Identifying the authoritative judgments of stuttering: Comparisons of self-judgments and observer... By: Ingham, Roger J., Cordes, Anne K., Journal of Speech, Language & Hearing Research, 10924388, Jun97, Vol. 40, Issue 3

IDENTIFYING THE AUTHORITATIVE JUDGMENTS OF STUTTERING: COMPARISONS OF SELF-JUDGMENTS AND OBSERVER JUDGMENTS

Reliable and accurate stuttering measurement depends on the existence of unambiguous descriptions or exemplars of stuttered and nonstuttered speech. The development of clinically meaningful and useful exemplars, in turn, requires determining whether persons who stutter judge the same speech to be stuttered that other observers judge to be stuttered. The purpose of these experiments, therefore, was to compare stuttering judgments from several sources: 15 adults who stutter, judging their own spontaneous speech; the same adults who stutter, judging each other's speech; and a panel of 10 authorities on stuttering research and treatment. Judgments were made under several conditions, including self judgments made while the speaker was talking and self- and other-judgments made from recordings in continuous and interval formats. Results showed substantial differences in stuttering judgments across speakers, judges, and judgment conditions, but across-task comparisons were complicated by low self agreement for many judges. Some intervals were judged consistently by all judges to be Stuttered or Nonstuttered, across multiple conditions, but many other intervals were either not assigned replicable judgments or were consistently judged to be Nonstuttered by the speaker who had produced them but were not assigned consistent judgments by other judges. The implications of these findings for stuttering measurement are considered.

KEY WORDS: stuttering, measurement, self-judgments, agreement, intervals

Recurring concerns about the clinical significance of perceptual judgments of stuttering (e.g., Cooper, 1986; Goldberg, Culatta, Ingham, Cooper, & Brutten, 1992; Ham, 1989), and growing evidence that judgments will differ across observers (see, e.g., Cordes & Ingham, 1994a, 1995; Ham, 1989; Ingham, Cordes, Ingham, & Gow, 1995), imply that standardized judgment procedures and training programs must be developed if researchers and clinicians are to be able to make valid, consistent, and useful judgments of stuttering. Some recent studies have used time-interval based measures in basic research about the nature of stuttering (e.g., Smith et al., 1993), and several studies of stuttering measurement have shown that interval-judgment procedures may lead to markedly better inter- and intrajudge agreement than that usually obtained with procedures requiring the identification of individual stuttering events or stuttered syllables (see Cordes & Ingham, 1994a; Ingham, Cordes, & Gow, 1993). The same studies are also demonstrating, however, that interval measurement itself, without judge training, does not solve the problems of reliability and validity for stuttering. One difficulty with these mixed results is that development of judge training programs presents a circular problem: Trusted referents for stuttered and nonstuttered speech must be developed as a prerequisite to developing training programs, but it is not clear how to identify those referents without training programs.

One potential source for standard or referent judgments of stuttering might be agreed exemplars of stuttered speech recorded from a variety of persons who stutter (see Cordes & Ingham, 1996). Precisely whose agreement should be sought to identify these exemplars, however, is a question

that raises another series of issues. Inexperienced judges might be selected, to provide judgments representative of those that an average, untrained observer would regard as stuttering. Inexperienced observers, however, have been shown repeatedly to display poor intrajudge and interjudge agreement for many stuttering judgment tasks (see Cordes & Ingham, 1994a; Ingham & Cordes, 1992; Young, 1975, 1984), making them poor sources for exemplar judgments. Experienced speech-language pathologists or researchers, whose judgments would be based on specialized academic or clinical knowledge of stuttering, might also be selected. Many experienced judges display exceptionally high self-agreement, and experienced judges often display superior levels of interjudge agreement when compared with less experienced judges (Cordes, 1995; Cordes & Ingham, 1995; Ingham, 1995). Experienced observers, however, including authorities at well-known research clinics, have also been shown to have poor levels of agreement among themselves in identifying stuttering (Cordes & Ingham, 1995; Ham, 1989; Ingham & Cordes, 1992; Kully & Boberg, 1988). These differences across judges, and certainly across research centers, limit both the number and the usefulness of agreed exemplars that can be identified from judgments made by authoritative observers.

Judges who stutter are another possibility as the source for exemplar intervals. Their judgments might be argued to carry face validity because of their unique experience and perspective on stuttering. Surprisingly, there is very little information available about judgments of stuttering made by persons who stutter.[1] Tuthill's (1946) original study of stuttering judgments included 20 persons who stuttered as judges (judging speech from other people). Agreement levels were poor overall, with clinicians, inexperienced judges, and judges who stuttered showing similar levels of self-agreement (between 49% and 56.6%). A few previous studies have also investigated judgments of stuttering made by persons who stutter about their own speech. Martin and Haroldson (1986), for example, found large differences for some speakers between the speakers' judgments of "loss of control" while orally reading and the counts of words that an observer identified as stuttered. Some studies on the effects of self-delivered contingencies for stuttering have also included reports that some subjects who stutter, but not all, may display very low agreement with judgments of stuttering made by clinicians or researchers (see Cordes & Ingham, 1994a).

The notion that persons who stutter will provide the most valid source of judgments about their own stuttering has been forcefully argued by Perkins (1983,1990; Perkins, Kent, & Curlee, 1991). Indeed, Perkins claims that the only person qualified to identify stuttering is the speaker, because observers have no access to the definitive sensation of loss of control over the ability to produce an utterance. Data interpreted as showing a lack of perceived (Moore & Perkins, 1990) or acoustic (Kelly & Conture, 1988) differences between faked and real stuttering have been reported previously, thereby supporting claims that the distinctive features of stuttering may not be found in the speech but must be found in the speaker's experience. Kelly and Conture's analyses and conclusions, however, ignored the possibility of significant differences in the variance produced by the two classes of events, while Moore and Perkins' (1990) study was based on only one speaker. More importantly, neither of these studies compared judgments of stuttering that persons who stutter made while speaking with judgments they made while observing recordings of their own speech. In other words, it has not yet been established whether or not persons who stutter identify the same behaviors as stuttered while speaking and while observing a videotape of themselves speaking. This question has critical implications for a self-

judged theory of stuttering: If the events identified in real time as stuttered are also identified while observing after the fact, then arguments about the exclusive validity of self-judgments while speaking become largely moot.

The several experiments included within the present study provide one test of claims that stuttering judgments made by speakers who stutter will differ fundamentally from judgments made by observers. These experiments were designed to advance the search for valid referent judgments of stuttering by examining the stuttering judgments made under several different judgment conditions by persons with different types of experience with stuttering. The specific purposes of these experiments included the following: (a) to investigate whether persons who stutter will identify the same speech events as stuttering in their own speech while speaking and while observing an audiovisual recording of themselves speaking; (b) to investigate the effects of requiring persons who stutter to repeat their judgments of stuttering, using audiovisual recordings of their own speech, until they met a stability (intrajudge agreement) criterion; and (c) to compare judgments made by persons who stutter about their own speech, judgments made by persons who stutter about each other's speech, and judgments made by highly experienced authorities on stuttering research and treatment.

Method

These studies included three separate but interdependent experiments, as shown in Table 1. The Concurrent Judgments Experiment investigated self-judgments of stuttering made under three different conditions: one real-time judgment task (Real-Time Task) while subjects were speaking and two tasks using videorecordings (Interval Task and Continuous Task). The Stability Experiment further investigated self-judgments of stuttering made from videorecordings. Finally, the Multispeaker Experiment investigated off-line judgments of stuttering made by two groups of judges with different types of extensive experience with stuttering: judges who stuttered themselves, and judges who were researchers or clinicians who studied stuttering.

Subjects Concurrent Judgments Experiment: Speakers I

Five adults who had stuttered since early childhood, all males, participated in this experiment. None had any other complicating neurological or other problems, and none had received any treatment for their stuttering within 5 years of this study. All were paid volunteers who had responded to an advertisement requesting subjects for studies of stuttering measurement and treatment.

Stability Experiment: Speakers II

The 5 adults who had participated in the Concurrent Judgments Experiment, plus an additional 10 adults who stuttered, served as both speakers and judges. The additional participants were 8 men and 2 women who met all other selection criteria as described above.

Multispeaker Experiment: Speakers II and Authorities

The 15 adults who had participated in the Stability Experiment, plus a group of 10 researchers and clinic directors with extensive experience with stuttering, participated in the Multispeaker Experiment. The same researchers and clinicians had participated in an earlier study of stuttering judgments (Cordes & Ingham, 1995). All had substantial publication records and recognized expertise in the area of stuttering, and all had documented histories of approaching stuttering from a perspective that required observer identification of stuttering behaviors (see Cordes & Ingham, 1995, for additional detail).

Materials

Judgment stimuli for the Stability and Multispeaker Experiments were developed from initial audiovisual recordings of each speaker that were made during the Concurrent Judgments Experiment (for Speakers I) or during an initial recording session (for the additional 10 speakers). Each subject sat alone in a sound-attenuated booth, wearing a lavalier microphone and facing a Hi-8 digital video camera, and was recorded during an uninterrupted spontaneous monologue. All recordings showed the subject's head and shoulders only.

One 5-min recording from each subject was subsequently transferred to a laser videodisk. Custom software, in conjunction with a Pentium system computer, allowed access to the laserdisk either (a) in a continuous playback format, which presented a 5-min sample without interruption, or (b) in an interval-recording format, which presented the 60 consecutive nonoverlapping 5-s intervals of speech from a 5-min sample. All judgments on the laserdisk samples were recorded via a mouse button press using customized software for later analysis. For the Multispeaker Experiment, a videotape was developed that contained 360 5-s intervals of recorded speech, each followed by a 3-s silent interval. These intervals were the first 24 intervals from each 5-min speech sample on the videodisk, or the first 2 min of each sample.

Procedure Concurrent Judgments Experiment

The first session of the Concurrent Judgments Experiment combined the audiovisual recording task, described above, with the Real-Time Task. Individual speakers (n = 5) were given instructions that included the following:

While you are talking we want you to record (by pressing a mouse button) each and every occasion of stuttering. If you stutter then press the mouse button as soon as that stutter begins and hold it down throughout the duration of that stutter. You should try to hold the button down throughout each stuttering. Of course, sometimes your stutterings might be so brief that you will only have time to press the button down and release it almost immediately, which is fine.

Stuttering will not be defined for you. You should be aware, however, that not all disruptions or interruptions in speech are stutterings. Some "disfluencies" are quite normal and acceptable in the speech of both persons who stutter and persons who do not stutter. Please do not count normal or acceptable disruptions as stuttering. We are interested in your best judgments about whether you believe your speech contained stuttering; normal disfluencies should not be judged as stutterings.

The purpose of the experiment was also carefully explained to each subject, with special emphasis on the importance of obtaining an accurate record of stuttering judgments. Speakers were told that it would be important for the experimenters to learn whether subjects could perform this task, and that they should inform the experimenter if they did not feel that their button presses accurately reflected their perception of their stuttering. It was further explained that the cost to the project of obtaining false or inaccurate judgments was considerable, including damaging the possibility of being able to train judges to identify stuttering behavior accurately. After this explanation, the participant was provided with exposure to this speaking task for increasingly longer periods (30 s; 60 s; 150 s; 300 s), with increases contingent on his decision that during two consecutive speaking trials his judgments had remained accurate. The final recording from the 300-s condition (5 min) provided the final speech sample from each speaker.

Approximately 3-6 weeks after completing the RealTime Task, each speaker returned to the same experimental setting to complete the two off-line judgment tasks (Continuous Task and Interval Task). These tasks were completed within one session, with a break of approximately 10 min between tasks. Task order was randomized across subjects.

The Continuous Task required judgments to be made from the speaker's uninterrupted 5-min recording.[2] Instructions were essentially the same as those used for the Real-Time Task, except that subjects were making judgments about the 5-min sample of their own speech that had been recorded in the first session, rather than about their own ongoing speech. Subjects were also instructed that they would judge the recording four times and would then be given a rest, and that they might then be asked to repeat the task. The Continuous Task was repeated at least four times, until trial-to-trial variability met a stability criterion that was defined as not more than 5% variation in self-agreement levels. Judges repeated the Continuous Task at least four times and until there was not more than a 5% variation in agreement between, for example, trials 4 and 5 as compared with the agreement between trials 3 and 4. Subjects could exhibit any absolute level of trial-to-trial agreement or disagreement in order to meet the criterion, as long as that level was neither improving nor worsening: If there was little variability in the level of trial-to-trial agreement, or no increasing or decreasing trend in agreement, then it was assumed that the judge had settled on his best or most stable judgments. Subjects were told that they would repeat each task on some variable number of trials, rather than being made explicitly aware of the stability criterion. Self-agreement was derived from an interval-by-interval analysis made possible by customized software that superimposed 5-s intervals onto the judgments made during each trial with each 5-min recording (see Data Analysis). The intervals were identical to those used in the Interval Task (see below).

The Interval Task was also an audiovisual judgment task, except that the speaker's 5-min sample was presented to the subject as 60 5-s intervals, in chronological order and with a 3-s blank interval after each speech interval. The judgment procedure has been described previously (e.g., Cordes & Ingham, 1994b, 1996): The subject pressed a mouse button during each 3-s blank recording interval to indicate whether the preceding 5s interval was judged to contain stuttering or not contain stuttering. The instructions for this task were, in all other respects, identical to those used in the Continuous Task. The Interval Task was also repeated until the 5% stability criterion was met for trial-to-trial agreement.

Stability Experiment

Speakers in the Stability Experiment who had not participated in the Concurrent Judgments Experiment (n = 10) began with an initial recording session as described above (see Materials). Approximately 4-6 weeks later, after the laser videodisk had been prepared, they returned to complete the Continuous Task and the Interval Task, as described above. For the subjects who had participated in the Concurrent Judgments Experiment (Speakers I), the Continuous and Interval conditions of that study provided data for comparisons in the Stability Experiment.

Multispeaker Experiment

Approximately 4-6 weeks after the Stability Experiment was completed, each speaker returned to observe the Multispeaker videotape (see Materials). Instructions for the Multispeaker Task were similar to those used for the Interval Task, except that the judges watched the tape only once, instead of being required to meet a stability criterion. Judges were also instructed that they were to rewind individual intervals if they were unsure of their judgments; that is, the level of repetition was changed from the set of all intervals to the individual interval. All judges returned on a final occasion approximately 8 weeks later to repeat the entire Multispeaker Task with the 360 stimulus intervals in a different random order.

The Multispeaker Experiment also included judgments made by previously identified authorities on stuttering research and treatment (see Cordes & Ingham, 1995). Following the procedures used in that study, each authoritative judge was mailed a copy of the Multispeaker videotape, plus written instructions, response forms, and a survey about judgment conditions. The authorities made judgments for each interval following the same instructions used by the speaker-judges (i.e., modified Interval-judgment instructions that allowed rewinding of individual intervals). Authorities were also asked, approximately 8 weeks later, if they would be willing to repeat the task; all agreed and completed a second set of judgments with a different randomization of the 360 intervals.

Data Analyses

All conditions of this study were designed to identify the most accurate and most stable judgments possible. Speakers were required to report that their judgments were accurate in the Real-Time Task; judges in the Stability Experiment were required to meet a stability criterion for intrajudge agreement; and judges in the Multispeaker Experiment were instructed to rewind individual intervals as often as necessary to be confident of their judgments. Analyses were completed interval-by-interval, in terms of the 60 5-s intervals available from each of the 15 speakers (900 total intervals; for additional detail about analysis methods see Cordes & Ingham, 1994b; Cordes, Ingham, Frank, & Ingham, 1992; Ingham et al., 1993). As mentioned earlier, analyses for the Continuous Task were derived by superimposing 5-s intervals onto the judgments made for each recording and then converting those judgments into a record of stuttered intervals (intervals that included any part of a stuttering event judgment) and nonstuttered intervals (intervals that did not include any part of a stuttering event judgment). For the RealTime Task, there was necessarily only one judgment per interval from each speaker, but more than one judgment was available from each judge for each interval in all other tasks.

Therefore, for the Continuous and Interval Tasks, most data analyses were based on judgments made in the final two trials of each task, and for the Multispeaker Task, analyses were made of judgments from both occasions. Within these tasks, data were first summarized from all judges, and then an interval-by interval agreement criterion of 90% or better for the two trials or occasions was used to identify judges who displayed satisfactory self-agreement. Across-task comparisons were made only for individual judges who met this criterion level of self-agreement within tasks, in order to maximize the validity of across-task comparisons. Final analyses compared the three potentially "authoritative" sets of stuttering judgments: those provided by the speakers about their own speech, those provided by other speakers who stutter, and those provided by clinical and academic authorities on stuttering.

Results Concurrent Judgments Experiment

The Concurrent Judgments Experiment investigated self-judgments of stuttering made under three different conditions: one real-time judgment task (Real-Time Task) while subjects were speaking and two tasks using videorecordings (Interval Task and Continuous Task).

Judgments of Stuttering

All speakers learned to make stuttering judgments during their own spontaneous speech that they reported to be accurate representations of their perception of their own stuttering. This required either one or two 2-hour sessions, for a total of between 2 and 4 hours per speaker. All speakers required at least one return to shorter trials as they progressed from 30-s trials to 300-s trials (see Method). Each speaker confirmed that his button presses during his final 5-min recording reflected accurately his judgments about the occurrence and duration of stuttering during those 5 min.

Stuttered and nonstuttered self-judgments from this study are summarized in Table 2. During the Real-Time Task, the speakers judged 131 of the 300 total intervals (5 speakers x 60 intervals) to be Stuttered (S) and 169 to be Nonstuttered (N). Continuous Task judgments resulted in only 109 S intervals and 161 N intervals, with 30 intervals judged inconsistently across the last two trials (stuttered in the second to last and nonstuttered in the last, or vice versa) and labeled disagreed (D). Interval Task judgments identified fewer S intervals (98) than either Real-Time or Continuous Task judgments and fewer D intervals (23) than Continuous Task judgments; proportionately more intervals were judged N in the Interval Task.

Table 2 also shows that only two judges, S1 and S4, achieved self-agreement of 90% or better for both offline tasks (at least 54 agreed intervals, or fewer than 6 Disagreed intervals, out of the total of 60). One judge (S5) achieved 93.3% self-agreement for the Continuous Task (56 agreed intervals), with a lower value for the Interval Task; another (S2) achieved 93.3% self-agreement for the Interval Task but less than 90% self-agreement for the Continuous Task. The fifth judge, S3, did not achieve 90% self-agreement for either task. Comparisons across the three self-judgment tasks, presented below, were made only for judges who achieved 90% agreement or better in the relevant conditions, in order to maximize the validity of any conclusions drawn (see Method).[3]

Across-Task Comparisons

Three speakers met the intrajudge agreement criterion for Real-Time and Continuous Task judgments, the two tasks that required self-judgments of uninterrupted speech (Table 3). S1 made essentially identical judgments in these two tasks, but S4 and S5 showed noticeable differences across these two tasks. Only 5 of the 8 intervals that S4 had judged nonstuttered in RealTime (62.5%) were so judged in the Continuous Task, for example, and only 11 of the 17 intervals that S5 judged stuttered in Real-Time (64.7%) were judged stuttered in the Continuous Task.

Comparisons between Real-Time judgments and offline Interval judgments, and between the two off-line tasks (Continuous and Interval) are also shown in Table 3. Judge S1 displayed high consistency across all tasks. Judge S4 showed relatively low consistency across tasks for Real-Time S intervals when those judgments were compared with Intervals judgments, just as he had shown for the Real-Time versus Continuous Task comparison (only 41 of 52 Real-Time S intervals were also judged S in the Interval Task). S2 also showed relatively low consistency across tasks for Real-Time S intervals (19 of his 27 Real-Time S intervals were actually judged Nonstuttered in the Interval Task), but he did show relatively high consistency across tasks for Real-Time N intervals (32 of his 33 Real-Time N intervals were also judged Nonstuttered in the off-line Interval Task). S4 judged only 50% of his Interval-Task N intervals as Nonstuttered during the Continuous Task (although other comparisons for S4 are much higher).

Overall, information from Tables 2 and 3 suggests that only Speaker S1 could be considered to have consistently made the same judgments in all three of these tasks. One other subject, S4, displayed satisfactory selfagreement within both the Interval Task and the Continuous Task. The judgments made by S4 in the Interval Task, however, differed from those he made in the Real-Time and Continuous Tasks, with the difference appearing to be toward judging more intervals as nonstuttered in the Interval Task.[4]

Stability Experiment

As described above, this experiment assessed the effect of repeated judgments on the speaker's level of judgment agreement.

Table 4 summarizes the judgments made by each judge in the first two trials of each condition (Continuous and Interval Tasks; see Method) and in the last two trials of each condition (i.e., for the last two trials of those that met the stability criterion). It appears from this table that repeating the judgment task until a stability criterion was met had essentially no effect: The number of agreed Stuttered or Nonstuttered intervals did not change substantially for individual judges from the first two trials to the last two. It also appears that the number of agreed intervals identified in the Continuous Task was similar to the number of agreed intervals identified in the Interval Task (although judges S2, S6, and S12 did show at least 10% more agreed intervals in the Interval Task than in the Continuous Task). The apparent similarities were confirmed by Wilcoxon T tests (with protected alpha levels of .05/3 = .0167) that found no significant differences between first and last pairs of trials in the Continuous Task (T_{obs} = 52.5), between first and last pairs for

the Interval Task (T_{obs} = 36.5), or between the last pairs in the Continuous Task as compared with the last pairs in the Interval Task (T_{obs} = 21).[5]

Multispeaker Experiment

Finally, the Multispeaker Experiment collected judgments made by the speakers about their own speech, judgments made by persons who stutter about the speech of other persons who stutter, and judgments made by recognized academic and clinical authorities. Data are summarized for all judges, with across-task comparisons made primarily for those judges who met the 90% intrajudge agreement criterion in the relevant tasks.

Speakers as Judges of Their Own Speech

Table 5 summarizes the self-judgments made within the Multispeaker Task and compares them with judgments made for the same intervals within the Interval Task. Overall, 95.3% of these intervals were judged consistently in the last two trials of the Interval Task (179 Stuttered and 164 Nonstuttered), and 88.3% (169 Stuttered and 149 Nonstuttered) were judged consistently across the two occasions of the Multispeaker Task. This difference in agreement across the two Tasks was statistically significant (T_{obs} = 14; $T_{crit(n = 15, alpha = .025)}$ = 22).

Persons Who Stutter Judging Each Other's Speech

As judged by each speaker, the Multispeaker Task tape contained 360 intervals, 24 of his or her own speech and 336 from other speakers. Figure 1 shows that individual judges considered between 55 and 240 of the 360 intervals to be stuttered on Occasion One, and between 56 and 252 to be stuttered on Occasion Two. The first columns of Table 6 summarize these judgments by speaker, with the speakers' own self-judgments now removed from these analyses. Of the 360 total intervals, 101 were agreed by the 14 other speakers to be Stuttered and 136 were agreed to be nonstattered; thus, overall interjudge agreement for these judges was 65.8% (i.e., at least 80% of judgments were in agreement, either stuttered or nonstuttered, for 237 of the 360 intervals, or 65.8% of intervals). Intrajudge agreement was calculated for each judge for each speaker in terms of the number of intervals from that speaker (out of 24) that were assigned the same judgment on the two occasions. Individual judges met the 90% intrajudge agreement criterion for between 3 and 12 of the other 14 speakers and, as Table 6 shows, there were between 4 and 10 judges whose judgments satisfied the 90% intrajudge agreement criterion for individual speakers.

Interjudge agreement was recalculated for each speaker using only those judges who met the 90% intrajudge agreement criterion for that speaker (Table 6). Interjudge agreement was significantly higher for these judges than for the whole group of 14 judges: 73.3% of intervals agreed (111 Stuttered and 152 Nonstuttered) as compared with 65.8% (T_{obs} = 25).

Authorities' Judgments

Figure 1 also shows that individual authority judges considered between 144 and 281 of the 360 intervals to be stuttered on Occasion One, and between 143 and 254 to be stuttered on Occasion Two. The third and fourth columns of Table 6 summarize these judgments for each speaker. Of

the 360 total intervals, 166 were agreed by the authorities to be Stuttered and 109 were agreed to be Nonstuttered; thus, overall interjudge agreement for these judges was 76.4% (275 of the 360 intervals). As shown in Table 6, the judgments made by between 2 and 10 of the authority judges satisfied the 90% intrajudge agreement criterion for individual speakers. Individual authority judges achieved 90% intrajudge agreement for between 5 and 15 of the speakers (as compared with only 3 to 12 speakers for the persons who stuttered acting as judges of each other's speech; see above).

Interjudge agreement was also recalculated for each speaker using only those judges who met the 90% intrajudge agreement criterion for that speaker (Table 6). The largest interjudge agreement level in Table 6, 77.2%, was achieved by the authority judges who met the intrajudge agreement criterion, although the difference between this level and the interjudge agreement reached by the full group of authorities was not significant ($T_{obs} = 42.5$). Table 6 also shows that interjudge agreement levels were similar for the authority judges who met the 90% intrajudge agreement criterion and for the other judges who stuttered who met the 90% intrajudge agreement criterion (77.2% vs. 73.3% interjudge agreement). However, the numbers of intervals agreed to be stuttered (169 vs. 111) and nonstuttered (109 vs. 152) by the members of these two groups were significantly different ($chi^2 = 22.0$; $df = 2$; $p < .001$).

Across Task Comparisons and Identification of Agreed Exemplars

Finally, one of the purposes of this study was to combine data from the many judgment conditions to attempt to determine whose judgments might be used to identify trustworthy exemplars of stuttered and nonstuttered speech. Analyses of the speakers' self-judgments in the Concurrent Judgments and Stability Experiments had shown that unacceptably low self-agreement within conditions, and low self-agreement across conditions, often made it difficult to determine whether a speaker believed that a given interval did or did not include stuttering.[6] As shown in Table 5, for example, the speakers' own Interval Task judgments matched their judgments for the same intervals within the Multispeaker Task for as few as 6 of their 24 intervals; 7 of the 15 judges assigned the same judgments in these two Tasks to fewer than 80% of their 24 intervals. Similarly, low intrajudge and interjudge agreement levels within the Multispeaker Experiment made it difficult to determine for many intervals whether other speakers who stuttered or authorities agreed that those intervals did or did not include stuttering.

Across-task comparisons were made, therefore, using only those speakers and judges who met the 90% self-agreement criterion in the four interval-recording tasks (the Interval Task, plus self-judgments, judgments by other judges who stutter, and judgments by authorities within the Multispeaker Experiment).[7] As shown in Table 7, 7 of the 15 speakers met these criteria and were included in these analyses. Of the 168 intervals from these speakers, 58 were agreed in all four tasks to be Nonstuttered, and 60 were agreed in all four tasks to be Stuttered; thus, 65.4% of these intervals were agreed across all conditions. The patterns of disagreements across conditions, as summarized in Table 7, are also noteworthy. Of the 58 intervals that were not judged identically across all four of these tasks, 10 were judged Nonstuttered by the speakers in Interval and Multispeaker conditions, and by the other judges who stuttered, but were Disagreed by the authorities; 20 were judged Stuttered by the speakers in Interval and Multispeaker Task conditions, and by the authorities, but were Disagreed by the other judges who stuttered. For 7

intervals, the speaker assigned a judgment of Nonstuttered in all self-judgments, but neither the other judges who stuttered nor the authorities could agree that they were Nonstuttered.

Discussion

These experiments highlight several difficulties with depending exclusively on either speakers or observers to identify stuttering or to identify recorded exemplars of stuttered speech. The speakers who participated in the Real-Time Task during the Concurrent Judgments Experiment reported uniformly, after at least 2 hours of practice with increasingly longer speaking trials, that they were confident that their judgments reflected their perceptions of stuttering. All speakers who participated in the Stability Experiment met the stability criterion, providing judgments by the end of each condition that appeared to be as consistent as they could be, or that gave no suggestion that continued practice would have changed their judgments. Despite this self-reported confidence and empirical stability, however, intrajudge agreement was often unsatisfactory for many judges, a problem that made across-task comparisons essentially meaningless in many cases. Moreover, judgments from the three self-judgment conditions in the Real-Time and Stability Experiments (Real-Time, Continuous, and Interval Tasks) often differed from each other, making it difficult to validate self-judgments from the Real-Time Task, for example, by reference to other judgments. Differences between Real-Time judgments and off-line judgments might reflect fundamental differences between judging oneself while speaking and judging from a recording, of course, but it must be presumed that the RealTime and off-line judgments are reliable in order to make a meaningful comparison (see Footnote 7). Even if it could be established that Real-Time judgments were reliable there still remains the possibility that the speaker simply does not or cannot recognize certain occurrences of stuttering while absorbed in the act of speaking. Martin and Haroldson (1986) raised this possibility in their report of a study that compared judgments of "loss of control" by persons who stuttered and observer-judged occurrences of stuttering; they described instances where the speaker displayed behaviors that were almost certainly instances of stuttering that the speaker did not record.[8]

The Stability Experiment also demonstrated that repeated judgments had no significant effect on the numbers of Stuttered or Nonstuttered intervals that speakers identified in their own speech, or on the numbers of intervals agreed in consecutive trials (when data from the first two trials of each condition were compared with data from the last two trials of each condition). It had seemed a reasonable hypothesis that extended practice with a novel task, such as judging one's own stuttering from a videotape, might result in improved intrajudge and across-condition agreement, or that the reliability (and therefore the validity) of a speaker's self-judgments might be improved by the simple procedure of allowing judgments to stabilize. These hypotheses were not supported; instead, results of this study replicate and extend earlier findings that practice without feedback had essentially no effect on judgments of stuttering (e.g., Cordes et al., 1992). Response-contingent training procedures for observers have been shown to be more effective at improving agreement and accuracy levels for interval judgments of stuttering than control conditions that required repeated judgments in the absence of feedback (e.g., Cordes & Ingham, 1996; Ingham et al., 1993). The possibility exists, therefore, that training procedures might improve speakers' off-line judgments of their own stuttering in the same way, but it is not at all clear how observers could dictate to a speaker what the correct judgments about his own speech should be, when so many differences exist in the judgments made by different judges in different

conditions. The Multispeaker Experiment included in this study provided further evidence of these differences, clearly replicating the dramatic disagreements in stuttering judgments by recognized authorities that were reported earlier by Cordes and Ingham (1995). Judgments made by other judges who stutter were distributed even more widely than judgments made by the authorities (Figure 1:55-240 stuttered judgments on the first occasion for the other speakers and 144-281 for the authorities). Even among the authorities, however, judges at the top of the range categorized approximately twice as many intervals as stuttered as did judges at the bottom of the range. The Multispeaker Experiment also replicated the substantial differences across research centers, and the relative agreement within research centers, that was reported by Cordes and Ingham (1995).[9]

Overall, it seems clear that the speakers involved in the three experiments reported here responded in several ways to these self-judgment tasks. Only one speaker, S1, made consistently similar judgments while speaking and while observing his speech. He provides evidence that some speakers who stutter may identify behaviors while speaking that are the same as the behaviors they identify as stuttering while observing their speech. For speakers such as S1, therefore, who are consistent in judging their own speech, self-judgments might be given precedence over observer judgments when the two differ. For other speakers in these experiments, however, judgments made while speaking and judgments made while observing recordings differed substantially, and recording conditions (Continuous, Interval, and interval judgments within the Multispeaker Task) appeared to affect judgments as well. It was also clear that many speakers and other judges provided judgments that did not meet the intrajudge agreement criterion of 90%. Across-task comparisons of their judgments would have had little validity because of their within-task instability, and it would be difficult to argue that these speakers' judgments should be given precedence over any observer's judgments because of the self-contradictory nature of the self-judgments.

In terms of identifying exemplar intervals of stuttering, or identifying the "authoritative" judgments of stuttering, the comparisons between self-judgments and judgments made by authorities on stuttering may prove to be among the most important findings of this study. For the 7 speakers who received judgments that met the intrajudge agreement criterion, one summary of the Multispeaker Experiment might be that there was no evidence of consistent differences between self-judgments and judgments made by other groups of observers: 140 of the 158 intervals were judged the same way, either Stuttered or Nonstuttered, in at least three of the four interval-recording conditions, with the fourth being Disagreed (rather than agreed in the opposite direction). For many of these intervals, however, the judges who stuttered (both the speaker and the other judges) agreed on a judgment of Nonstuttered while the authorities were disagreed. Such a pattern suggests that behaviors labeled nonstuttered by a speaker might reasonably be considered to be nonstuttered, even if some authorities call them stuttered. There were also many intervals for which the individual speaker and the authorities agreed on a judgment of Stuttered while the other judges (who stuttered) could not agree that they were Stuttered or Nonstuttered. This pattern suggests that judgments of the presence of stuttering made by a speaker or by highly experienced researchers and clinicians should be respected even if other observers do not detect the presence of stuttering.

One additional question involves whether these intervals can be said to contain abnormal speech, or whether they received a few stuttered judgments from some judges because they resembled the normally disfluent speech produced by normally fluent speakers. A general finding emerging from many studies is that there are some perceptual differences between the stutter-free speech of persons who stutter and the speech of normally fluent speakers; it also appears that these differences may be related to speech rate (Prosek & Runyan, 1982, 1983). Far less is known about the normal disfluencies that may or may not exist in the speech of speakers who stutter. For some theorists (Bloodstein, 1995; Smith, 1990), the proposed continuity between stuttering and normal disfluencies suggests that any events that judges cannot agree are either stuttered or nonstuttered are likely to have nonnormal features that still make them distinctive. If this is the case then, arguably, there is still reason to consider the disagreed intervals as likely to have pathological features. Other theorists would argue that, although the differences may sometimes be blurred, there is reason to believe that persons who stutter produce speech events that can be unambiguously identified as stuttering and other events that can be unambiguously identified as normal disfluencies (Martin & Haroldson, 1981; Perkins, 1990). The findings in support of this position are equivocal (Curlee, 1981; MacDonald & Martin, 1973). One approach to this question would take advantage of some of the findings of this study: One possibility that the experimenters are currently exploring is to determine whether judges, experienced or inexperienced, can distinguish between those intervals that judges cannot agree are stuttered or nonstuttered and other intervals that judges agree are normal disfluencies. Early data suggest that they cannot.

Perhaps the most important issue, though, is one that needs to be addressed from a very different perspective: the level of clinical importance that is attached to what the speaker considers (reliably) to be stuttered or nonstuttered, regardless of whether it is perceived while speaking or perceived while observing speech, and regardless of the judgments made by other observers. It does appear, based on these results, that some observers and some speakers can agree completely about whether a large number of intervals are stuttered or nonstuttered, a promising result for the development of valid and reliable exemplar intervals. This finding also poses a challenge to Perkins's claims that there are, by definition, differences between the judgments of stuttering made by speakers while they are talking and the judgments of stuttering made after the fact, whether by the speaker or by some other observer. Perhaps more important, however, are the large numbers of intervals in this study that were not assigned the same judgments in all conditions. Many of them were assigned relatively consistent judgments by the speaker that were in conflict with the judgments made by some or all of the authorities; these differences between what the person who stutters sees as stuttering and what the professional sees as stuttering must be addressed if our discipline is to serve its consumers appropriately.

Acknowledgments

This paper was supported by Research Grant Number 5 R01 DC 00060-05 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. Preliminary data from some of the investigations described in this paper were included in presentations by the authors at the 1995 Annual Convention of the American Speech-Language-Hearing Association, Orlando, Florida. The authors express their sincerest appreciation to the subjects, researchers, and clinicians who participated in these studies. Special thanks to Richard

Moglia and the late Peter Frank, who provided invaluable technical and statistical support for this and other studies conducted within this grant.

1. There are probably many experimenters who stutter and who provide judgments of stuttering, but those judgments have been neither systematically documented nor compared with those from experimenters who do not stutter.
2. Recordings showed the subject's head and shoulders only, as stated above. It was also verified in pilot work that there were no visual or auditory cues on the tapes that could reveal when button presses had occurred during the Real-Time judgment task.
3. Across-task interval-by-interval comparisons for all judges, and complete data for all judges from the Multispeaker Task, are available from the authors on request.
4. It was clear from these 5 speakers that Real-Time judgments could not be consistently related to judgments made by speakers observing recordings of the same speech that they had judged in Real-Time. Arranging for the subjects who participated in the Stability Experiment to also complete the Real-Time Task would not have altered this conclusion: Some speakers may make similar judgments in Real-Time and off-line conditions, but for others there was virtually no relationship.
5. The last comparison was also completed for just the eight judges who met the intrajudge agreement criterion of 90% for the two conditions; they did not display a significant difference across conditions ($T_{obs} = 17.5$). These two tests were considered to be alternatives of each other, rather than concurrent tests: considering them concurrent tests and dividing the alpha level to .0125 would not have influenced these negative results.
6. In the absence of an independent method of verifying the accuracy of Real-Time judgments of self-perceived stuttering, the only alternative is to employ reasonable inference based on the speaker's off-line judgments. If a speaker produces consistent off-line judgments of his own stuttering then, at the very least, it is possible to relate that speaker's off-line judgments of stuttering to his real-time judgments for the same speech sample. However, if a speaker cannot make consistent off-line judgments of his own stuttering, then it is not even possible to relate off-line judgments to Real-Time judgments for the same speech sample. As discussed above, only two of the five speakers in the Concurrent Judgments Experiment made consistent off-line judgments of their stuttering. Unaddressed by this study, of course, is whether the findings might have been changed by training speakers to make consistent off-line judgments.
7. Data from a given speaker were included in these final analyses only if that speaker had met the 90% intrajudge agreement criterion for both the Interval condition and the self-judgments within the Multispeaker condition. For the other judges who stuttered and for the authorities, intrajudge agreement was established interval-by-interval and summarized speaker-by-speaker; judgments from a given speaker were included in these final analyses from those judges who met the 90% interval-by interval agreement criterion for that speaker.
8. "On nine occasions, SS held his breath for at least 4 s prior to 'exploding' a word or syllable, but did not depress his hand switch" (Martin & Haroldson, 1986, p. 188) to register a loss of speech control. Perkins (1990) has defined such self-judged occasions of loss of speech control as definitive of stuttering.
9. This study actually systematically replicated the Cordes and Ingham (1995) study and its findings. As Figure 1 shows, the distribution of the same 10 authorities' judgments on

these 360 intervals was remarkably similar to the distribution they displayed in the Cordes and Ingham (1995) study on 720 intervals from different speakers. There was, in fact, a correlation of 0.90 between the authorities' stuttered interval counts in both studies; that is, the pattern of judgments was consistent regardless of which speakers they were asked to judge. It was also noteworthy that the exceptionally high level of interval-by-interval intrajudge agreement that these judges displayed in the 1995 study (83.2-98.3%) was sustained in the present study (87.5-96.4%).

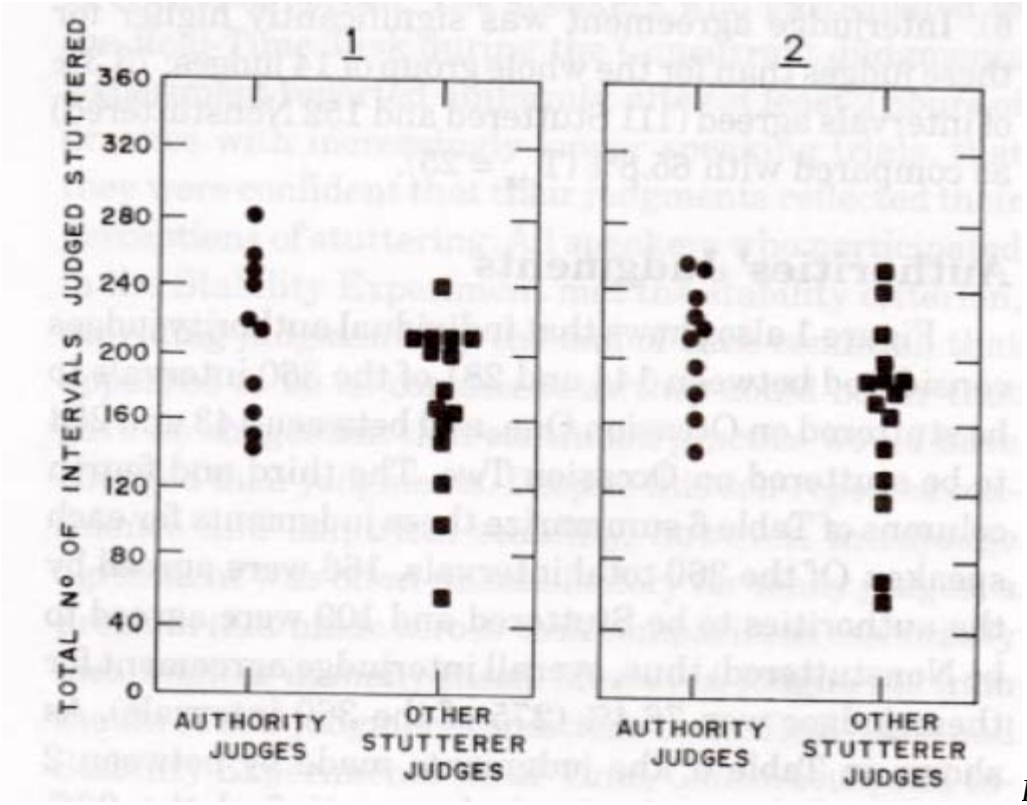

Figure 1. Number of intervals judged stuttered by each judge in the Authorities group (n = 10) and in the Other Stutterers group (n = 15) on Occasion 1 (left) and Occasion 2 (right), out of 360 intervals in the Multispeaker Task. Each data point represents the number of stuttered intervals identified by one judge.

Table 1.

Outline of the judges and tasks included in the three experiments. Speakers I refers to a group of 5 adults who stuttered; Speakers II refers to a group of those 5 plus 10 additional adults who stuttered; Authorities refers to a group of 10 experienced researchers and clinic directors.

```
Legend for Chart:

A - Real-Time Task
B - Continuous Task
C - Interval Task
D - Multispeaker Task, 2 occasions
```

	A	B	C
Concurrent Judgments Experiment	Speakers I (n = 5)	Speakers I (n = 5)	Speakers I (n = 5)
Stability Experiment		Speakers II (n = 15)	Speakers II (n = 15)
Multispeaker Experiment			Speakers II (n = 15)
			Authorities (n = 10)

Table 2.

Self-judgments of stuttering made in the Concurrent Judgments Experiment, expressed as the number of intervals that did (Stuttered) and did not (Nonstuttered) include at least part of a stuttering-duration judgment in the Real-Time Task and in the last two trials of the Continuous Task, and the number of intervals judged stuttered in both (S) or neither (N) of the last two trials of the Interval Task. Intervals assigned conflicting judgments in the last two trials of the Continuous and Interval Tasks are labeled Disagreed (see text).

	Real-Time		Continuous			Interval		
Judge	S	N	S	N	D	S	N	D
S1[ci]	8	52	9	49	2	10	49	1
S2[']	27	33	7	43	10	5	51	4
S3	27	33	30	21	9	25	28	7
S4[ci]	52	8	47	8	5	42	14	4
S5[c]	17	43	16	40	4	16	37	7

c Interval-by-interval intrajudge agreement (percent of intervals judged consistently, either stuttered or nonstuttered, in the final two trials) was above 90% for the Continuous Task.

i Interval-by-interval intrajudge agreement (percent of intervals judged consistently, either stuttered or nonstuttered, in the final two trials) was above 90% for the Interval Task.

Table 3.

Joint distributions showing numbers of intervals self-judged to be Stuttered (S), Nonstuttered (N), and Disagreed (D) in pairwise comparisons of Real-Time, Continuous, and Interval Tasks, for judges who met the 90% intrajudge agreement criterion in those Tasks (see text). Column totals (sigma) correspond to the task condition totals shown in Table 2.

	S1	S2	S4	S5

Real-Time								
Continuous	S	N		S	N	S	N	
S	8	1	(a)	45	2	11	5	
N	0	49		3	5	4	36	
D	0	2		4	1	2	2	
sigma	8	52		52	8	17	43	

Real-Time							
Interval	S	N	S	N	S	N	
S	8	2	5	0	41	1	(a)
N	0	49	19	32	7	7	
D	0	1	3	1	4	0	
sigma	8	52	27	33	52	8	

Continuous								
Interval	S	N	D		S	N	D	
S	9	1	0	(a)	39	1	2	(a)
N	0	47	2		5	7	2	
D	0	1	0		3	0	1	
sigma	9	49	2		47	8	5	

a. Empty cells represent combinations where the speaker did not meet the intrajudge agreement criterion in one or bath of the judgment tasks. Speaker S3 did not meet the intrajudge agreement criterion in either task and is not represented in this table.

Table 4.

Numbers of intervals self-judged to be Stuttered (S) and Nonstuttered (N) within the Stability Experiment, out of 60 total intervals per speaker, shown for the first two trials (First trials) in each condition and for the last two of the trials (Last trials) that met the stability criterion in each task (see text). Intervals not represented here were judged inconsistently across the two relevant trials (i.e., were Disagreed intervals). Totals are provided for all judges and for judges who met the 90% intrajudge agreement criterion.

Legend for Chart:

A - Speaker
B - Continuous Task--First trials--S
C - Continuous Task--First trials--B
D - Continuous Task--Last trials--S
E - Continuous Task--Last trials--N
F - Interval Task--First trials--S
G - Interval Task--First trials--N
H - Interval Task--Last trials--S
I - Interval Task--Last trials--N

A	B	C	D	E	F	G	H	I
S1[ci]	10	49	9	49	10	47	10	49

S2[i]	9	40	7	43	9	48	5	51
S3	33	18	30	21	26	25	25	28
S4[ci]	47	5	47	8	43	11	42	14
S5[c]	15	44	16	40	16	36	16	37
S6[i]	55	2	44	6	57	0	54	2
S7[ci]	7	51	8	50	7	53	8	52
S8[ci]	56	3	58	2	56	1	59	1
S9[c]	53	4	52	5	37	17	39	15
S10[ci]	51	6	50	6	53	7	52	7
S11	43	7	43	6	42	5	32	18
S12[i]	15	23	12	37	11	35	9	47
S13[ci]	3	53	2	54	4	50	1	55
S14[ci]	23	31	24	32	19	37	21	35
S15[ci]	53	3	54	2	53	7	51	5
Total								
All Judges	473	339	456	350	443	379	424	416
Total								
Criterion Judges	318	249	320	237	322	296	312	318

c Interval-by-interval intrajudge agreement (percent of intervals judged consistently, either stuttered or nonstuttered) was above 90% for the final two trials of the Continuous Task.

i Interval-by-interval intrajudge agreement (percent of intervals judged consistently, either stuttered or nonstuttered was above 90% for the final two trials of the Interval Task.

Table 5.

Self-judgments of stuttering for the 24 intervals per speaker that were included in the Multispeaker Task, expressed as the number of those intervals that received self-judgments of stuttered in both (S) or neither (N) of the last two trials in the Interval condition and in both (S) or neither (N) of the two occasions within the Multispeaker Task. Intervals not represented here were judged inconsistently across the two relevant trials (i.e., were Disagreed intervals in that condition).

Legend for Chart:

A - Speaker
B - Interval Task--S
C - Interval Task--N
D - Multispeaker Task--S
E - Multispeaker Task--N
G - Across-Task match[a]--n
H - Across-Task match[a]--(%)

A	B	C	D	E	F	G
S1	3	20	3	21	23	(95.8)
S2	4	20	2	16	18	(75.0)
S3	13	8	12	9	22	(91.7)

S4	18	6	19	4	22	(91.7)
S5	8	13	9	13	21	(87.5)
S6	22	0	16	3	16	(66.7)
S7	3	21	2	22	23	(95.8)
S8	23	1	20	4	21	(87.5)
S9	17	5	17	3	20	(83.3)
S10	21	2	20	2	21	(87.5)
S11	15	7	14	4	15	(62.5)
S12	2	22	3	14	16	(66.7)
S13	0	22	2	21	21	(87.5)
S14	7	16	8	12	18	(75.0)
S15	23	1	22	1	23	(95.8)

a Number (n) and percentage (%) of intervals (out of 24 total) that received the same judgment in the Interval Task and in the Multispeaker Task, either agreed Stuttered in both trials of both Tasks, agreed Nonstuttered in both trials of both Tasks, or judged Disagreed in both Tasks.

Table 6.

Numbers of intervals agreed to be Stuttered (S) and Nonstuttered (N) in the Multispeaker Experiment, expressed as the number of intervals judged stuttered or nonstuttered in at least 80% of all available judgments (two judgments per judge, one for each occasion; see Method) from the other judges who stutter and from the Authorities. Data are provided for all judges and only for those judges who met the intrajudge criterion for that speaker. The number who met the intrajudge criterion for each speaker is also shown (n).

Legend for Chart:

A - Speaker
B - Others who stutter (n = 14), S
C - Others who stutter (n = 14), N
D - Others who stutter intrajudge > 90%, S
E - Others who stutter intrajudge > 90%, N
F - Others who stutter intrajudge > 90%, (n)
G - Authorities (n = 10), S
H - Authorities (n = 10), N
I - Authorities intrajudge > 90%, S
J - Authorities intrajudge > 90%, N
K - Authorities intrajudge > 90%, (n)

A	B	C	D	E	F	G	H	I	J	K
S1	2	20	2	20	9	3	18	3	17	6
S2	1	11	1	10	5	4	8	3	8	6
S3	10	8	12	9	7	13	2	14	1	3
S4	5	0	5	0	9	18	0	21	0	6
S5	2	12	0	16	7	6	10	4	10	3
S6	9	2	15	0	6	20	0	21	0	8
S7	2	19	2	21	8	3	19	3	19	7
S8	20	2	19	2	10	20	2	20	2	10
S9	10	0	23	0	5	23	0	23	0	7
S10	14	3	4	3	14	17	3	11	4	7
Sll	4	7	1	13	4	9	2	15	0	2
S12	0	19	0	22	7	1	16	2	16	6

S13	2	16	2	19	6	1	13	1	14	6
S14	3	16	3	16	9	6	15	6	17	8
S15	17	1	21	1	7	22	1	22	1	8
Total	101	136	111	152		166	109	169	109	
	(237 = 65.8%)		(264 = 73.3%)			(275 = 76.4%)		(278 = 77.2%)		

Table 7.

Numbers of intervals judged consistently to be Stuttered or Nonstuttered in four conditions (Interval Task plus the three sections of the Multispeaker Experiment), for those speakers and judges where intrajudge agreement was 90% or better (see text).

The following chart reads as follows:

Row 1: Speaker; Stuttered; Nonstuttered

Row 2: Pattern in remaining intervals

S1[a]	2	16
3: N, except authorities D 3: inconsistent patterns		
S4[a]	5	0
13: S, except other-stutterers D 4: self-judged N, other-stutterers and authorities S or D 2: inconsistent patterns		
S7	2	19
2: N, except authorities D 1: inconsistent pattern		
S8	19	0
5: inconsistent patterns		
S10	11	1
7: S, except other-stutterers D 5: inconsistent patterns		
S13	0	13
5: N, except authorities D 3: self-judged N, other-stutterers and authorities D 3: inconsistent patterns		
S15	21	1
2: inconsistent patterns		
Total	60 58	58

a Two judges also completed the Real-Time Task; all intervals that were judged consistently across the four tasks were judged the same way in the Real-Time Task.

References

Bloodstein, O. (1995). A handbook on stuttering (5th ed.). San Diego, CA: Singular.

Cooper, E. B. (1986). Treatment of disfluency: Future trends. Journal of Fluency Disorders, 11, 317-327.

Cordes, A. K. (1995,December). The development and current status of time-interval measures of stuttering frequency: Part 2. Paper read at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL.

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

Cordes, A. K., & Ingham, R. J. (1995). Judgments 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., Frank, P., & Ingham, J. C. (1992). Time-interval analysis of interjudge and intrajudge agreement for stuttering event judgments. 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.

Goldberg, S. A., Culatta, R., Ingham, R. J., Cooper, E. B., & Brutten, G. J. (1992,November). Criteria for the selection of stuttering intervention techniques and programs. Paper read at the Annual Convention of the American Speech-Language-Hearing Association, San Antonio, TX.

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

Ingham, R. J. (1995,December). The development azzd current status of time-interval measures of stuttering frequency: Part 1. Paper read at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL.

Ingham, R. J., & Cordes, A. K. (1992). Interclinic differences in stuttering event counts. Journal of Fluency Disorders, 17, 171-176.

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

Ingham, R. J., Cordes, A. K., Ingham, J. C., & Gow, M. L. (1995). Identifying the onset and offset of stuttering events. Journal of Speech and Hearing Research, 38, 315326.

Kelly, E. M., & Conture, E.G. (1988). Acoustic and perceptual correlates of adult stutterers' typical and imitated stutterings. Journal of Fluency Disorders, 13, 233-252.

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.

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 judgments of authentic and simulated stuttering. Journal of Speech and Hearing Disorders, 55, 383-391.

Perkins, W. H. (1983). Onset of stuttering: The case of the missing block. In D. Prins & R. J. Ingham (Eds.), Treatment of stuttering in early childhood: Methods and issues (pp. 1-20). San Diego, CA: College-Hill Press.

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.

Prosek, R. A, & Runyan, G. M. (1982). Temporal characteristics related to the discrimination of stutterers' and nonstutterers' speech samples. Journal of Speech and Hearing Research, 25, 29-33.

Prosek, R. A, & Runyan, G. M. (1983). Effects of segment and pause manipulations on the identification of treated stutterers. Journal of Speech and Hearing Research, 26, 510-516.

Smith, A. (1990). Towards a comprehensive theory of stuttering: A commentary. Journal of Speech and Hearing Disorders, 55, 398-401.

Smith, A., Luschei, E., Denny, M., Wood, J., Hirano, M., & Badylak, S. (1993). Spectral analyses of laryngeal and orofacial muscles in stutterers. Journal of Neurology, Neurosurgery, and Psychiatry, 56, 1303-1311.

Tuthill, C. E. (1946). A quantitative study of extensional meaning with special reference to stuttering. Speech Monographs, 13, 81-98.

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

Young, M. A. (1984). Identification of stuttering and stutterers. In R. F. Curlee & W. H. Perkins (Eds.), Nature and treatment of stuttering: New directions (pp. 13-30). San Diego, CA: College-Hill.

Received June 17, 1996

Accepted November 13, 1996

~~~~~~~~

By Roger J. Ingham University of California, Santa Barbara and Anne K. Cordes The University of Georgia Athens

Contact author: Roger J. Ingham, Department of Speech and Hearing Sciences, University of California, Santa Barbara, CA 93106-7050. Email: sphlingh@ucsbuxa.ucsb.edu

---
~~~~~~~~