Published in final edited form as:
*Stammering Res*. 2005 January 1; 1(4): 364–374.

# The effect of using time intervals of different length on judgements about stuttering

**Peter Howell**
Department of Psychology, Centre for Human Communications, Institute of Cognitive Neuroscience, and Institute of Movement Neuroscience, University College London, Gower St., London WC1E 6BT

## Abstract

Conventional clinical procedures for assessment of stuttering are reported to have poor reliability. Time interval analysis procedures have been reported to produce greater reliability than the conventional procedures. In time interval procedures, successive intervals of the same duration are extracted from a sample of speech and judged by participants as stuttered or fluent. There is a problem insofar as the amount of speech judged stuttered depends on the length of the interval used. This problem is illustrated in an experiment in which 1-s and 5-s intervals were drawn from the same samples of speech and judged by participants as stuttered or fluent. It is also shown that the problem of lack of sensitivity when longer intervals are used is more acute for individuals who exhibit severe stuttering. Since ability to detect changes in stuttering rate is dependent on the length of interval used (as well as stuttering severity), the procedure can highlight or disguise changes in stuttering rate depending on parameterization of interval length and choice of participants to study. Thus, use of different length intervals across studies can distort whether particular treatments have an effect on speech control. Therefore, it is concluded that time interval analysis, as it is currently used, is an unsatisfactory procedure. If a standard-length interval could be agreed, comparison across studies or analyses would be possible.

### Keywords

Stuttering assessment; time interval analysis procedure

## 1. Introduction

A conventional method for assessing stuttering is to make a count of the disfluent events directly from recordings (by using, for instance, a manually operated counter). Studies have shown that agreement between judges using such methods is only around 60% (e.g., Curlee, 1981; Martin & Haroldson, 1981). Ingham and his colleagues have proposed that time-interval (TI) procedures produce higher agreement. In TI procedures applied to stuttering, listeners hear (and in some cases see) a fixed-length extract of speech and designate it as stuttered (STUT) or fluent (FLU). The Ingham group claim that such procedures ".. could certainly lead to different ways of overcoming the problem of judgement reliability for stuttering" (Ingham, Cordes & Gow, 1993, p.512). If this claim is true, then TI procedures should be used for clinical assessment in preference to the standard procedures. The current study examines the claim that TI procedures provide a reliable indication about stuttering.

p.howell@ucl.ac.uk.


Europe PMC Funders Author Manuscripts

Reliability can be defined as the relationship between observed and true scores (Cordes, 1994). Ingham and co-workers used expert judges to establish the true responses to each time interval. They employed audio-visual recordings of speech from a selected sample of speakers who stutter. The samples of speech from each speaker were divided up into adjacent, non-overlapping 4-s intervals and these were played to judges in semi-random order. Each interval was presented for judgement to the panel of four experienced judges who independently assessed each sample twice. For each 4-s sample, these judges were instructed to identify whether the sample contained stuttering or not (Ingham et al., 1993, p. 506). These data were then used to locate which intervals the experienced judges agreed on (a criterion of 7/8 judgements given the same response was used for this purpose) and to establish what the agreed response should be (the response given on the majority of occasions for the agreed intervals). 110 of the 143 intervals tested (77%) were agreed by the experienced judges and 64 of these were judged to be STUT.

A major potential source of bias in such procedures stems from the fact that intervals of different lengths have been used across different TI studies. Howell, Staveley, Sackin and Rustin (1998) pointed out that a) when long intervals are used, the chance of the interval containing signs of stuttering increases, and b) long intervals will tend to result in a ceiling effect with all intervals judged as being stuttered. These effects limit the use of TI procedures for assessing whether a treatment reduces stuttering rate. To make such an assessment, a researcher may decide to have participants' speech judged before and after treatment. If an interval-length is used which results in a ceiling effect before and after treatment, it would not be possible to detect a change due to the treatment. Changes due to the treatment may have been evident if a shorter interval had been used. A worked example shows this. Say there are two samples of speech, each 100 s in length. One contains 30 stutterings and the other 20 stutterings and in both cases these are evenly distributed over the sample[1]. Assuming a speech rate of five syllables per second (Perkins, 2001) this would correspond to stuttering rates of 6%, and 4%, respectively. If the samples are partitioned into 5-s TIs, there will be an average of one stutter for the 20 stutter sample and 1.5 for the 30 stutter sample. If these are reliably judged, 100% of intervals will be designated STUT for both samples. This would suggest that there is no difference in stuttering rate between the samples. Different outcomes would be expected if the same samples were partitioned into 1-s TIs. The 1-s TIs made from the first sample would include 30 stuttered intervals and the second sample 20 stuttered intervals and observers making accurate judgements would reflect a 10% difference in stuttering rate between the two sets of intervals. The ceiling effect when the longer intervals are used shows that these intervals may fail to detect any fluency-enhancing effects of treatment procedures, which would have been evident if shorter intervals had been used. There are indications that this is more than a hypothetical possibility. For instance, Ingham, Moglia, Frank, Costello-Ingham and Cordes (1997) assessed the effects of frequency-shifted feedback on the speech of people who stutter using 5-s TIs. The long interval appears to have a) resulted in a ceiling effect in both the pre- and post-treatment conditions (both conditions showed around 100% intervals judged to be STUT), and b) led in turn to failure to find effects of the treatment that the majority of participants reported (i.e., increased fluency under frequency-shifted feedback) (Howell et al., 1998).

This analysis shows that the authors might very well have found fluency-enhancing effects of frequency-shifted feedback, in line with what their participants reported, if they had used a shorter (more sensitive) TI. Clearly, a procedure that can result in misleading conclusions

[1]The interval starts and ends (both 1-s and 5-s) are imposed irrespective of where utterances start. Providing interval durations are as long as the utterance duration or longer, there will be no effect due to the tendency of stutterings to occur at the beginning of sentences,

about treatment outcome is not satisfactory. To provide evidence on this, the study compared STUT/FLU judgements of the same material when it was segmented into 5-s and 1-s TIs. The earlier analysis predicts that more speech intervals will be judged STUT when the longer (5-s) intervals are used than when the shorter (1-s) intervals are used. The influence of interval length on speakers with different severity of stuttering is also examined. The above analysis predicts that speakers with more severe stutters will be less affected (as a higher proportion of their intervals will be judged stuttered whatever the length of interval used).

## 2. Method

### Speech samples

Eight 2-minute samples of speech from the UCLASS archive of stuttered speech (Howell & Huckvale, 2004) were selected for use in this study. These represent one sample each from eight male speakers ranging in age from 7 years 7 months to 17 years 9 months. The samples were chosen to cover a broad range of both age and stuttering rate. The samples can be down-loaded and examined by visiting http://speech1.psychol.ucl.ac.uk/index.htm. The samples employed were 0030_17y9m.1, 0061_14y8m.1, 0078_16y5m.1, 0095_7y7m.1, 0098_10y6m.1, 0138_13y3m.1, 0210_11y3m.1, 0234_9y9m.1 (as listed in the first column of Table 1). Further details about these (and other) speech samples from the UCLASS archive can be obtained from Howell and Huckvale (2004).

### Procedure

A MATLAB program was written to divide a file into either 1-s or 5-s intervals. The MATLAB program played one set of intervals and recorded listeners' responses to each interval as STUT or FLU. The program had an option that allowed intervals to be replayed at random with no replacement, or in sequence. The responses of the experts were used to obtain intervals that were agreed as fluent (FLU) or stuttered (STUT)[2]. These provided criteria against which the responses of naïve judges were assessed. These are described in more detail below.

### Selection of intervals for assessment and experts' criteria judgements

Four expert judges, each of whom had 10 or more years' intensive experience of assessing stuttered speech, were employed. The procedure for assessing the intervals was broadly similar to that which Ingham et al. (1993) adopted (see the start of the introduction). The differences in procedure were 1) that the experts judged each sample of speech both when it was segmented into 1-s, and into 5-s, TIs, 2) that judgements about the 1-s, and 5-s TIs were made twice, once when the TIs were sampled at random without replacement, and once when the TIs were presented in sequential order (the reasons for the latter are given below), and 3) they were instructed as to what events should be used to judge an interval as STUT, (this ensured that the task was less open-ended than in other studies).

The reason that judgements were made in sequential and random order was that in initial work with the experts, they indicated that their designations might well have differed if they had had the preceding interval available as context. (Howell, Sackin & Glenn, 1997 also found that judgements were better if some of the prior context, the preceding word in their report, was given before the judged extract.) The segmentation process led to two artifacts: 1. Sometimes an interval started at a point in speech that gave the sound an apparent hard

[2]The expert judges also indicated which of the intervals contained part-word repetitions or prolongations alone. These are intended for use in future studies and are not reported here as the procedures and treatment of results parallel those that Ingham and co-workers used in their studies.

onset which led judges to rate it as STUT. 2. Truncation at the end of an interval sometimes made an interval sound as if the first sound was part of a repetition sequence. In both cases the sound would have been designated STUT while hearing the preceding context would have led the judge to revise their view and designate the interval as FLU. There is a five times greater chance of these artifacts affecting 1-s than 5-s intervals. A 5-s TI was dropped when the judgement about one or more of the five 1-s intervals that made up the 5-s interval was judged STUT while the 5-s interval itself was judged FLU (indicative of these problems) in more than one of the eight judgements made by the experts.

It is also possible that there are cases where stutterings were not apparent on 1-s intervals but were on 5-s intervals. An example would be where a prolonged phone is split between two adjacent 1-s TIs leaving each section of the prolonged phone below the duration-threshold of a prolongation that then leads to each interval judged FLU. A 5-s TI was dropped when the judgement about the 5-s interval was STUT while all the five 1-s intervals it contained were judged FLU (indicative of such a problem) for more than one of the eight judgements made by the experts.

To ensure consistency in what events were considered as STUT, the experts were told:

1. not to count revisions, whole word repetitions or retraces as STUT unless there were other signs of stuttering (e.g., prolongation or repetition of part of a word).
2. to count silence, laughs, breathing noises and filled pauses as FLU. Intervals with these events were included in the results of the naïve judges who were also told to judge them as fluent.

The experts were also allowed to indicate any intervals that they considered to be ambiguous with repect to STUT/FLU status. One situation where this arose was when the duration of a sound was prolonged only marginally as this could have been done for emphasis or might have been a brief period of disfluency.

Ingham et al.'s (1993) procedure depended on the experts using their own judgement as to what was, or was not, stuttered. This would inevitably lead to difference of opinion and, for this reason, the procedure employed here where judges were told what events to consider STUT was considered preferable.

The experts next indicated when there were extraneous sounds, as these might have affected productions and/or judgements in intervals that contained them. These usually arose because the participant knocked the microphone or some other object in the recording environment. There were also occasional prompt questions by the researcher making the recordings (when, for instance, the speaker ran out of things to say). These intervals were noted and intervals with these events were not included in the results of the naïve judges.

During any one session, judgements were made about either the 1-s or 5-s set of intervals (interval length) and the intervals were either presented in random, or sequential, order (order type). The speech of each person who stutters was assessed in turn. The interval length by order type judgements were carried out in different random orders by the expert judges, and the judgements were separated by at least a week to prevent carry-over effects on judgements. The experts indicated which rejection criteron applied to an excluded interval. The numbers of 5-s intervals excluded by the different criteria discussed above are summarised in Table 1. 86 5-s intervals were excluded in total (corresponding to 430 1-s intervals).

Note that there were more cases where the 5-s interval was judged FLU but one of the constituent 1-s intervals was judged STUT (24) than vice versa (10). This suggests that the

hard onset and spurious repetition artifacts were more prevalent than cases where the more extensive 5-s context allowed disfluencies to be detected that were missed when the shorter 1-s segments were used.

Some of the 1-s intervals that were agreed by the experts occurred within a 5-s interval that included other 1-s intervals that the experts did not agree on (six cases in total). All constituent 1-s intervals of a 5-s interval were dropped when one or more of the 1-s intervals were not agreed. This permitted direct comparison - i.e. all 1-s intervals and the 5-s interval they comprised were agreed.

The exclusion criteria affected speakers differentially and this depended on the severity of their disorder. Table 2 gives some details of how the exclusion criteria affected individual speakers and how this relates to stuttering incidence in the 5-s intervals that remain. The overall rejection rate of 5-s samples was 46% (i.e., the 43% in Table 1 and the additional 3% where 5-s agreed intervals contained at least one 1-s interval not agreed on). The same applies (though to a lesser extent) to Ingham et al. (1993). As the main point here is to evaluate the TI procedure for intervals that are precisely defined, more strict criteria were applied (leading to the higher rejection rate than in Ingham et al., 1993).

The experts' judgements served two roles: 1) to select intervals for testing with the naïve judges as described in the next section; 2) to determine whether the selected intervals were responded to correctly by naïve judges (correspond with experts' responses) or not (did not correspond with experts' responses).

### Assessment of intervals by naïve judges

The naïve judges were undergraduates aged betwen 20 and 22 from a variety of humanity disciplines who reported that they had no experience of judging stuttered speech (speech science students were explicitly excluded). Eight naïve judges assessed the sets for each interval length (1-s and 5-s intervals) in random order for each speaker separately in the same way as the experts. The two assessments of the same material (1-s or 5-s intervals) were done at least a week apart. TI judgements were made with intervals presented as with the experts. All intervals were judged (i.e., material in the row labeled 'initial duration' in Table 2) but the results are only reported for those intervals that the experts agreed on (i.e., material in the row labeled 'after exclusion' in Table 2). Using all material ensured that there was at least one FLU interval agreed by the experts for each of the speakers, so judges should have used both available responses and the context in which they made these judgements was the same as that of the experts so that judgements could be compared (Parducci, 1965).

## 3. Results

### Prediction one

The experts' response designations (STUT or FLU) for the agreed intervals, were used as the criterion against which to assess the accuracy of the naïve judges. The expert judgements for 1-s intervals excluded all five 1-s intervals when the 5-s interval they comprise was agreed by the experts to be FLU but one of its constituent 1-s TI was judged STUT and also excluded all five 1-s intervals when the 5-s interval they comprise was agreed by the experts to be STUT but all the constituent 1-s TI were judged FLU. Cases were, however, included where 5-s intervals were agreed to be STUT by the expert judges which contained one or more 1-s intervals that were agreed by the experts to be FLU (though at least one 1-s interval has to be expert-agreed STUT). This arises when the experts agree that there is a stuttering of less than 5-s in length and leads to agreed FLU 1s-TI from the expert judges for where

there were no agreed FLU 5-s TIs for speakers 2 and 4 (see below where responses of each judge to each speaker are presented).

Responses from 5-s intervals were converted to responses to 1-s intervals so that the results on different interval lengths could be compared directly. To do this, 1) a 5-s FLU interval was considered to be made up of five 1-s FLU intervals, and 2) a 5-s STUT interval was considered to be made up of five 1-s STUT intervals. The second assumption operationalizes the view that too much material is designated STUT when longer intervals are used (i.e., the main topic addressed in this paper). After this response translation, comparison can be made between intervals of different length. Table 3 gives the mean percentage correct (and SD) for the naïve judges separately for each speaker and separately for both interval lengths using the experts' responses to the agreed intervals as the criterion (more extensive data from individual judges are given in Table A.1 at the end of this article).

The mean percentages for each speaker from Table 3 are presented in histogram form in Figure 1. Speakers are indicated on the abscissa, blue bars represent 1-s judgements and red bars represent 5-s judgements. It can be seem that for all but one speaker (S3), there were more 5-s intervals judged STUT than 1-s intervals judged STUT. Thus for seven out of eight of the speakers, the judgments about STUT across 1-s and 5-s intervals were in the expected direction according to prediction one.

To test the first prediction statistically, STUT intervals alone were examined to see whether more TIs were judged STUT when the longer (5-s) intervals were used than when the shorter (1-s) intervals were used. A mixed model Analysis of Variance was employed with the within-groups factor of interval length (1-s versus 5-s) and the between-groups factor of speaker (the eight speakers whose speech was judged) and the dependent variable was proportion of intervals judged stuttered). There was a significant effect of interval length ($F(1,56) = 20.4$, $p < .001$) which arose because more speech was judged STUT when 5-s intervals were used than when 1-s intervals were used. There was also a significant effect of speaker ($F(7,56) = 5.410$, $p < .001$) which indicated total stuttering rate (across 1-s and 5-s TI) differed (i.e., showed the speakers differed in severity of stuttering). There was also an interaction between interval length and speaker ($F(7,56) = 3.27$, $p < 0.01$). This arose because the effect on stuttering rate of changing interval length (stuttering rate increase going from 1-s to 5-s intervals) depended on speaker. Inspection of individual speaker data revealed that more severe stutters showed less effect than milder ones. This effect is explored in more detail in the next section.

## Prediction two

The second prediction tested was that the participants who had a more severe stutter had less chance of losing intervals than milder ones. This prediction was based on the fact that the first two exclusion criteria in Table 1 would apply less to speakers with a severe stutter, as few of their intervals did not include a real stutter before and after these exclusion criteria were applied. Essentially this implies that the first two exclusion criteria were less applicable to speakers with a severe stutter than speakers with a milder stutter. This predicts that there should be a negative correlation between amount of speech lost from the expert judges and the percentage of TIs judged STUT after the exclusion criteria were applied (i.e. a one-tail prediction). The Pearson product moment correlation coefficient was -.512 which was in the correct direction but not significant ($p = .10$, one tail) which is not surprising given the small N. The related correlation coefficient between amount of speech lost and total length of those 5-s intervals designated STUT, correlated negatively $r = -.818$, $p = .013$ with an N of 8. This gives qualified support to the view that speakers with a less severe stutter lose more intervals due to the exclusion criteria than speakers with a more severe

stutter. Thus, TI assessments using 5-s intervals affect speakers with different stuttering severity differentially.

### Examination of data from individual judges and speakers

The data for the individual judges for each speaker are given in Table A.1. These data show which length intervals are judged more consistently by the naïve judges, and some additional information concerning variability between judges. Looking at interval length first, the right-most section gives the proportion of the total correct responses the naïve judges made (relative to the experts) separately for 1-s (first column of this section) and 5-s (middle column of this section) intervals and the signed difference between the two (5-s - 1-s). The majority of the latter signed differences are positive, which indicates that naïve judges were more consistent with the experts for long intervals.

Looking at variability across interval length (Table A.1), there appears to be higher numbers of false positives (i.e., calling FLU intervals STUT) in naïve judges' ratings of the 1-s than the 5-s intervals. For example, judge four assessing speaker three had 29 out of the total 75 1-s intervals rated as STUT, which included 17 false positives. So 17/29 (more than 58%) of this judge's STUT responses are wrong. For 5-s intervals for this same speaker and judge, 11 out of the total 15 intervals were rated as STUT, including three false positives. So 3/11 (27.27%) of this judges STUT responses were false positives for the 5-s intervals. 48 speaker by judge sets of data were available for whom this calculation was possible (there were no agreed fluent intervals judged for speakers two and four). For these data, the percentage of false positives was 52.6% for the 1-s intervals and 16.9% for the 5-s intervals which was highly significant by related t test ($t(47) = 10.5$, $p < .001$) Thus judgements about 1-s intervals are more prone to false positive STUT responses than 5-s intervals.

## 4. Discussion

The main result is that estimated stuttering rate depends on interval length with longer intervals more likely to be judged STUT. The second main finding was that the effect of interval size depended on the speaker's stuttering severity (more severe stutterers tend to have more intervals consistently judged 'stuttered' than milder ones). The experiment controlled for decison context between the expert judges that provided the criteria responses and the naïve judges by having both sets of judges assess all materials. If only expert-agreed intervals had been assessed, different range and frequency effects would have applied to the different materials and this would have affected the responses given (Parducci, 1965). The results suggest that users of TI procedures should not have free choice over interval length; otherwise the results across clinics and with different clients are not comparable. Moreover use of long intervals (5-s and over) is not recommended for detecting changes in stuttering frequency across conditions as the procedures are insensitive even to large changes in stuttering rate (this applies to Ingham et al., 1997).

The study points to major issues of reliability of the TI procedure such as difficulty in selecting a good number of samples, particularly in fluent speech, with the rejection criteria being as they are; large proportion of false positives; inability of longer intervals to measure differences in more severe stuttering. Although TI procedures can be automated and would then provide a relatively efficient method for assessing speech, these problems rule out using these procedures in clinics. TI procedures may have other uses with respect to stuttering, however. The method as applied in the current study has provided intervals which are completely fluent or contain one type of stuttering. These could be used for training and testing material for procedures that automatically count stuttering events (Howell & Huckvale, 2004). Also, the intervals can be used to establish what acoustic information is salient for detecting stutterings.

Europe PMC Funders Author Manuscripts

## Acknowledgments

This research was supported by programme grant 072639 awarded by the Wellcome Trust.

## References

Cordes AK. The reliability of observational data: I. Theories and methods for speech-language pathology. Journal of Speech and Hearing Research. 1994; 37:264–278. [PubMed: 8028308]

Curlee RF. Observer agreement on disfluency and stuttering. Journal of Speech and Hearing Research. 1981; 24:595–600. [PubMed: 7035743]

Howell P, Huckvale M. Facilities to assist people to research into stammered speech. Stammering Research. 2004; 1:130–242. [PubMed: 18418475]

Howell P, Sackin S, Glenn K. Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. Journal of Speech, Language and Hearing Research. 1997; 40:1073–1084. [PubMed: 9328878]

Howell P, Staveley A, Sackin S, Rustin L. Methods of interval selection, presence of noise and their effects on detectability of repetitions and prolongations. Journal of the Acoustical Society of America. 1998; 104:3558–3567. [PubMed: 9857514]

Ingham RJ, Cordes AK, Gow ML. Time-interval measurement of stuttering: Modifying interjudge agreement. Journal of Speech and Hearing Research. 1993; 36:503–515. [PubMed: 8331907]

Ingham RJ, Moglia RA, Frank P, Costello-Ingham J, Cordes A. Experimental investigation of the effects of frequency-altered feedback on the speech of adults who stutter. Journal of Speech, Language and Hearing Research. 1997; 40:361–372. [PubMed: 9130204]

Martin RR, Haroldson SK. Stuttering identification: Standard definition and moment of stuttering. Journal of Speech and Hearing Research. 1981; 24:59–63. [PubMed: 7253630]

Parducci A. Category judgement: A range-frequency model. Psychological Review. 1965; 72:407–418. [PubMed: 5852241]

Perkins WH. Stuttering: A matter of bad timing. Science. 2001; 294:786. [PubMed: 11681317]

Europe PMC Funders Author Manuscripts

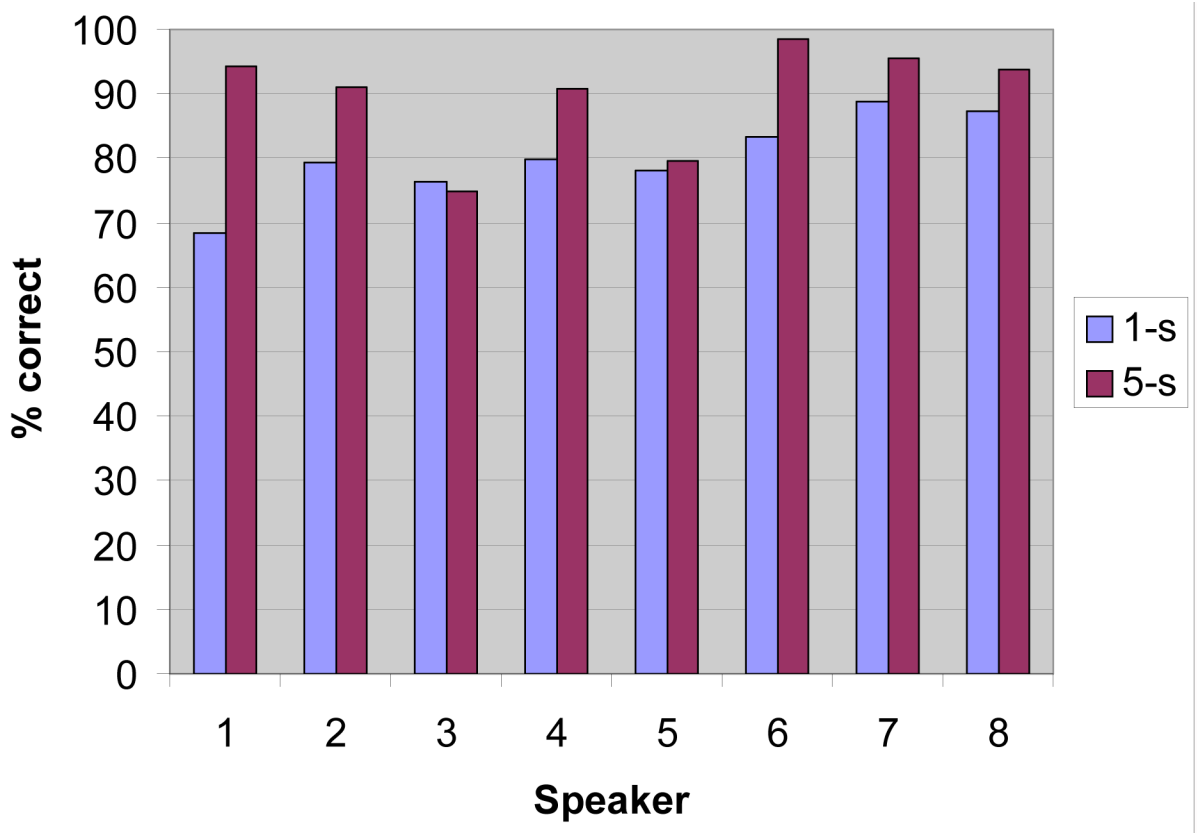


**Figure 1.**
Percent correct judgements of naïve judges (relative to experts' agreed responses). Results are shown for individual speakers (labeled along the abscissa) separately for 1-s and 5-s TIs.

Europe PMC Funders Author Manuscripts

**Table 1**

Summary of number of 5-s intervals that were excluded (under the column labeled N) and the percentage of total speech available this represents (under the column labeled %) for the exclusion criteria indicated at the left of each row.

| | N | % |
|---|---|---|
| Artefacts of segmentation process: | | |
| 5-s FLU, one of the five 1s-intervals agreed STUT | 24 | 12 |
| 5-s STUT, all of the five 1s-intervals agreed FLU | 10 | 5 |
| Ambiguous | 28 | 14 |
| Extraneous noises | 24 | 12 |
| Total | 86 | 43 |

Europe PMC Funders Author Manuscripts



**Table 2**

Speakers are labeled at the top of the Table. The first row gives the duration (in s) of the original file (based on 5-s intervals) and results for across speakers (labeled N). The second row indicates the duration (again in s) after all exclusion criteria were applied. The amount of speech lost (in s) is given in the third row (i.e. the difference between what was available initially and after the exclusion criteria were applied) and row four gives this as a percentage of the total material available. Row five gives the TIs that were agreed to be STUT (in s) for the data after the exclusion criteria were applied and row six represents this as a percentage of all material (row two).

| | **Speaker** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **N** |
| Initial duration | 125 | 130 | 125 | 125 | 120 | 130 | 125 | 120 | 1000 |
| After exclusion | 55 | 105 | 75 | 75 | 55 | 40 | 85 | 50 | 540 |
| Lost | 70 | 25 | 50 | 50 | 65 | 90 | 40 | 70 | 460 |
| %age lost | 56 | 19 | 40 | 40 | 54 | 69 | 32 | 58 | 46 |
| TIs where judged STUT | 40 | 105 | 45 | 75 | 5 | 20 | 70 | 45 | 405 |
| %age stuttered | 73 | 100 | 60 | 100 | 9 | 50 | 82 | 90 | 75 |

Europe PMC Funders Author Manuscripts

**Table 3**

Mean percentage correct and standard deviation across judges for 1-s and 5-s intervals and for each speaker.

| | 1-s | | 5-s | |
|---|---|---|---|---|
| **Speaker** | **Mean** | **SD** | **Mean** | **SD** |
| S1 | 68.41 | 6.87 | 94.32 | 4.70 |
| S2 | 79.40 | 2.83 | 91.07 | 5.37 |
| S3 | 76.33 | 3.62 | 75.00 | 4.71 |
| S4 | 79.83 | 3.93 | 90.83 | 4.95 |
| S5 | 78.18 | 5.99 | 79.55 | 4.42 |
| S6 | 83.44 | 5.17 | 98.44 | 4.42 |
| S7 | 88.82 | 2.52 | 95.59 | 5.21 |
| S8 | 87.25 | 3.37 | 93.75 | 5.18 |

Europe PMC Funders Author Manuscripts





**Table A.1**

The three sections (going from left to right) give results for 1-s, and for 5-s, TIs and comparisons between 5-s and 1-s TIs. Speaker (S1-S8) and naïve judge (J1-J8) are labeled at the left of each row. The three rows under the two sections headed 'Correct/total 1-s' and 'Correct/total 5-s', are, form left to right, number of responses the naïve judge got correct relative to the expert for a) fluent TI (FLU), b) stuttered TI (STUT) and c) all TI (A). The section headed 'Comparison of 5-s and 1-s' gives (going left to right), the total proportion of TI that were judged correct for 1) 1-s, and 2) 5-s intervals and 3) the signed difference between the 5-s and 1-s proportions.

| Correct/total 1-s | | | | Correct/total 5-s | | | Comparison 5-s and 1-s | | |
|---|---|---|---|---|---|---|---|---|---|
| | **FLU** | **STUT** | **A** | **FLU** | **STUT** | **A** | **PropA 1-s** | **PropA 5-s** | **DiffA (5-s - 1-s)** |
| S1J1 | 39/49 | 3/6 | 42/55 | 8/8 | 3/3 | 55/55 | 0.76 | 1.00 | 0.33 |
| J2 | 33/49 | 4/6 | 37/55 | 8/8 | 2/3 | 50/55 | 0.67 | 0.91 | 0.24 |
| J3 | 39/49 | 3/6 | 42/55 | 7/8 | 3/3 | 50/55 | 0.76 | 0.91 | 0.15 |
| J4 | 34/49 | 3/6 | 37/55 | 8/8 | 3/3 | 55/55 | 0.67 | 1.00 | 0.33 |
| J5 | 35/49 | 4/6 | 39/55 | 8/8 | 3/3 | 55/55 | 0.71 | 1.00 | 0.29 |
| J6 | 34/49 | 3/6 | 37/55 | 8/8 | 2/3 | 50/55 | 0.67 | 0.91 | 0.14 |
| J7 | 33/49 | 4/6 | 37/55 | 7/8 | 3/3 | 50/55 | 0.67 | 0.91 | 0.14 |
| J8 | 26/49 | 4/6 | 30/55 | 8/8 | 2/3 | 50/55 | 0.55 | 0.91 | 0.36 |
| | | | | | | | | | |
| S2J1 | 33/42 | 52/63 | 85/105 | | 20/21 | 100/105 | 0.81 | 0.95 | 0.14 |
| J2 | 35/42 | 53/63 | 88/105 | | 19/21 | 95/105 | 0.84 | 0.90 | 0.06 |
| J3 | 31/42 | 47/63 | 78/105 | | 17/21 | 85/105 | 0.74 | 0.80 | 0.06 |
| J4 | 33/42 | 50/63 | 83/105 | | 20/21 | 100/105 | 0.79 | 0.95 | 0.16 |
| J5 | 34/42 | 50/63 | 84/105 | | 19/21 | 95/105 | 0.80 | 0.90 | 0.10 |
| J6 | 33/42 | 52/63 | 85/105 | | 20/21 | 100/105 | 0.81 | 0.95 | 0.14 |
| J7 | 32/42 | 49/63 | 81/105 | | 20/21 | 100/105 | 0.77 | 0.95 | 0.18 |
| J8 | 33/42 | 50/63 | 83/105 | | 18/21 | 90/105 | 0.79 | 0.86 | 0.07 |
| | | | | | | | | | |
| S3J1 | 46/60 | 13/15 | 59/75 | 4/6 | 7/9 | 55/75 | 0.79 | 0.73 | 0.06 |
| J2 | 50/60 | 12/15 | 62/75 | 5/6 | 5/9 | 50/75 | 0.83 | 0.67 | 0.16 |
| J3 | 43/60 | 11/15 | 54/75 | 4/6 | 7/9 | 55/75 | 0.72 | 0.73 | 0.01 |
| J4 | 43/60 | 12/15 | 55/75 | 3/6 | 8/9 | 55/75 | 0.73 | 0.73 | 0.00 |
| J5 | 44/60 | 12/15 | 56/75 | 5/6 | 7/9 | 60/75 | 0.75 | 0.80 | 0.05 |
| J6 | 46/60 | 13/15 | 59/75 | 5/6 | 6/9 | 55/75 | 0.79 | 0.73 | 0.06 |

| | Correct/total 1-s | | | Correct/total 5-s | | | Comparison 5-s and 1-s | | |
|---|---|---|---|---|---|---|---|---|---|
| | FLU | STUT | A | FLU | STUT | A | PropA 1-s | PropA 5-s | DiffA (5-s - 1-s) |
| J7 | 45/60 | 13/15 | 58/75 | 6/6 | 6/9 | 60/75 | 0.77 | 0.80 | 0.03 |
| J8 | 43/60 | 12/15 | 55/75 | 5/6 | 7/9 | 60/75 | 0.73 | 0.80 | 0.07 |
| | | | | | | | | | |
| S4J1 | 32/39 | 30/36 | 62/75 | | 13/15 | 65/75 | 0.83 | 0.87 | 0.04 |
| J2 | 30/39 | 26/36 | 56/75 | | 15/15 | 75/75 | 0.75 | 1.00 | 0.25 |
| J3 | 31/39 | 30/36 | 61/75 | | 13/15 | 65/75 | 0.81 | 0.87 | 0.06 |
| J4 | 33/39 | 30/36 | 63/75 | | 13/15 | 65/75 | 0.84 | 0.87 | 0.03 |
| J5 | 29/39 | 26/36 | 55/75 | | 14/15 | 70/75 | 0.73 | 0.93 | 0.20 |
| J6 | 31/39 | 28/36 | 59/75 | | 14/15 | 70/75 | 0.79 | 0.93 | 0.14 |
| J7 | 32/39 | 30/36 | 62/75 | | 13/15 | 65/75 | 0.83 | 0.87 | 0.04 |
| J8 | 31/39 | 30/36 | 61/75 | | 14/15 | 70/75 | 0.81 | 0.93 | 0.12 |
| | | | | | | | | | |
| S5J1 | 39/53 | 0/2 | 39/55 | 7/10 | 1/1 | 40/55 | 0.71 | 0.73 | 0.02 |
| J2 | 47/53 | 2/2 | 49/55 | 9/10 | 1/1 | 50/55 | 0.89 | 0.91 | 0.02 |
| J3 | 43/53 | 1/2 | 44/55 | 8/10 | 1/1 | 45/55 | 0.80 | 0.82 | 0.02 |
| J4 | 42/53 | 1/2 | 43/55 | 7/10 | 1/1 | 40/55 | 0.78 | 0.73 | -0.05 |
| J5 | 40/53 | 2/2 | 42/55 | 8/10 | 1/1 | 45/55 | 0.76 | 0.82 | 0.06 |
| J6 | 40/53 | 1/2 | 41/55 | 8/10 | 0/1 | 40/55 | 0.75 | 0.73 | -0.02 |
| J7 | 39/53 | 1/2 | 40/55 | 7/10 | 1/1 | 40/55 | 0.73 | 0.73 | 0.00 |
| J8 | 44/53 | 2/2 | 46/55 | 9/10 | 1/1 | 50/55 | 0.84 | 0.91 | 0.07 |
| | | | | | | | | | |
| S6J1 | 26/34 | 6/6 | 32/40 | 4/4 | 4/4 | 40/40 | 0.80 | 1.00 | 0.20 |
| J2 | 30/34 | 6/6 | 36/40 | 4/4 | 4/4 | 40/40 | 0.90 | 1.00 | 0.10 |
| J3 | 27/34 | 6/6 | 33/40 | 4/4 | 4/4 | 40/40 | 0.83 | 1.00 | 0.17 |
| J4 | 28/34 | 6/6 | 34/40 | 4/4 | 4/4 | 40/40 | 0.85 | 1.00 | 0.15 |
| J5 | 25/34 | 5/6 | 30/40 | 4/4 | ¾ | 35/40 | 0.75 | 0.88 | 0.13 |
| J6 | 26/34 | 6/6 | 32/40 | 4/4 | 4/4 | 40/40 | 0.80 | 1.00 | 0.20 |
| J7 | 30/34 | 6/6 | 36/40 | 4/4 | 4/4 | 40/40 | 0.90 | 1.00 | 0.10 |
| J8 | 28/34 | 6/6 | 34/40 | 4/4 | 4/4 | 40/40 | 0.85 | 1.00 | 0.15 |

Europe PMC Funders Author Manuscripts



| | Correct/total 1-s | | | Correct/total 5-s | | | Comparison 5-s and 1-s | | |
|---|---|---|---|---|---|---|---|---|---|
| | **FLU** | **STUT** | **A** | **FLU** | **STUT** | **A** | **PropA 1-s** | **PropA 5-s** | **DiffA (5-s - 1-s)** |
| S7J1 | 46/53 | 29/32 | 75/85 | 2/3 | 14/14 | 80/85 | 0.88 | 0.94 | 0.06 |
| J2 | 49/53 | 27/32 | 76/85 | 3/3 | 14/14 | 85/85 | 0.89 | 1.00 | 0.11 |
| J3 | 48/53 | 29/32 | 77/85 | 2/3 | 13/14 | 75/85 | 0.91 | 0.93 | 0.02 |
| J4 | 46/53 | 29/32 | 75/85 | 3/3 | 14/14 | 85/85 | 0.88 | 1.00 | 0.12 |
| J5 | 49/53 | 29/32 | 78/85 | 3/3 | 14/14 | 85/85 | 0.92 | 1.00 | 0.08 |
| J6 | 44/53 | 27/32 | 71/85 | 3/3 | 12/14 | 75/85 | 0.84 | 0.93 | 0.09 |
| J7 | 46/53 | 29/32 | 75/85 | 2/3 | 14/14 | 80/85 | 0.88 | 0.94 | 0.06 |
| J8 | 48/53 | 29/32 | 77/85 | 3/3 | 14/14 | 85/85 | 0.91 | 1.00 | 0.09 |
| | | | | | | | | | |
| S8J1 | 28/33 | 14/17 | 42/50 | 0/1 | 9/9 | 45/50 | 0.84 | 0.90 | 0.06 |
| J2 | 30/33 | 14/17 | 44/50 | 1/1 | 8/9 | 45/50 | 0.88 | 0.90 | 0.02 |
| J3 | 28/33 | 13/17 | 41/50 | 1/1 | 8/9 | 45/50 | 0.82 | 0.90 | 0.08 |
| J4 | 30/33 | 13/17 | 43/50 | 1/1 | 8/9 | 45/50 | 0.86 | 0.90 | 0.04 |
| J5 | 29/33 | 16/17 | 45/50 | 0/1 | 9/9 | 45/50 | 0.90 | 0.90 | 0.00 |
| J6 | 30/33 | 13/17 | 43/50 | 1/1 | 9/9 | 50/50 | 0.86 | 1.00 | 0.14 |
| J7 | 29/33 | 16/17 | 45/50 | 1/1 | 9/9 | 50/50 | 0.90 | 1.00 | 0.10 |
| J8 | 30/33 | 16/17 | 46/50 | 1/1 | 9/9 | 50/50 | 0.92 | 1.00 | 0.08 |

Note: As indicated in the text, there were cases where 5-s intervals were agreed to be STUT by the expert judges which contained one or more 1-s intervals that were agreed by the experts to be FLU (though at least one 1-s interval was expert-agreed STUT). This arose when the experts agreed that there was a stuttering of less than 5-s in length. This results in agreed FLU 1-s TI from the expert judges for subjects 2 and 4 where there were no agreed FLU 5-s TIs.