# *A Comparison of Listener and Speaker Perception of Stuttering Events*

# *Comparaison du bégaiement que perçoivent des auditeurs et des locuteurs*

*John A. Tetnowski*
*University of Louisiana at Lafayette*
*Lafayette, Louisiana*
*and*
*Anne J. Schagen*
*Portland Public Schools*
*Portland, Oregon*

**Abstract**

**This study compared judgments of stuttering made by six persons who stutter (PWS) with the highly agreed upon judgements of stuttering made by eight independent listeners. Listeners' judgements were based on at least seven of the eight judges agreeing on whether stuttering occurred or not on a given segment. Judgements of whether stuttering occurred were made immediately by the speakers as suggested by Moore and Perkins (1990). Comparisons of speakers' judgements of stuttering to listeners' judgements of stuttering revealed a Cohen's Kappa of 0. 276 (significant beyond the .001 level). Analysis of results shows high listener/speaker agreement on nonstuttered words (as identified by speakers), but low listener/speaker agreement on stuttered words (as identified by speakers).**

**Abrégé**

**La présente étude compare les jugements sur le bégaiement posés par six personnes qui bégaient elles-mêmes avec ceux sur lesquels s'entendaient largement huit auditeurs indépendants. Au moins sept des huit juges devaient être d'accord sur la présence ou l'absence de bégaiement dans un segment donné pour que l'on retienne leur jugement. À la suggestion de Moore et Perkins (1990), les locuteurs ont jugé sur-le-champ s'ils avaient bégayé ou non. Les comparaisons des avis des deux groupes révèlent une concordance Kappa de Cohen de 0,276 (largement supérieure au niveau 0,01). Une analyse des résultats indique que les auditeurs et les locuteurs s'entendent largement sur les mots qui n'ont pas été bégayés (selon ce qu'ont dit les locuteurs), mais qu'ils s'entendent peu quant aux mots qui ont été bégayés (selon ceux qu'ont identifiés les locuteurs).**

**Key words: stuttering, stuttering identification, self-perception, speaker perception, listener identification**

The assessment of stuttering is most commonly based upon judgements of stuttering made by listeners, typically, speech-language pathologists. Words or syllables perceived as stuttered are tallied and compared to the total number of words or syllables spoken. Despite the widespread use of this technique, agreement as to where stuttering actually occurs is quite low among judges (Curlee, 1981; MacDonald & Martin, 1973; Young, 1975, 1977). More recent attempts at improving the reliability of stuttering judgements come from a series of studies by Ingham and Cordes (Cordes & Ingham, 1994a, 1995, 1996, 1999; Cordes, Ingham, Frank, & Ingham, 1992; Ingham & Cordes, 1997; Ingham, Cordes, & Gow, 1993). These studies used a time-interval analysis to identify stuttering, rather than the more traditional method of identifying each stuttered word or syllable within a speech sample (i.e., an "event-based" approach). The studies by Ingham, Cordes, and their colleagues show higher reliability between judgements made

by listeners, but are based on five-second intervals of speech, rather than on individual words or syllables. Despite the improved reliability, the clinical and practical utility of this technique does not warrant its acceptance at this time (Yaruss, 1997). The results of studies that examine listener reliability continue to raise serious questions about the validity of observational data when evaluating or treating persons who stutter (PWS).

An alternative method of stuttering identification, one that has yet to receive a great deal of investigation, is based on the speaker's own perception of their stuttering. There is some theoretical support for this view. Perkins, Kent, and Curlee (1991) propose a neuropsycholinguistic theory of stuttering that incorporates many observable and several internal (and less observable) aspects that comprise the stuttering event. These include stress, time pressure, and other feelings that can only be identified by the speaker, and only at the time of actual stuttering. Their hypothesis further claims that working memory, cognitive systems, language systems, segmental systems, speech motor control systems, and paralinguistc systems must be in synchrony in order to produce fluent speech. When all components/systems required for speech are not integrated at the same time, the speaker falls out of synchrony and disfluency occurs. The determination of whether the disfluency is a stuttered or a nonstuttered disfluency is based upon internal perceptions of the speaker related to timing and struggle behaviours. If this theory is true, the only way to identify stuttering is to know the inner perceptions of the speaker at the time when fluency breaks down. Presently, the only way to gain knowledge of this "inner feeling" about time pressure (i.e., stuttering) is to gather information from the speaker at the time of stuttering.

Researchers who support this view of stuttering indicate that it is a personal event, only identifiable by the speaker (Moore & Perkins, 1990; Perkins, 1990; Perkins et al., 1991). They believe that valid measures of stuttering can only be made by speakers because only speakers have access to inner feelings of "time pressure." Listeners do not have access to the same information, therefore, when we compare judgements of stuttering made by listeners and judgements of stuttering made by speakers, judgements of stuttering (speakers versus listeners) are based on different input. Presently, this requires two different methods of data collection. Since the listener judgement technique mirrors current clinical practice, it is fruitful for investigation. Speaker judgement research is considerably more limited. As such, the following section will review methods of stuttering identification by speakers and listeners.

## Speaker Judgements of Stuttering

Recent works that explore speaker judgements of stuttering and their validity are based on the works of Moore and Perkins (1990) and Perkins (1990). According to Moore and Perkins (1990), the only method that allows for accurate and valid judgements of stuttering by the speaker is for those judgements to be made immediately by the speaker (i.e., "on-line", or in "real-time"). Surprisingly, few studies that consider on-line judgements of stuttering by a speaker are reported in the literature. Moore and Perkins conducted one of the few controlled studies in this area. Their study involved a single participant who was a confirmed stutterer. The experimenters recorded the participant's speech during a reading task and asked her to alert the experimenters as to when she stuttered. The participant signaled that she had stuttered by pressing a button that provided a signal (red light) to the experimenters. The experimenters then provided a signal (green light) to the participant to continue talking. Moore and Perkins hypothesized that if the participant could not continue speaking without difficulty, then her stuttering must be "authentic". In 100% of the instances when the participant signaled the experimenters that she had stuttered, she could not continue speaking without a continuation of stuttering behavior. Again, based on the participant's inability to continue speaking, these on-line judgements of stuttering were deemed valid and authentic. Later in the experiment, the participant listened to recorded samples of her stuttering and was asked to imitate the stuttering as closely as possible. After she produced the "faked" stuttering, she paused and continued speaking upon receiving a signal from the experimenter (green light). Since the participant could continue speaking following the "faked" stuttering episodes, these speaking situations were considered to be valid examples of "faked" stuttering. The participant was later asked to read the same set of readings again. During this reading, when the participant signaled to the experimenter that she had stuttered, the experimenter played a recorded speech sample for the participant to hear. The sample played back to the participant was either the stuttered recording that she had just produced, or a recording of "faked" stuttering that had been recorded in the previous segment of the study. The participant was then asked to determine whether the recording was "real" or "faked" stuttering. The speaker was able to identify 93% of her "real" stuttered utterances when

reviewed within one minute of their actual production. The same procedure was completed on the next day and she could only identify 73% of her "real" stuttering events. These findings indicate that a speaker is capable of identifying their own stuttering most accurately within a short time period following its actual occurrence (i.e., speaker identification of stuttering is most accurate when completed on-line).

## Listener Judgements of Stuttering

A final phase of the experiment by Moore and Perkins (1990) tested the stuttering identification skills of a group of listeners who heard recorded utterances of the participant and were asked to identify which utterances contained "real" stuttering, and which utterances contained "faked" stuttering. Listeners could accurately determine episodes of "real" stuttering with only 54 % accuracy, considerably lower than the accuracy levels judged by the speaker. Moore and Perkins (1990) conclude that only a speaker can be an accurate judge of their own stuttering, especially if made at the time of speaking or very shortly thereafter.

Recently, another problem involving listeners making "on-line" judgements of stuttering has been brought to light. Research has indicated that on-line perceptual judgements of stuttering by listeners are slightly less accurate than judgements of stuttering made from repeated listenings of audio and video-taped segments of speech while following a transcript of the participant's speech (Yaruss, Max, Newman, & Campbell, 1998). Therefore, methods for stuttering identification by listeners should allow the listeners to hear recordings several times (rather than listening to speech only once, as is necessary in on-line identification).

At least one study has compared self-judgements of speakers who stutter and judgements of stuttering made by listeners using a time-interval method (Ingham & Cordes, 1997). This study compared only consistent judgements of stuttering made by speakers with judgements of stuttering made by "expert" listeners. Expert listeners were drawn from a pool of speech-language pathologists who, through their research and clinical experiences, were deemed as experts by the authors. Judgements of stuttering were based on five-second intervals of speech and not individual words or syllables. The study required speakers' judgements of stuttering to be made as they spoke (i.e., on-line), while listeners' judgements were made as they listened to recorded samples of speech. The on-line judgements by the speakers were based on only one judgement (only one judgement can necessarily be made for each speaker), while the experts' judgements were based on the availability of more than one-time listening. The study revealed agreement levels of over 70% between speakers and "expert" judges. Although the 70% agreement levels are much better than the agreement levels of earlier studies, the results used time-interval analysis of speaking, rather than the much more commonly used event-based (i.e., syllable-by-syllable) judgements. [1]

In a related study, experimenters attempted to use only "exemplars" of stuttering when exploring listener judgement of stuttering (Cordes & Ingham, 1996). The results indicate that listener judgements of stuttering are more reliable when using "highly agreed upon" judgements of stuttering (i.e., exemplars). Once again this study was completed using time interval judgements of stuttering. However, the Moore and Perkins (1990) article remains the key study that measured a speaker's self-identification of stuttering based on words or syllables stuttered (event-based) and compared their reliability with judgements made by listeners.

Finally, a brief study reported in the literature by Martin and Haroldson (1986) also looked at listener and speaker agreement on when stuttering occurred. Their study recorded a speaker reading a short passage. The speaker pushed a button when they experienced a "loss of control" which was recorded along with the reading. Listeners then listened to the recording and were instructed to push a button each time the participant stuttered. The signals of the listeners were then compared with the signals of the speakers (i.e., if the button was activated within +/- one second of each others' judgements of stuttering or loss of control). Results indicate 69%, 67%, 67%, 62%, and 55% agreement, respectively, across five readings of the same passage. It appears that 60-70% agreement is about the expected level of agreement when comparing judgements of stuttering between speakers and listeners.

### Listener Training

One final issue deals with how listeners are trained. Research has shown that researchers and judges trained at alternate clinical settings are likely to identify stuttering differently (Cordes & Ingham, 1995; Ham, 1989; Kully & Boberg, 1988). This factor is likely to effect judgements of stuttering by a group of listeners. Thus, in order to gain maximum agreement, listeners in a group study should all be trained by the same method. Based on this issue, as well

as concerns regarding judgements made by the speaker who stutters and those of independent listeners, the intent of the present study was to compare highly agreed upon listener judgements of stuttering with speaker judgements of stuttering while using the traditional event-based (word-by-word) approach. The specific purposes of this experiment included the following: (a) to investigate the correlation between a speaker's self-judgement of stuttering and highly agreed upon judgements of stuttering by listeners, and (b) to investigate the significance of that correlation.

## Method

### *Participants*

Six persons who stuttered (PWS) served as participants in this study. Participants ranged in age from 18 to 47 years. From this point forward, these participants will be referred to as speakers. Each speaker had a moderate or greater level of stuttering severity based upon the Stuttering Severity Instrument for Children and Adults (SSI-3, Third Edition; Riley, 1994). All speakers were free of other speech and language deficits based on clinical records and subjective judgements made by the researchers during casual conversation. Speaker age, gender, and severity of stuttering information are summarized in Table 1.

The judges, referred to as "listeners", were eight graduate students currently enrolled in the Speech and Hearing Sciences Program at Portland State University. All listeners had completed a training sequence as part of their graduate course in stuttering disorders. The training sequence was based on two weeks of class time (three times, 50 minutes per week). During this time, the listeners learned how to identify stuttering and practiced identifying stuttering while observing videotapes of stuttering events. None of the listeners had a personal history of stuttering. All listeners passed a bilateral hearing screening at 20 dB HL for 500, 1,000, 2,000, and 4,000 Hz.

**Table 1. Speakers' Gender, Age, and Stuttering Severity.***

| Speaker | Gender | Age | Stuttering Severity |
|---|---|---|---|
| 1 | M | 18 | Moderate |
| 2 | M | 47 | Moderate |
| 3 | M | 28 | Moderate |
| 4 | M | 39 | Moderate |
| 5 | M | 28 | Moderate |
| 6 | M | 30 | Severe |

* Stuttering severity was determined through the use of the Stuttering Severity Instrument for Children and Adults (Riley, 1994).

### *Procedure*

Each speaker read a list of 25 sentences that were generated for this study. The sentences contained all phonemes represented in the English language. Semantic and syntactic complexity was not controlled. The order of presentation for reading was randomly varied among participants. Each sentence was printed twice on a single sheet of paper. The upper sentence was printed in a large font (36 point) with normal spacing and punctuation. The lower sentence was printed in the same fashion, but contained slashes between words, so that the sentence was broken down into single words, a space between each word and a space prior to the first word. The wording and spacing was designed in this fashion following methods used by MacDonald and Martin (1973); an example of the sentences is provided below:

The dog is barking at the little boy.

/ /The / /dog / /is / /barking/ / at / /the / /little/ /boy/.

Each word and space was termed a decision point. The 25 sentences generated a pool of 342 potential stuttering points (words or spaces between words) for each speaker, or a total of 2,052 points for the six speakers combined.

### *Speaker Perception Task*

Each speaker was scheduled for an individual reading time. All speakers volunteered to participate in this study. No monetary or other reward was provided for participation in this study. Before beginning the task, speakers were asked to read in their "normal" voice. Speakers were given a list of sentences and asked to read them into a microphone in a normal manner. The speakers were aware that they were being recorded. Speakers were requested to stop after reading each single sentence and then mark each sheet with a red "X" over any word, or any space between words

where they thought they stuttered. This task was completed after each individual sentence was finished, based upon earlier findings that indicated speakers are most accurate in identifying their stuttering at a time very close to the time when they had produced it (Moore & Perkins, 1990). After completing one sentence, the speaker was cued to move on to the next sentence, and so on. Specific directions or training of how to identify stuttering were not provided to the speakers. They were simply asked to identify when they believed they had produced a stuttering event.

### *Listener Perception Task*

Listeners were scheduled at individual times to review the recordings. All listeners were graduate students in a speech and hearing sciences program, had normal hearing, and had completed stuttering identification training as part of a graduate course in stuttering disorders. All listeners volunteered to participate in this study. No monetary or other reward was provided for participation in this study. As noted by past studies, this method should improve listener agreement (Cordes & Ingham, 1995; Ham, 1989; Kully & Boberg, 1988). Listeners were given the same reading script that the speakers received and were instructed to mark a red "X" over any word or space between words where they perceived a stuttering event. No definition of stuttering was provided to the listeners at this time. They were simply asked to identify when they heard a stuttering event (approximately the same directions that the speakers received). Listeners reviewed the sentences in a quiet room and were allowed to listen to a sentence as often as they liked before making their judgements. Each listener reviewed all 25 sentences generated by each of the six speakers during an individual session.

### *Instrumentation*

All recordings were made in a sound treated room on a Digital Audio Tape (DAT) recorder (SONY, Model PCM-2300) on high quality digital tape. Mouth-to-microphone distance was kept steady at a distance of 20 cm using a unidirectional microphone (Audio-Technica).

Listeners reviewed and identified stuttering events from the recordings in a quiet room using the same audiotape equipment. Recordings were played through a high quality amplifier and high quality studio monitor speakers. Individual listeners were allowed to adjust the volume to a level that was comfortable.

### *Analysis Technique*

The responses of both speakers and listeners were transferred to a spreadsheet with potential stuttering points numbered on the left margin. Before the final analysis was completed, all points of low agreement were eliminated. That is, only stuttered points that were highly agreed upon were considered for analysis. This was done by eliminating all points where more than one judge disagreed as to whether stuttering occurred, or whether the point contained no stuttering. In other words, at least seven of the eight judges must have agreed whether a word or space was stuttered or not stuttered by the speaker. This technique (use of exemplars) was used to eliminate as many unreliable judgements of stuttering as possible (Cordes & Ingham, 1996).

Those judgements that were highly agreed upon by listeners were compared to the judgements of stuttering or nonstuttering made by the participants at the time of their reading. Analysis was made using Cohen's Kappa (Cohen, 1960), an agreement index based in probability theory that yields a measure of reliability to control for the likelihood of chance agreement. Cohen's Kappa is the ratio of observed agreements (expressed as a proportion), less the expected chance agreements (also expressed as a proportion), divided by the total possible agreement less the proportion expected by chance. The formula is expressed as: $k = o - c / 1 - c$, where, "o" is the observed agreement expressed as a proportion and "c" is the proportion of agreement expected by chance. Therefore, it is an estimate of true agreement. Cohen's Kappa is an index of agreement measured as a proportion corrected for inflation due to chance. It is an estimate of reliability (uninflated) based on probability theory (Cohen, 1960) and has been recommended as an appropriate analysis technique for interpreting listener agreement of stuttering (Lewis, 1994).

Since the purpose of the present study was focused on assessing agreement between on-line judgements of stuttering made by speakers, and perceptual judgements of stuttering made by listeners, two slightly different methods of data collection were employed. The selection of methods was justified because we were interested in obtaining the most reliable means of identifying stuttering in an event-based method. This required the speaker's task to be a single identification of stuttering made by the PWS at the time of stuttering (Moore & Perkins, 1990). The listener task allowed for multiple listenings of tape recorded samples and a transcript (Yaruss, Max, Newman, & Campbell, 1998).

Further, the present study employed the use of only highly agreed upon judgements of stuttering (exemplars) when comparing listener to speaker judgements of stuttering (Cordes & Ingham, 1996). Finally, training biases were reduced by using only listeners who were trained through the same procedure (Cordes & Ingham, 1995; Ham, 1989; Kully & Boberg, 1988).

## Results

A total of 1,970 judgements were analyzed. The judgements not analyzed were due to less than seven of eight agreements by the listeners. One sentence was also eliminated due to Speaker 1 skipping one sentence completely. Thus, 82 total points were eliminated (70 points where at least seven of the eight listeners did not agree and another 12 points due to Speaker 1 skipping one sentence).

The Cohen's Kappa value generated for this study was 0.276 with a level of significance of $p = .001$, indicating highly significant agreement between speakers' and listeners' perception of where stuttering occurred. The results of this study show a significant correlation between speaker judgements of stuttering and highly agreed upon listener judgements of stuttering.

Exploration of the data can reveal the utility of this study. Of the 1,970 points analyzed, listeners agreed (i.e., at least seven of eight listeners) that 1,953 points were not stuttered and only 17 points were stuttered. Speakers, in turn, judged 1,910 points to be not stuttered and 60 points to be stuttered. For this study, the speakers judged more stuttering than the listeners perceived. This was the case even though stringent agreement criteria (at least seven out of eight listeners agreed) were used.

**Table 2. Total and Percentage of Stuttered Points Marked by Speakers.**

| Speaker | Stuttered Points | Total Points | % Stuttered Points |
|---|---|---|---|
| S1 | 13 | 330 | 3.9 |
| S2 | 13 | 342 | 3.8 |
| S3 | 5 | 342 | 1.5 |
| S4 | 4 | 342 | 1.2 |
| S5 | 3 | 342 | 0.9 |
| S6 | 47 | 342 | 13.7 |

Occurrences of stuttering, as identified by the speaker, are listed in Table 2. Despite the severity levels of individual speakers, only one speaker showed a substantial

**Table 3. Total Stuttering Points and Means Marked by Listeners and Points Marked as Stuttering by the Speakers.**

| | Listener | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Speaker** | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | Mean | Self |
| S1 | 10 | 10 | 7 | 6 | 6 | 2 | 15 | 5 | 7. 63 | 13 |
| S2 | 11 | 13 | 11 | 11 | 2 | 2 | 17 | 11 | 9. 75 | 13 |
| S3 | 7 | 10 | 7 | 8 | 9 | 6 | 9 | 8 | 8. 00 | 5 |
| S4 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0. 63 | 4 |
| S5 | 3 | 11 | 3 | 3 | 2 | 0 | 10 | 1 | 4. 13 | 3 |
| S6 | 21 | 41 | 20 | 29 | 11 | 21 | 29 | 23 | 24. 38 | 47 |
| **Total** | **53** | **86** | **48** | **58** | **30** | **31** | **81** | **49** | | |

amount of stuttering (13. 7%) based on his own perception of instances of stuttering. It should be noted that the percentage of stuttering is lower than might be expected. This can be explained by the method of calculation. Speakers were given the opportunity to identify stuttering on either a word, or a space between words, whereas the methods for counting stuttering in the SSI-3 (Riley 1994) only count the number of syllables spoken. Therefore, the percentage of stuttering shown in Table 2 is approximately one-half of the percentage of stuttered syllables that we would expect to calculate in measures like the SSI-3 (Riley, 1994).

Comparisons of individual listeners' judgements of stuttering are summarized in Table 3. Four of the six speakers (S1, S2, S4, S6) perceived more stuttering than the mean number of stuttering episodes observed by listeners. Even when considering individual listeners, only one listener (L7) perceived more stuttering than Speaker 1 had judged for himself; only one listener (L7) perceived more stuttering than Speaker 2; all eight listeners perceived more stuttering than Speaker 3; no listeners perceived more stuttering than Speaker 4; two listeners (L2, L7) perceived more stuttering than Speaker 5, and no listeners perceived more stuttering than Speaker 6. In most cases, there was more unobserved stuttering by the listeners than self-perceived stuttering by the speaker. One individual speaker and one individual listener were the exceptions to this trend. Speaker 3 (S3) iden-

**Table 4. Speakers' Perceptions of Nonstuttered (N) and Stuttered (S) Points.**

| | | | *Number of Speaker's Judgments in Agreement with Listeners' Judgments (LA)* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LA=0 | LA=1 | LA=2 | LA=3 | LA=4 | LA=5 | LA=6 | LA=7 | LA=8 |
| | | **Speaker** | | | | | | | | | |
| **Speaker** | **N/S** | **Total** | | | | | | | | | |
| S1 | N | 317 | 0 | 0 | 3 | 4 | 0 | 2 | 2 | 5** | 301** |
| | S | 13 | 7 | 4 | 2 | 0 | 0 | 0 | 0 | 0* | 0 |
| S2 | N | 329 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 9** | 317** |
| | S | 13 | 2 | 0 | 1 | 0 | 1 | 2 | 5 | 2* | 0 |
| S3 | N | 337 | 0 | 1 | 1 | 0 | 0 | 2 | 2 | 7** | 324** |
| | S | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4* | 0* |
| S4 | N | 338 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0** | 336** |
| | S | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0* | 0 |
| S5 | N | 339 | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 4** | 321** |
| | S | 3 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0* | 0* |
| S6 | N | 295 | 0 | 5 | 4 | 2 | 2 | 1 | 8 | 16** | 257 |
| | S | 47 | 21 | 10 | 6 | 0 | 1 | 2 | 2 | 5* | 0* |

* Indicates when at least seven out of eight judges agreed with the speaker that stuttering had occurred.
** Indicates when at least seven out of eight judges agreed with the speaker that stuttering had not occurred.

tified less stuttering in his own speech than any of the individual listeners. In addition, Listener 7 (L7) identified more stuttering than the speaker for four of six individual speakers. Finally, Listener 6 (L6) indicated far fewer episodes of stuttering than most other judges.

The number of listeners who agreed with judgements by speakers is summarized in Table 4. Inspection of this table indicates strong agreement where stuttering did not occur. For example, of the 317 judgements of "nonstuttered" speech (N) for Speaker 1 (S1), 301 segments were also judged as "nonstuttered" by all eight of the listeners, and five more were judged as "nonstuttered" by seven of eight listeners. In other words, at least seven out of eight judges agreed with the speaker on 306 out of the 317 "nonstuttered" judgements of Speaker 1. Further inspection of the Table 4 shows a similar trend for all speakers. A comparison of "stuttered" judgements (S) by speakers and the number of listeners that agreed can also be seen in Table 4. Seven or more judges agreed where stuttering occurred on zero out of 13 judgements for Participant 1; two out of 13 judgements for Participant 2; four out of five judgements for Participant 3; zero out of four judgements for Participant 4; zero out of three judgements for Participant 5; and five out of 47 judgements for Participant 6. Not a single segment that was identified as "stuttered" by the speakers was judged as stuttering by all eight listeners. The total percentage of agreement overlap between judges and speakers is still only 12. 9% (11 out of 85 judgements of stuttering where at least seven out of eight judges agreed with the self-perception of the speaker). The higher agreement between judges and speakers as to when stuttering *did not* occur (1904 out of 1955 judgements of not stuttering, or 97. 4%) appears to be a much easier task. Levels of agreement on where stuttering did not occur are quite high, but levels of agreement on where stuttering did occur were quite low. This agreement of when stuttering *did not* occur inflated the Cohen's Kappa figure of 0.276.

The conclusion drawn from data presented in Tables 3 and 4 indicates that event-based stuttering identification, a commonly used element in stuttering severity ratings, is extremely different when it is a self-perception (generated by the speaker), versus when perception is generated by a listener. This holds true, even when the listener-perception measure is based only on highly agreed upon judgements of stuttering (agreed on by at least seven out of eight listeners).

## Discussion

This study measured the concordance, or agreement of judges between speakers' self-perception of instances of stuttering and listeners' perceptions of instances of stuttering from the same data set. Using only highly agreed upon judgements of stuttering, the correlation between listener and speaker judgements of stuttering revealed a Cohen's Kappa of 0.276; this measure is significant beyond $p < .05$.

The results of this study do not support earlier findings by Moore and Perkins (1990) that listener and speaker judgements of stuttering do not coincide with each other. Judgements of stuttering/nonstuttering do coincide beyond a chance level. There are several possible explanations for these findings. One of the reasons may be the use of "exemplars," that is, highly agreed upon judgements of stuttering. Research has clearly shown that more definitive examples of stuttering will improve listener agreement (Cordes & Ingham, 1996). Another reason for the significant correlation may be the high level of agreement on "nonstuttered" words. It appears that the Cohen's Kappa level of 0.276 may be inflated due to agreement points where stuttering *did not* occur. Previous studies in stuttering identification only compared agreement on stuttered words, and did not factor in the potential agreement on nonstuttered words. As with past studies, there still is low agreement (12. 9%) on the stuttered words (if that is all we are attempting to count).

### *Differences Between Speaker and Listener Judgements of Stuttering*

Despite the overall level of agreement, differences between speaker and listener judgements of just "stuttering" did appear in our study. This finding coincides with earlier studies that found listeners did not agree where stuttering occurred in the speech of stutterers (Curlee, 1981; MacDonald, & Martin, 1973; Young, 1975, 1977). Untrained clinicians, working clinicians, and even "experts" in stuttering continue to work at low levels of reliability when trying to identify stuttering events (Cordes & Ingham, 1995; Ham, 1989; Ingham & Cordes, 1997). These results particularly hold true for judgements made in event-related tasks (rather than in the "time-interval analysis" methods preferred by Ingham, Cordes, and their colleagues). It should be noted that time-interval analyses have shown higher reliability coefficients, but these ratings are based upon whether stuttering was or was not present during a five-second interval of connected speech, rather than on single

words or syllables. The results of our study further support the extreme difficulty that exists in obtaining consistent listener-based perceptions of where stuttering occurs (Cordes, 1994; Cordes & Ingham, 1994b).

Some might argue that that no listener related method for identifying stuttering is valid (Moore & Perkins, 1990; Perkins, 1990; Perkins et al., 1991). This problem with validity could contribute to the low agreement levels that exist in most stuttering identification studies. It is possible that self-perception of stuttering may be the only valid means of assessing stuttering. Simply put, our definition of stuttering may be flawed. Wingate's (1964) definition of stuttering is still widely accepted as a standard definition of stuttering; however, definitions like this place considerable emphasis on a listener's perception of what is truly stuttering (Martin & Haroldson, 1986). Perkins (1990) and Perkins et al. (1991) have made a case for a speaker's perception of stuttering as being the only valid method of stuttering identification. Moore and Perkin's (1990) study supported this view, but was conducted on only a single participant. However, this area of inquiry would appear to be an important area for further research.

It is interesting to note that most listeners identified considerably less stuttering than the speakers (see Table 3). This could be a result of listeners not having access to the "inner feelings" of stuttering that are only available to speakers (Moore & Perkins, 1990; Perkins, 1990; Perkins et al., 1991). We might expect that listeners would only be able to identify the most visible or the most audible stuttering behaviours. This is supported by the higher agreement levels obtained when only using "exemplars" of stuttering (Cordes & Ingham, 1996). It is likely that these exemplars are simply the most obvious episodes of stuttering; however, more subtle stuttering is still difficult to identify.

### *Individual Listener Differences*

The individual listener differences noted earlier (please refer to Table 3) are difficult to explain, since all listeners were trained in the same manner. For example, Listener 6 (L6) identified significantly fewer stuttering points than did other listeners. L6 identified a lower number of stuttering episodes than five of the six speakers. A potential cause could include a more stringent criteria for stuttering that L6 adopted, despite the training received earlier by all listeners. This appears to be the case since L6 identified the least stuttering for five out of six speakers (S1, S2, S3, S4, S5). During their training, individual listeners were not required to show competency at stuttering identification tasks. It is also possible that training listeners (Martin & Haroldson, 1986), or using only listeners that were held to a high standard of intra-rater agreement could eliminate these individual differences (Ingham et al., 1993). At least one report has indicated that we can predict which listeners will be the most accurate judges of stuttering behaviours (Tetnowski, Ham, & Walker, 1994). Knowledge of these listener characteristics under controlled conditions could help eliminate individual outliers, such as L6. Larger groups of listeners could also help in eliminating the effects of a single listener on data obtained. Future studies that compare individual listener ratings of stuttering could help us understand the difficulty involved in the identification of stuttering.

## Conclusions

The listener/speaker agreement levels for stuttering behaviours can be improved through the use of specific methods (e.g., using only highly agreed upon judgements of stuttering by listeners when comparing to speaker judgement of stuttering). Researchers like Ingham and Cordes (1997) have comprehensively explored the concept of listener and speaker agreement of stuttering. Their results showed higher than expected agreement levels between "experts" in the field of stuttering and judgements made by the speakers. It should be noted again that all of these stuttering judgements were made on five-second intervals of connected speech, and not on the more commonly used word-by-word or syllable-by-syllable (event-related) identification methods. Our study did not seek to argue whether the time analysis measures, or the event-related method of stuttering identification method is better. It merely sought to compare listener versus speaker judgements of stuttering. The fact remains that we still cannot obtain high degrees of agreement on stuttering identification tasks as long as we do not inflate agreement levels through counting "nonstuttered" events. More research should continue in the area of speaker perception of stuttering, including the use of time-interval techniques, and any other techniques that can assist in the accuracy of this extremely difficult task.

### *A Caveat*

In summary, the authors would like to refer to a 1990 article published by Oliver Bloodstein, which we think makes a valid point. In his article, "On pluttering, skivering, and floggering," Bloodstein states that if we can define a

behaviour, we should be able to count it. Not until we have a valid statement as to what stuttering is, can we count and measure it accurately. It is certainly possible to gain high reliability on a concept that is not highly valid. In other words, our struggle to gain more reliable techniques for stuttering identification may be 'barking up the wrong tree' (Martin & Haroldson, 1986). We acknowledge that the significant levels of agreement found in our study may be due to the mechanics of the analysis. Efforts were made to increase reliability as much as possible (use of only highly agreed upon judgements, use of only judges that were trained in the same way, etc.). The problems of reliability may really be a problem with our search for a valid definition of stuttering.

We can only reach validity through newer and more refined techniques, or we will continue to argue the same arguments over and over about the nature of stuttering. Alternative techniques may allow more input by the speaker and, thus, allow greater understanding of stuttering. Qualitative research techniques, such as conversational analysis, ethnographic interviewing, and lamination have been used successfully in other areas of speech-language pathology such as aphasia (Simmons-Mackie & Damico, 1996), and in other fields such as anthropology and sociology (Goodwin, 1995, 1986; Spradley, 1979). These studies have revealed insights into the nature of communication and communication breakdowns. The use of alternative research paradigms in stuttering has only recently been suggested (Tetnowski & Damico, 2001). Techniques such as these may help us to gain a more valid and reliable means of assessing stuttering behaviours as well.

## Footnote

[1] For further arguments regarding the use of event-based judgements of stuttering, please see Yaruss (1997).

## Author Note

Please address all correspondence to John A. Tetnowski, PhD, University of Louisiana at Lafayette, Department of Communicative Disorders, PO Box 43170, Lafayette, Louisiana, 70504-3170 USA.

## References

Bloodstein, O. (1990). On pluttering, skivering, and floggering: A commentary. *Journal of Speech and Hearing Disorders, 55*, 392-393.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cordes, A. K. (1994). The reliability of observational data: I. Theories and methods for speech-language pathology. *Journal of Speech and Hearing Research, 37*, 264-278.

Cordes, A. K., & Ingham, R. J. (1994a). Time-interval measurement of stuttering: Effects of interval duration. *Journal of Speech and Hearing Research, 37*, 779-788.

Cordes, A. K., & Ingham, R. J. (1994b). The reliability of observational data: II. Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research, 37*, 279-294.

Cordes, A. K., & Ingham, R. J. (1995). Judgements of stuttered and nonstuttered intervals by recognized authorities in stuttering research *Journal of Speech and Hearing Research, 38*, 33-41.

Cordes, A. K., & Ingham, R. J. (1996). Time-interval measurement of stuttering: Establishing and modifying judgment accuracy. *Journal of Speech and Hearing Research, 39*, 298-310.

Cordes, A. K., & Ingham, R. J. (1999). Effects of time-interval judgment training on real time measurement of stuttering. *Journal of Speech, Language, and Hearing Research, 42*, 862-879.

Cordes, A. K., Ingham, R. J., Frank, P., & Ingham, J. C. (1992). Time-interval analysis of interjudge and intrajudge agreement for stuttering event judgements. *Journal of Speech and Hearing Research, 35*, 483-494.

Curlee, R. F. (1981). Observer agreement on disfluency and stuttering. *Journal of Speech and Hearing Research, 24*, 595-600.

Goodwin, C. (1995). Co-constructing meaning in conversations with an aphasic man. In S. Jacoby & E. Ochs (Eds.), *Research in language and social interaction (Special issue of Construction), 28*, 233-260.

Ham, R. E. (1989). What are we measuring? *Journal of Fluency Disorders, 14*, 231-243.

Ingham, R. J., & Cordes, A. K. (1997). Identifying the authoritative judgements of stuttering: Comparisons of self-judgements and observer judgements. *Journal of Speech and Hearing Research, 40*, 581-594.

Ingham, R. J., Cordes, A. K., & Gow, M. L. (1993). Time-interval measurement of stuttering: Modifying interjudge agreement. *Journal of Speech and Hearing Research, 36*, 503-515.

Kully, D., & Boberg, E. (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders, 13*, 309-318.

Lewis, K. E. (1994). Reporting observer agreement on stuttering event judgments: A survey and evaluation of current practice. *Journal of Fluency Disorders, 19*, 269-284.

MacDonald, J. D., & Martin, R. R. (1973). Stuttering and disfluency as two reliable and unambiguous response classes. *Journal of Speech and Hearing Research, 16*, 691-699.

Martin, R. R., & Haroldson, S. K. (1981). Stuttering identification: Standard definition and moment of stuttering. *Journal of Speech and Hearing Research, 24*, 59-63.

Martin, R. R., & Haroldson, S. K. (1986). Stuttering as involuntary loss of control: Barking up a new tree. *Journal of Speech and Hearing Disorders, 51*, 187-190.

Moore, S. E., & Perkins, W. H. (1990). Validity and reliability of authentic and simulated stuttering. *Journal of Speech and Hearing Disorders, 55*, 383-391.

Perkins, W. H. (1990). What is stuttering? *Journal of Speech and Hearing Disorders, 55*, 370-382.

Perkins, W. H., Kent, R. D., & Curlee, R. F. (1991). A theory of

neuropsycholinguistic function in stuttering. *Journal of Speech and Hearing Research, 34*, 734-752.

Riley, G. D. (1994). *Stuttering Severity Instrument for Children and Adults* (3rd ed.). Austin, TX: Pro-Ed.

Simmons-Mackie, N. N., & Damico, J. S. (1996). The contribution of discourse markers to communicative competence in aphasia. *American Journal of Speech-Language Pathology, 5*, 37-43.

Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart & Winston.

Tetnowski, J. A., & Damico, J. S. (2001). A demonstration of the advantages of qualitative methodologies in stuttering research. *Journal of Fluency Disorders, 26,* 17-42.

Tetnowski, J. A., Ham, R. E., & Walker V. G. (1994, November). *Variables associated with unit-by-unit identification of stuttering.* Poster session presented at the annual meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

Yaruss, J. S. (1997). Clinical measurement of stuttering behaviors. *Contemporary Issues in Communication Science and Disorders, 24*, 33-44.

Yaruss, J. S., Max, M. S., Newman, R., & Campbell, J. (1998). Comparing real-time and transcript-based techniques for measuring stuttering. *Journal of Fluency Disorders, 23*, 137-151.

Young, M. A. (1975). Observer agreement for marking moments of stuttering. *Journal of Speech and Hearing Research, 18*, 530-540.

Young, M. A. (1977). An extension of a familiar index of observer agreement. *Journal of Speech and Hearing Research, 20*, 72-80.

*Manuscript received: March 21, 2000*
*Accepted: February 11, 2001*