of 15 000 words</td>
<td>0.92</td>
<td>0.98</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>Kalinowski et al. (1999)</td>
<td>Two trained research assistants</td>
<td>10% of total sample (600 syllables)</td>
<td>0.84</td>
<td>0.93</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Stuart et al. (1996)</td>
<td>Two trained research assistant</td>
<td>10% of total sample (3000 syllables)</td>
<td>0.73 (syllable by syllable)</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Zimmerman et al. (1997)</td>
<td>Two: the first author and independent trained researcher</td>
<td>10% of total sample (135 samples)</td>
<td>0.90</td>
<td>0.92</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Eichhardt et al. (1998)</td>
<td>Two graduate speech-language pathologists</td>
<td>10 samples (no data related to the totality of syllables)</td>
<td>0.68</td>
<td>0.76</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Armson and Stuart (1998)</td>
<td>Two trained research assistants (trained by an American Speech-Language-Hearing Association (ASHA)-certified clinician with over 20 years of experience)</td>
<td>12% of the total sample (no data related to the totality of syllables)</td>
<td>0.78</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Kalinowski et al. (1999)</td>
<td>Two trained research assistant</td>
<td>10% of the total sample (1800 syllables)</td>
<td>0.78 (syllable by syllable)</td>
<td>0.86 (syllable by syllable)</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Langevin et al. (2006)</td>
<td>Three Dutch and four Canadian research assistants (trained with a training programme)</td>
<td>11% of the total sample from Dutch participants (473 samples); 15% of the total sample from Canadian participants (54 samples)</td>
<td>0.97</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Armson et al. (2006)</td>
<td>Two trained by clinician with over 25 years of clinical experience in the area of stuttering</td>
<td>13% of the total sample (600 syllables)</td>
<td>0.85 (Cohen’s kappa), 0.99 (Spearman rank-order correlation)</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>O’Donnell et al. (2008)</td>
<td>Two: the first author and a research assistant, (both been trained to identify stuttering episodes by a clinician with 30 years of clinical experience in the area of stuttering)</td>
<td>10% of the total sample (210 speech samples)</td>
<td>0.86</td>
<td>0.89</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Study</td>
<td>Judges</td>
<td>Number of samples/syllables assessed</td>
<td>Inter-judge</td>
<td>Intra-judge</td>
<td>Statistical method</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Armson and Kiefte (2008)</td>
<td>Two research assistants</td>
<td>13% of the total sample (600 syllables)</td>
<td>0.93</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Max and Baldwin (2010)</td>
<td>Two judges</td>
<td>4000 words from a total of 60 000 words</td>
<td>0.89</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Bauerly and De Nil (2011)</td>
<td>Information not available</td>
<td>25% of total data (no data related to the totality of syllables)</td>
<td>0.98 (for reading samples of PWS); 0.96 (for speaking samples of PWS)</td>
<td>Not calculated</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>Huinck and Rietveld (2007)</td>
<td>Three raters</td>
<td>17% of the total sample for inter-judge reliability (952 samples); 12% of the total sample for intra-judge reliability (952 samples)</td>
<td>0.95</td>
<td>0.95</td>
<td>ICC</td>
</tr>
<tr>
<td>Antipova et al. (2008)</td>
<td>Two: the first author and an independent/research clinician</td>
<td>10% of the total sample (150 samples)</td>
<td>0.91</td>
<td>0.99</td>
<td>ICC</td>
</tr>
<tr>
<td>Langevin et al. (2010)</td>
<td>Four trained research assistants</td>
<td>13 samples chosen from a totality of 126 samples</td>
<td>0.98 (CI = 0.94, 0.99)</td>
<td>Not calculated</td>
<td>ICC</td>
</tr>
<tr>
<td>Unger et al. (2012)</td>
<td>Trained research assistants (number not mentioned) and the first author</td>
<td>Information not available</td>
<td>1.00</td>
<td>Not calculated</td>
<td>ICC</td>
</tr>
<tr>
<td>Beilby et al. (2012)</td>
<td>Two SLT</td>
<td>2000 syllables (totality of the sample)</td>
<td>0.91 (one-way independent group random effect model analysis)</td>
<td>0.89</td>
<td>ICC</td>
</tr>
<tr>
<td>Gallopa and Runyanb (2012)</td>
<td>Two</td>
<td>900 syllables (totality of the sample)</td>
<td>0.83</td>
<td>Not calculated</td>
<td>ICC</td>
</tr>
<tr>
<td>Mackood et al. (1995)</td>
<td>Two: one trained speech-language pathology graduate student and one trained research assistant</td>
<td>30% of total sample (3000 syllables)</td>
<td>88%</td>
<td>95%</td>
<td>Agreement percentage</td>
</tr>
<tr>
<td>Hubbard (1998)</td>
<td>Two SLT</td>
<td>1538 syllables (the totality of the sample)</td>
<td>98% (point-to-point agreement for stuttering occurrence and non-occurrence)</td>
<td>98% (point-to-point agreement for stuttering occurrence and non-occurrence)</td>
<td>Agreement percentage</td>
</tr>
<tr>
<td>Vincent et al. (2012)</td>
<td>Two: one SLT and one graduate student in speech-language pathology, who had been trained in identifying stuttering moments</td>
<td>12 390 syllables (the totality of the sample)</td>
<td>95%</td>
<td>99%</td>
<td>Agreement percentage</td>
</tr>
</tbody>
</table>
### Appendix B

#### Table B1. Results summary of selected studies related to agreement data in which training was applied

<table>
<thead>
<tr>
<th>Study and judges</th>
<th>Inter-judge agreement</th>
<th>Intra-judge agreement</th>
<th>Accuracy</th>
<th>Factor</th>
<th>Statistical method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingham et al. (1993b): 23 undergraduate students</td>
<td>Group A: 81% for occasion 1 (so-called pre-training) and 83% for occasion 2 (so-called post-training)</td>
<td>Group A: a mean of 91%</td>
<td>Group B: 5.0 errors on occasion 1 and 1.6 errors on occasion 2</td>
<td>Training</td>
<td>Agreement percentage</td>
</tr>
<tr>
<td></td>
<td>Group B: 71% for occasion 1 and 86% for occasion 2</td>
<td>Group C: 75% for occasion 1 and 65% for occasion 2</td>
<td>Group C: a mean of 84%</td>
<td>Sample presentation order</td>
<td>Agreement percentage</td>
</tr>
<tr>
<td>Cordes and Ingham (1994b): 30 speech-language pathology and audiology students divided into six groups</td>
<td>TH: mean of 80% on occasion 1 and 79% on occasion 2</td>
<td>TH: mean of 82% on occasions 1–2, 88% on occasions 3–5</td>
<td>Training</td>
<td>Agreement percentage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean of 81% on occasions 1 and 2, 78% on occasions 4 and 5</td>
<td>mean of 83% on occasions 1–2, 86% on occasions 4–5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean of 81% on occasion 1 and 80% on occasion 2, 80% on occasion 4 and 78% on occasion 5</td>
<td>mean of 83% on occasions 1–2</td>
<td>Control: no tendency to change the total number of errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mean of 81% on occasion 1 and 80% on occasion 2, 82% on occasion 4 and 83% on occasion 5</td>
<td>mean of 81% on occasions 1–2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cordes and Ingham (1996): 10 undergraduate students</td>
<td>Stuttered intervals: Pre-training: 48% (occasion 1), 41% (occasion 2) for training speakers; 63% (occasion 1 and 2) for non-training speakers</td>
<td>Stuttered intervals: Pre-training ranged from 67% to 89%</td>
<td>Phase II: first three trials’ accuracy between 45% and 99%; last three trials’ accuracy not lower than 91%</td>
<td>Training</td>
<td>Agreement percentage</td>
</tr>
<tr>
<td></td>
<td>Post-training: 96% (occasion 3) and 99% (occasion 4) for training speakers; 97% (occasion 3) and 99% (occasion 4) for non-training speakers</td>
<td>Post-training ranged from 84% to 99%</td>
<td>Maintenance of accuracy during occasions 3 and 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-stuttered intervals: Pre-training: 98% (occasion 1) and 96% (occasion 2) for training speakers; 100% (occasion 1) and 99% (occasion 2) for non-training speakers</td>
<td>Non-stuttered intervals: Pre-training ranged from 89% and 100%Post-training ranged from 43% and 96%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disagreed intervals: Pre-training: 88% (occasion 1) and 77% (occasion 2) for training speakers; 85% (occasion 1) and 87% (occasion 4) for non-training speakers</td>
<td>Disagreed intervals: Pre-training ranged from 73% and 80%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post-training: 45% (occasion 4 and 5) for training speakers; 46% (occasion 3) and 99% (occasion 5) for non-training speakers for stuttered intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study and judges</td>
<td>Inter-judge agreement</td>
<td>Intra-judge agreement</td>
<td>Factor</td>
<td>Statistical method</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>--------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>Cordes and Ingham (1999): 20 university students semi-randomly allocated to two judge groups</td>
<td>Group 1: Pre-training: ranged from 71% and 76%; Post-training: ranged from 82% and 91%; Group 2: Pre-training: ranged from 75% and 87%; Post-training: ranged from 83% and 88%</td>
<td>Group 1: Pre-training: ranged from 81% and 87%; Post-training: ranged from 89% and 92%; Group 2: Pre-training: ranged from 86% and 92%; Post-training: ranged from 92% and 94%</td>
<td>Number of stuttering events; Stuttering duration; Number of intervals; Interval-by-interval inter-judge; Interval-by-interval intra-judge; Accuracy of training</td>
<td>Agreement percentage</td>
<td></td>
</tr>
<tr>
<td>Einarsdóttir and Ingham (2008): 20 female preschool teachers randomly allocated to an experimental group and to a control group</td>
<td>Experimental group: not calculated; Control group: not calculated</td>
<td>Experimental group: not calculated; Control group: not calculated</td>
<td></td>
<td>Agreement percentage</td>
<td></td>
</tr>
<tr>
<td>Study and judges</td>
<td>Inter-judge agreement</td>
<td>Intra-judge agreement</td>
<td>Factor</td>
<td>Statistical method</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>--------</td>
<td>--------------------</td>
<td></td>
</tr>
<tr>
<td>Cordes et al. (1992): three groups of judges (18 judges): six undergraduate students (UG), six graduate students (G) and six experienced clinicians (E)</td>
<td>Undergraduate: unavailable data</td>
<td>Graduate: unavailable data</td>
<td>Experience: unavailable data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Undergraduate: mean of between 60% and 85%, for 0.2 s (agreement for total counts of stuttered intervals); Graduate: mean of between 82% and 94%, for 0.2 s (agreement for total counts of stuttered intervals); Experience: mean of between 88% and 93%, for 0.2 s (agreement for total counts of stuttered intervals)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingham et al. (1993a): 34 graduate and undergraduate students</td>
<td>Experienced: mean of 82.1 on Set 1 and 57.1 on Set 2 for AV; mean of 77.2 on Set 1 and 70.0 on Set 2 for A</td>
<td>Inexperienced: mean of 78.6 on Set 1 and 78.6 on Set 2 for AV; mean of 81.4 on Set 1 and 74.3 on Set 2</td>
<td>Experienced: mean ranged from 76% to 94% for AV; from 79% to 94% for A; Inexperienced: mean ranged from 79% to 96% for AV; from 79% to 93% for A</td>
<td>Audiovisual conditions versus audio-only conditions; Judges experience; Inter-clinic comparison</td>
<td></td>
</tr>
<tr>
<td>Cordes and Ingham (1994a): 12 undergraduate students and 12 graduate students and clinicians</td>
<td>Experienced: mean of 85% for all judges</td>
<td>Inexperienced: mean of 85% for all judges</td>
<td>Experienced: not calculated; Inexperienced: not calculated</td>
<td>Interval duration; Judges experienced</td>
<td></td>
</tr>
<tr>
<td>Cordes and Ingham (1995): 10 researchers and clinical researchers</td>
<td>84% on occasion 1 and 86% on occasion 2, for all judges</td>
<td>Ranged from 83% and 98% for all judges</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingham et al. (1995); four researchers</td>
<td>Ranged from 50% of 1 s intervals agreed to 77% of 6 s intervals agreed</td>
<td>Not calculated – Agreement percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingham et al. (1997): two judges</td>
<td>Ranged from 86% to 93%</td>
<td>Not calculated – Agreement percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingham and Cordes (1997): 10 experienced researchers and clinic directors</td>
<td>76% (overall inter-judge agreement)</td>
<td>Not calculated – Agreement percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finn (1997): five graduate students in speech–language pathology</td>
<td>89% met the criteria of 80% agreement</td>
<td>Not calculated – Agreement percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fox et al. (2000): three judges</td>
<td>Mean interval-by-interval were 93%</td>
<td>Not calculated – Agreement percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brundage et al. (2006): 41 students and 31 clinicians</td>
<td>Mean of 87% for clinicians</td>
<td>Mean of 89% for students</td>
<td>Experience Agreement percentage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Godinho et al. (2006): two: one experimenter and one research assistant</td>
<td>Ranged between 75% and 97% across speakers</td>
<td>Ranged from 81% to 100% across speakers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>