RELIABILITY OF THE TIME INTERVAL DISFLUENCY TEST ON AN INTER-OBSERVER AGREEMENT BASIS

by

CAROL BENNETT BUETZER

B. S., Kansas State University, 1972

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF ARTS

Department of Speech

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1973

Approved by:

Bruce C Flanagan

Major Professor

ILLEGIBLE DOCUMENT

THE FOLLOWING DOCUMENT(S) IS OF POOR LEGIBILITY IN THE ORIGINAL

THIS IS THE BEST COPY AVAILABLE

ACKNOWLEDGEMENTS

I wish to express a special thanks to Dr. Bruce Flanagan, whose guidance and encouragement as major professor has been tremendously invaluable. A significant acknowledgement is also made to other members of the thesis committee: Professor Norma D. Bunton, Head of the Department of Speech; Professor Harry R. Rainbolt; and Professor Thomas M. Longhurst.

My colleague, Carol Coburn Hammond, deserves special thanks for her assistance in recording subjects and tabulating data in the present research.

The men who participated as subjects in the experiment deserve many thanks for their patience and consideration.

A special appreciation is extended to Dr. J. B. Lingwall, Professor of Speech Pathology and Audiology, Kansas University; and Dr. J. N. Neeley, Head of the Department of Speech Pathology and Audiology, Kansas University, for their assistance and use of Kansas University's Speech and Hearing Clinic facilities.

For his faithful support, assistance and understanding, I will forever be grateful to my loving husband, Eric.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Recently, a procedure to measure disfluencies was developed at Kansas State University which was titled the Time Interval Disfluency test (TID). Preliminary examination suggests that this test is considerably less time consuming, especially with spontaneous speech samples. This report is concerned with an examination of the interobserver reliability of the TID test.

Foremost among previous attempts to develop specific disfluency measurement instruments, was a method developed by Johnson et. al (1961) known as the Iowa Speech Disfluency Test. The speech behavior classified as disfluencies were: interjections of sounds, syllables, words or phrases; part-word repetitions; phrase repetitions; revisions; incomplete phrases; broken words; and prolonged sounds. In this study measurements were made of rate and disfluency in samples of oral reading and speaking. Subjects consisted of 100 male and 100 female adult speakers, of whom 50 in each group were classified as stutterers and 50 as nonstutterers. Each subject performed three speaking tasks: Job task; Responses to Thematic Apperception Test Cards; and Oral Reading. Disfluencies were identified in each case from a verbatim transcript while listening to a play-back of the recording. The recording was replayed as often as necessary to insure an accurate identification of disfluencies. The subjects' verbal output was defined as the number of words spoken. Speaking and reading rates were computed in terms of words per minute by calculating the ratio of verbal output to reading or speaking time as measured by a stop watch. A computation was made of the number of instances of each type of disfluency per 100 words for each of the three speaking tasks, giving a Disfluency Category Index. The formula used to compute this index was (ND/NW) 100=Disfluency Category Index.

> ND represents total no. of instances of disfluencies of the designated type in the speech sample of the subject.
>
> NW represents the number of words (verbal output) of the subject for the sample.

A disfluency index for each task was determined for each subject by obtaining the sum of the subjects' category indexes for the task. Johnson et. al., presents his data including: verbal output in the Job and tat tasks; measures of time spent on each task; ranges and deciles of distributions of oral reading and speaking rates for the three tasks; measures of disfluency; and coefficients of correlation among the three tasks for measures of each of the disfluency variables and of rate, in terms of tables. Johnson utilizes these findings as normative and comparative data for the Iowa Speech Disfluency Test, with respect to rate and disfluency in the speech and oral reading of adult male and female stutterers and nonstutterers. Young (1961) has also developed a method of measuring disfluencies. Disfluencies were classified as interjections, part-word repetitions, word-phrase repetitions, prolongations, and revisions. Subjects consisted of 50 college age male stutterers. Recordings were made of the Job Task to secure samples of speech. First a verbatim transcription was made and then the disfluencies were identified and classified. The tape was replayed as often as necessary to obtain sufficient accuracy. Young presents his data, including a summary of the five measures of disfluency, a frequency count for each type of disfluency, a measure of speaking time, and total frequency of all types of disfluency, in terms of mean frequency, range, and standard deviation. Sander (1961) conducted a study concerned with establishing the test-retest reliability for both reading and speaking tasks of the Iowa Speech Disfluency Test administered to a group of stutterers over an interval of 24 hours. He suggested a modification of the disfluency categories which were recommended by Johnson and Young, to consist of disfluencies defined as ones that involved a prolonged sound or a broken utterance, was interrupted by an interjection, or were involved in a sound, syllable or word repetition. Sander's Disfluent Word Index involves the eliciting of a recorded sample of reading and spontaneous speech. Subjects consisted of 40 stutterers, all but one participating in therapy at the University of Iowa Speech Clinic. Thirty-four of the stutterers were male, six were female. They ranged in age from 17 to 37 years with a mean age of 22.6 years. There was a 24-hour interval between the initial and subsequent administration of the reading and speaking tasks. The speaking situation was held constant for both days. The recording was then replayed as often as necessary to determine and classify the speaker's disfluencies. Rate of utterance for the 300-word passage and the 250 word speaking task was also determined. Sander presented the means, standard deviations, and ranges for total disfluencies, disfluent words, and time in a table. Results indicated that over the 24 hour period, half of the subjects showed no shifts in decile ranking, and over four-fifths of the subjects showed changes not greater than one decile. These results indicated rather high temporal reliability.

Each of the aforementioned studies; Johnson, Young, and Sander; utilize a method consisting of measuring rate of utterance and classifying disfluencies into five to eight classifications. That these methods are time-consuming and tedious is supported by the clinical experience of the current author, which tends to indicate that these measures take from 4 to 6 hours to complete an analysis. Johnson (1961) himself concedes: "Much time is needed, often several hours per subject, to obtain accurate transcriptions and identifications of the disfluencies in individual speech samples." (p. 19).

Aside from these methods of measuring disfluencies, perhaps the most prevalent means of determining the severity of stuttering and assessing progress in therapy is through use of rating scales. Although rating scales have been utilized for a considerable length of time, Lewis and Sherman (1951) were the first to devise a scale based on listener judgements of stuttering of recorded samples. They developed a scale for measuring severity of stuttering consisting of eight equal appearing intervals, ranging from 0.50-1.49 (Non-severe) to 7.50-8.49 (most severe). Shumak (1955) used two scales of estimating severity of stuttering. The stutterer's recorded personal evaluations of the severity of their stuttering on a 5 point scale (1-mild, 5-severe). The clinicians rated severity of stuttering on a 7 point scale (1-no stuttering, 7-extreme stuttering). Johnson, Darley, and Spriestersback (1963) devised a scale to rate the severity of stuttering consisting of an equal appearing interval scale from 1 (no stuttering) to 9 (very severe stuttering). This scale, known as the Iowa Rating Scale of Stuttering Severity, is probably the most widely used scale for rating severity even though Johnson stresses that it should be regarded as a rough measure. In 1963, Cullinan, Prather, and Williams conducted a study using ratings of severity on seven types of scales similar to those used by previous investigators (including five-, seven-, and nine-point scales). Cullinan and his associates had 27 samples of speech rated on seven types of scales with each being made by a different group of undergraduate students. The results of the study indicated that reliability coefficients for one judged ranged from 0.74 to 0.80 when only one rating was obtained. Indications were that reliability of the rating scale was not good enough to use in individual prediction.

Two more recent attempts to eliminate some of the subjectivity of rating scales have been research done by Riley (1972) and Wertheim (1972). Riley designed a scale for clinical and research use which yields a single numerical representation of severity with a range of 0-45. Frequency duration and physical concomitant are scored on a point system by the evaluator. Wertheim reports to be presently researching a stuttering measurement that would provide information on the situation, stability of the disorder, its qualitative pattern, its severity and the interrelationship between these parameters for a given individual. Both Riley and Wertheim call their new measures objective, however, both require some subjectivity on the part of one or more listener.

In view of the considerable variability and subjectivity of rating scales, and the length of time required to administer and score the aforementioned disfluency indexes, it would appear there is a need for a more efficient procedure that affords equivalent reliability. This author suggests that the Time Interval Disfluency Test appears to meet this criterion. With the procedure, a word was considered to be disfluent if it involved prolonged sounds, was classified as a broken word, was involved in sound, syllable or word repetition, or was interrupted by an interjection, which follows the disfluency definition advocated by Sander (1961). The test involved the eliciting and the recording of a sample of speech in three speaking situations: oral reading, conversation, and monologue. At 10-second intervals, a bleep (1000 Hz) was recorded simultaneously with the speech sample. Recording sheets consisted of four minute segments divided into 10-second intervals. The recording process consisted of the listener making one mark within a 10-second interval if a disfluency was heard during that time span. The listener indicated on the score sheet if the client was speaking, if he was disfluent, or if the examiner was speaking. The disfluency index was expressed in terms of the percent of time the subject was speaking and the percent of time he was disfluent. To better illustrate the TID test, a visual representation of a series of 10-second intervals and the scoring involved follows:

Disfluency		x	x		x				x		x			x	x			x			x	x		
Subject Speaking	x	x	x	x	x	x		x	x		x	x		x	x	x	x	x		x	x	x	x	x
Clinician Speaking	x						x						x											x
	1 minute						1 minute						1 minute						1 minute					

$\frac{\text{Subject Speaking}}{\text{Total blocks available}}$ = % time client was speaking

$\frac{\text{Disfluencies}}{\text{Subject Speaking}}$ = % disfluent

$\frac{\text{Examiner speaking}}{\text{Total blocks available}}$ = % time examiner was speaking

Therefore, the subject was speaking 83.3% of the time, the examiner was talking 16.6% of the time, and the subject was 41.6% disfluent.

This recording method was originally proposed by Bijou, Peterson and Ault (1968) for tabulating the frequency of occurrance of disruptive social behavior of pre-school children. This method of recording behavioral events in field situations consisted of registering the frequencies of occurrences and non-occurrences within a time interval. The observer makes a mark and only one mark, in each time interval in which the response occurred. In this procedure, it is obvious that the maximum frequency of a response is determined by the size of the time unit selected, i.e., if a 10-second interval was utilized, the maximum rate would be six responses per minute. If the frequency of occurrence of the desired behavior is high, smaller time intervals should be employed to obtain a high correspondence between the recorded and actual frequencies of occurrences. Johnston, Kelly, Wolf, and Harris (1966) have also reported a time series analysis procedure where the observation period is broken up into 10-second intervals. For each interval, the absence or presence of the behavior being studied is recorded. Since this procedure simply records whether or not a specific behavior occurred in an interval and ignores the number of times it occurred in the interval, it is important that the intervals are short enough so that variations in the behavior can be indicated. The frequency of the behavior should be used to determine the size of the interval. However, with manual recording, it is impractical to use intervals shorter than 10 seconds.

This procedure has been reported to be effective in recording spontaneous speech, tantrum behavior, and attending-to-work behavior. Schmidt and Ulrich (1969) conducted a study in which inappropriate behavior consisting of getting out of a desk seat, was recorded using a 10-second interval basis. The observer marked a box on a sheet with a 0 if the subject left her seat at any time during a 10-second interval over a 20 minute period. Schmidt and Ulrich were able to illustrate a change in base rate after reinforcement utilizing this procedure of recording behavior. In a study to ascertain the effects of tangible reinforcement of visual attending, Craig and Holland (1970), utilized a 10-second interval recording procedure. Two recorders observed seven deaf children individually for 10 seconds for standard sessions consisting of fifteen observations, recording if the child was or was not attending. Several studies conducted in the classroom have utilized a time sample recording technique consisting of recording if a particular behavior occurred during a 10-second interval: Cooper, Thomson, and Baer (1970); Broden, Bruce, Mitchell, Carter, and Hall (1970); and Bailey, Wolf, and Philipps (1970). Solomon and Wahler (1973) conducted a study with the intent to analyze experimentally peer reinforcement control of disruptive children's problem behaviors. They utilized ten second interval units to record six different target behaviors. The observers were instructed to record the first category occurring in a time interval. Thus for a ten second interval, only one of the possible six behavior categories could be scored. In a study of naturalistic treatment of an autistic child, Nordquiest and Wahler (1973) instructed observers to make coded checks for the occurrence of a behavior class within successive 10-second intervals over a 30 minute observation period. Any occurrence of a class, regardless of duration, was scored as a single unit. This method of recording behaviors clearly showed the change in the rate of inappropriate behaviors as a result of reinforcement.

These studies suggest that a time series analysis appears to be a successful way to measure several behaviors such as disruptive behavior, attending behavior, and autistic behaviors. All of these behaviors can be recorded in this manner during a specified period of time.

Assuming that stuttering is behavior which can be defined, observed, and recorded, the concern of the present study is to ascertain on an inter-observer basis if a time series analysis procedure is a reliable method for measuring stuttering behavior.

Reliability has generally been assessed in prior research involving stuttering behavior, in terms of test-retest, inter-observer agreement, and intra-observer agreement procedures. Tuthill (1946) attempted to see to what extent people agree when they label various speech phenomena as "stuttering". This study had to do with the extensional definition of stuttering by normal speakers, stutterers, clinicians, and experts in the field of speech pathology. The study consisted of three parts: Experiment A concerned the extensional definition of stuttering by three groups of subjects: stutterers, normal speakers unacquainted with the field of speech pathology and a group of clinicians. The subjects listened to a phonograph record, followed along on a transcript, and marked the places where "stuttering" occurred. The speech sample on each side of the record consisted of five samples of speech in the following order: Stutterer, normal speaker simulating stuttering, stutterer, normal speaker simulating stuttering, and normal speaker extemporizing; Experiment B consisted on the same type of definition using the following subjects: Speech Pathologists with Ph.D.'s, normal speakers unacquainted with speech pathology, mothers of preschoolers, stutterers, and extremely naive persons with a lower educational level. The same type of recording procedures were used as in Experiment A; and Experiment C was an attempt to see to what extent the variability and the agreement regarding stuttering would be altered if the subject had visual as well as auditory cues. Subjects consisted of a group of normal speakers who had participated in Experiment A and four groups of normal speakers unacquainted with the experiment or the issues involved. The subjects watched a sixteen-millimeter sound film of nine extemporaneous speakers (six stutterers, three normal speakers). Each subject held a signal key attached to a polygraph in his hand while viewing the film and pressed it when he thought "stuttering" had occurred. The sound film was shown with and without the visual image; the presentation varied with each group. Tuthill used intra-observer reliability in Experiment A, having each subject repeat the recording procedure a week after the first trial. He presented the reliability of each subject in terms of a coefficient of correlation, the ratio of the C to the T values (number of items checked both days divided by the total number of different items checked), and dividing the larger of two values into the smaller; Experiment B utilized inter-observer agreement (how well each observer agreed with all other observers) presented by a correlation conceived of as percentage of overlapping factors; and Experiment C measured inter-observer agreement utilizing percentage of overlapping factors. Results of the study indicated that normal speakers unacquainted with speech pathology indicated a significantly smaller number of items as instances of stuttering than do stutterers. The number of different items upon which each group agreed more than 75 percent were approximately the same. Twenty percent of all stutterings were agreed upon by 80 percent or more of all judges. Forty percent of the stutterings received 50 percent agreement or better. Average coefficients of correlation for the clinical, normal, and stuttering groups of judges were .72, .72, and .65 respectively. The agreement among a number of recognized experts in speech pathology was no greater than that among a number of persons unacquainted with the field. The addition of visual to auditory cues did not tend to increase the agreement materially, when practice and adaptation effects were taken into account.

Johnson (1961) in his research concerning the Iowa Speech Disfluency Test, established inter- and intra-examiner reliability through three separate investigations. Duffy obtained data from two observers who listened to 12 recorded speech samples of female stutterers and recorded disfluencies using the eight categories. He correlated the results of the two observers in identifying disfluencies in each category and obtained coefficients of correlation ranging from .90 to .99. Young, in order to establish intra-examiner reliability, analyzed 10 tape recorded speech samples of adult male stutterers twice, with intervening periods of from 3 weeks to 2 months with reference to 5 types of disfluent categories. Indexes ranged from .91 to 1.00, for the 10 samples, with an index of agreement of .97 for all 10 samples combined. Sander established intra-examiner reliability by making two analyses of the recording of 12 job tasks and 12 reading tasks performed by male stutterers. The analyses were separated by at least a one month interval. Percentage of agreement between the first and second scoring was 96% for both Job and reading tasks. Young (1961) in order to estimate intra-examiner reliability in making a fluency analysis, selected 10 tapes at random and a second fluency analysis of each was performed with a minimum of 2 weeks and a maximum of 3 months separating the analyses. Young used this same reliability procedure for both the Johnson

(1961) study and his own study in 1961. As previously mentioned, Sander (1961) demonstrated the reliability of the Iowa Speech Disfluency Test utilizing a test-retest procedure. A twenty-four hour interval separated the initial and subsequent administration of the reading and speaking tasks. Sander established self-scoring reliability using coefficient of agreement. Sander also utilized this same reliability procedure in both the Johnson (1961) study and his own study.

It is obvious that both test-retest reliability procedures and inter-agreement and intra-agreement reliability procedures are both effective and widely used as evidenced by the studies cited above. It has been the experience of the author and others working in the Kansas State University speech and hearing clinic that there is extensive use of behavioral observation procedures in the clinic setting. Many of the clinic personnel are knowledgeable about literature of time series analysis recording procedures. Through extensive research they are able to obtain inter-observer and intra-observer agreement of above 80%. The concern of the present study was to measure inter-observer agreement of observers who are exposed to brief, but explicit training procedures. The author is not only interested in accrueing data for formal assessment of the reliability of the TID test, but also to assess inter-observer reliability with a simple, brief training procedure. Hence, the present study is an attempt to ascertain the reliability of the Time Interval Disfluency Test on an inter-observer agreement basis.

METHOD

The procedures followed in this research may be presented in reference to subjects, recording procedures, disfluency definition and analysis, observers, observer training procedure and scorer reliability.

Subjects

Nine males, all of whom were considered by themselves and by their speech clinicians to be stutterers, served as subjects in this study. The disfluent subjects, whose ages ranged from 18 to 40 years, were drawn from Kansas State University, the University of Kansas, and Fort Riley Army Base, Fort Riley, Kansas. Eight of the nine subjects were participating in speech therapy, and four of the subjects were university students when the tape recordings of their speech were made.

Recording Procedure

Each subject was seated beside the examiner at a table in a quiet room with a microphone positioned in front of him. The examiner asked identifying information such as name, age, address, and phone number. This brief interview served to familiarize the subject with the experimental situation, and to adjust the recording level.

For the oral reading situation, the subject was handed fifty 5" by 7" index cards and was asked to read them as he normally would. Each card contained a paragraph excerpt from an encyclopedia, with different topics on each. The subject was instructed to read until asked to stop. The examiner utilized a stop watch for all speaking situations to insure obtainment of a complete four minute sample.

A monologue sample of speech was also obtained from each subject. The subject was handed fifty 5" by 7" index cards, consisting of a single colored photograph on each card. The pictures were gathered from various magazines and were chosen to elicit spontaneous speech. The pictures were presented in a random order to each subject. The subject was instructed to look at one picture at a time and talk about any facet of the picture he desired. He was encouraged to speak at his own rate and say as little or as much as he wished about each picture. The examiner stopped the subject after four minutes had passed.

The third speaking situation consisted of a conversation between the subject and the examiner. Questions and comments were posed by the examiner to elicit conversational speech. These questions (see Appendix A) inquired about such information as what the subject did in his leisure time, what hobbies or interests he had, and what places he had traveled to or would like to visit. The examiner not only asked questions but also responded appropriately to the subject's comments, as would occur in a typical conversation. Order of presentation of the three speaking situations was randomized for each subject. The examiner concluded the conversation at the end of four minutes.

A two-channel Ampex tape recorder, model no. AG 5002 and monitoring equipment, housed in an adjacent room, were used to record the speech samples. Ten second bleeps (1 K Hz tones) were recorded simultaneously on the speech sample tapes to produce standard tapes for analysis. A portable Wollensak cassette tape recorder, playing the ten second bleeps from a prerecorded cassette, was connected to channel B of the Ampex tape recorder. A four minute sample of each speaking situation was obtained from each subject.

Disfluency Definition

The definition of a disfluency was adopted from Sander (1961). "A word was considered to be disfluent if it involved prolonged sounds, was classified as a broken word, was involved in sound, syllable or word repetition, or was interrupted by an interjection. Words preceeded by interjections or involved in phrase repetition were not counted as disfluent words."

Disfluency Analysis

Utilizing the Time Interval Disfluency Test measure, in which a mark is made in a 10-second block if a disfluency occurred during that time, disfluencies were recorded for each speaking situation for each subject. Recording sheets (see Appendix B) consisted of four, 1 minute segments divided into 10-second intervals. The recording process for each of the three speaking situations; monologue, conversation, and oral reading; consisted of making one and only one mark within a 10-second interval if a disfluency was heard within that time. The listener indicated in the appropriate space on the score sheet if the subject was speaking, if he was disfluent, and/or if the examiner was speaking. The final result was then expressed in terms of percent of time the subject was speaking, and the percent of time he was disfluent.

Observers

The eleven observers selected for this study were drawn from graduate students enrolled in the course entitled Stuttering 283-834, at Kansas State University, during spring semester, 1973. Two of the eleven judges had previously been exposed to the Time Interval Disfluency Test.

Observer Training Procedure

Three separate listening sessions were held, in which the eleven observers analyzed the conversation sample for each subject during one listening session, the monologue sample at the second session, and the oral reading during the final listening session. The order of presentation of the speech samples of each subject was randomized for each listening session. It was felt that the total length of samples for all subjects, 108 minutes, was too lengthy a task to be completed at one session, without risk of fatigue being a factor in the analysis.

In order to facilitate accurate analysis of the samples, training sessions were utilized immediately prior to each analysis session. Each observer listened to two, 2-minute samples from three different subjects ranging in severity from mild to severe, and scored disfluencies according to the method utilized by the TID test. Immediate feedback was provided by having the observers compute their percentage of agreement with a master score sheet. (This master score sheet was developed by the examiner and two assistants. A disfluency was considered to have occured during a ten-second period when two of the three agreed that it had). Each of the eleven observers were required to reach 80% agreement with the master score sheet. The training session continued until this criterion was met. This procedure was implemented immediately before all of the listening sessions. Since a period of seven days separated each of the three sessions, the training sessions before the second and third listening sessions were regarded as refresher training sessions.

During the listening sessions, the listeners attended to the tape recorded speech samples of the nine subjects, taking short rest periods while the examiner changed the tapes, and recorded any disfluency occurrences for each 10-second interval on the TID test recording sheets. Percent of time the subject was speaking and the percent of disfluency was computed for each subject for each speaking task.

Scorer Reliability of the Disfluency Analysis

Scorer reliability was indicated by two methods: percentage of agreement of each observer to all other observers in terms of agreement of whether a 10-second interval was marked or unmarked, as a function of disfluency occurrence; and coefficient of correlation of the total disfluency percent computed for each of the speaking tasks from each subject by each of the eleven observers.

RESULTS

The results of this research are presented in reference to reliability of the TID test on an inter-observer basis. This section will contain information relevant to the percent of observer agreement for each speaking task: monologue, conversation, and oral reading; and coefficients of correlation for total percent of disfluency computed by each observer for the subjects.

Recall that scoring utilizing the TID test involves listening to monologue, conversation and oral reading samples, and recording one mark on the score sheet within a ten second interval block, if a disfluency was heard during that time span. The listener indicates in the proper place on the score sheet if the subject was speaking and if he was disfluent during a ten-second interval. The disfluency index was expressed as a percentage by dividing the number of ten-second intervals in which the subject was disfluent, by the total number of ten-second intervals in which speech was observed.

Inter-observer coefficients of agreement were computed by dividing the number of instances of agreement on two observer score sheets by the total number of agreements plus disagreements (i.e., the total number of recording opportunities). The quotient was multiplied by 100 and the resulting figure was the percentage of agreement between two observers. This procedure was utilized to compare each observer with every other observer. Hence, 55 coefficients of agreement were obtained per subject. The 25th, 50th, and 75th percentile points of the coefficients of agreement were computed for each subject, together with ranges of coefficients of agreement for individual subjects. The 25th percentile was the value located so that one-fourth of the coefficients of agreement fell below it and three-fourths of the agreements exceeded it. The 50th percentile was the median point, where one-half of the agreements fell below it and one-half exceeded it. The 75th percentile was the value so located that three-fourths of the agreements fell below it and one-fourth exceeded it. Tables I through III present percentages of coefficients of agreements among observers for subjects' disfluent intervals for the three speaking tasks: oral reading, conversation, and monologue. Table I indicates the percent of agreement among observers for the oral reading task. It shows that at the 25th percentile, coefficients of agreement ranged from 63-88 percent among subjects, at the 50th percentile, the range was 75-92 percent, and at the 75th percentile, coefficients of agreement ranged from 79-96 percent. The range for percentages of coefficients of agreement for individual subjects was fairly wide, ranging from 42-92 percent for subject 1, to a range of 79-100 percent coefficients of agreement for subject 2. Table II concerns percentage of agreements among observers for subjects' disfluent intervals for the conversation speaking task. The range of coefficients of agreement for the 25th percentile among subjects was 63-88 percent; for the 50th percentile, coefficients of agreement ranged from 75-92 percent; and for the 75th percentile, a range of 79-96 was computed. Coefficients of agreement for individual subjects for the conversation task ranged from 50-96 percent for subject 9 to 79-100 percent agreements for subject 8. Table III depicts the coefficients of agreements among observers for subjects' disfluent intervals for the monologue task. Coefficients of agreement across subjects ranged from 67-83 percent at the 25th percentile, 75-88 percent at the 50th

TABLE I. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Disfluent Intervals for the Oral Reading Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	63	75	79	42- 92
2	83	92	96	72-100
3	88	92	92	75-100
4	79	88	92	67-100
5	75	83	88	58- 96
6	83	88	92	75-100
7	79	83	88	71-100
8	83	88	92	71-100
9	67	75	79	63- 96

TABLE II. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Disfluent Intervals for the Conversation Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	75	79	83	67- 92
2	79	83	88	71-100
3	79	83	88	71-100
4	75	83	88	67- 96
5	71	79	83	38- 96
6	71	79	83	54-100
7	63	75	79	50- 92
8	88	92	96	79-100
9	75	79	83	50- 96

TABLE III. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Disfluent Intervals for the Monologue Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	67	75	79	54-100
2	79	83	88	63- 96
3	79	88	92	67-100
4	83	88	92	67-100
5	71	79	83	58- 92
6	67	75	79	50- 92
7	67	75	83	58- 96
8	75	83	88	50- 96
9	75	79	83	67- 96

percentile, and a range of 79-92 percent coefficients of agreement at the 75th percentile. Subjects 1 and 7 indicated the widest range for individual ranges of coefficients of agreement: 54-100 percent and 50-96 percent respectively.

The median coefficients of agreement among subjects by task are indicated in Table IV. The median coefficient of agreement for the reading task at the 25th percentile, was 79 percent, 88 percent at the 50th percentile, and 92 percent at the 75th percentile. For the conversation speaking task, the median coefficient of agreement at the 25th percentile was 75 percent; at the 50th percentile, median coefficient of agreement was 79 percent, and at the 75th percentile, the median coefficient of agreement was 83 percent. A median coefficient of agreement of 75 percent at the 25th percentile, 79 percent at the 50th percentile, and 83 at the 75th percentile was obtained for the monologue speaking task.

Tables V through VII present the summaries of percent of agreements among observers for subjects' speaking intervals for the oral reading, conversation, and monologue speaking tasks. Table V depicts the coefficients of agreement among observers for subjects' speaking intervals for the oral reading task. The range and median coefficients of agreement for the 25th, 50th, and 75th percentiles was 100 percent. Table VI presents the summary of percentages of agreements among observers for subjects' speaking intervals for the conversation task. The range of coefficients of agreement among subjects for the 25th percentile was 92-100 percent; for the 50th percentile coefficients of agreement ranged from 96-100 percent with a median agreement of 100 percent, and the range for the 75th percentile was 100 percent agreement.

TABLE IV. The Median Coefficients of Agreement Among Subjects by Task.

Task	Median Coefficients of Agreement		
	25th Percentile	50th Percentile	75th Percentile
Reading	79	88	92
Conversation	75	79	83
Monologue	75	79	83

TABLE V. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Speaking Intervals for the Oral Reading Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100
5	100	100	100	100
6	100	100	100	100
7	100	100	100	100
8	100	100	100	100
9	100	100	100	100

TABLE VI. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Speaking Intervals for the Conversation Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	96	100	100	92-100
2	96	96	100	92-100
3	92	96	100	83-100
4	100	100	100	100
5	96	96	100	88-100
6	100	100	100	100
7	92	96	100	83-100
8	100	100	100	96-100
9	100	100	100	100

The largest range of agreements for individual subjects was 83-100 percent for both subjects 3 and 7. Percentage of agreements among observers for subjects' speaking intervals for the monologue speaking task are presented in Table VII. For the 25th percentile, coefficients of agreement among all subjects ranged from 96-100 percent; for the 50th percentile, agreements ranged from 92-100 percent among subjects with a median coefficient of agreement of 100 percent; and agreements at the 75th percentile ranged from 96-100 percent. Only subject 1, with an individual range of 88-100 percent, had a range other than 100 percent agreement. This data shows that the median coefficient of agreement among observers for subjects' speaking intervals at the 50th percentile was 100 percent for the oral reading, conversation, and monologue speaking tasks.

The second measure utilized to compute reliability on an inter-observer agreement basis was the Pearson Product Moment Correlation procedure. This involved obtaining coefficients of correlation for the total disfluency percent computed for each speaking task from each subject by each of the eleven observers. Table VIII indicates that reliability coefficients for inter-observer agreement of percent of disfluency on the oral reading task ranged from .8820 to .9935 with a coefficient of .9479 at the 25th percentile, .9669 at the 50th percentile, and .9783 at the 75th percentile. For the conversation speaking task, a coefficient of .7886 was computed at the 25th percentile, .8352 at the 50th percentile, and .9007 at the 75th percentile. Coefficients of correlation for this task ranged from .5986 - .9712. Coefficients of .7986 at the 25th percentile, .8531 at the 50th percentile, and .8940

TABLE VII. The Summary of Percentage of Agreements Among Observers Utilizing the TID Test for Subjects' Speaking Intervals for the Monologue Task.

Subjects	Coefficients of Agreement			
	25th Percentile	50th Percentile	75th Percentile	Range
1	92	96	96	88-100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100
5	100	100	100	100
6	100	100	100	100
7	100	100	100	100
8	100	100	100	100
9	100	100	100	100

TABLE VIII. The Summary of Coefficients of Correlation of the Total Disfluency Percent Computed for Each Speaking Task from Each Subject by Each of the Eleven Observers.

Speaking Tasks	Coefficients of Correlation			
	25th Percentile	50th Percentile	75th Percentile	Range
Oral Reading	.9479	.9669	.9783	.8820-.9935
Conversation	.7886	.8352	.9007	.5986-.9712
Monologue	.7986	.8531	.8940	.6055-.9596

at the 75th percentile were obtained for the monologue task which had a range of .6055 - .9596. These correlations indicate that there was high agreement among observers as to the total percent of disfluency of the subjects.

DISCUSSION

The research finding of major importance was a demonstrable reliability of the TID test on an inter-observer agreement basis. It was hypothesized that a time series analysis procedure such as that utilized by the Time Interval Disfluency Test could be shown to be reliable by means of inter-observer agreement. Reliability was defined in terms of the degree to which independent observers agree on what they have observed in the same subject during the same observation session. Bijou, et. al., (1968) suggest that, although no absolute standards have been established, coefficients of agreement of 80% or more for observers recording the occurrence or nonoccurrence of specified behaviors in a time interval, are indicative of significant reliability.

Two measures were utilized to present reliability: coefficients of agreement among observers for each speaking task: oral reading, conversation, and monologue; and coefficients of correlation for total percent of disfluency computed by each observer for the subjects.

The results indicate that the median coefficient of agreement among observers for the 50th percentile among all subjects for the oral reading task was 88 percent, and median coefficient of agreement among all subjects for both conversation and monologue tasks was 79 percent. These coefficients are higher than the findings of Tuthill (1946) who indicated that a coefficient of correlation of .72 for intra-observer agreement was sufficiently reliable. Although the median coefficient of agreement among all observers for the conversation and monologue task was 79 percent (which is 1 percent below the level of agreement supported by Bijou et. al.) this is not considered to be a significant difference. Hence the median coefficient of agreement among observers, given Bijou's criteria, is reliable for each speaking task. The results also clearly indicate that observers demonstrated high reliability in indicating agreement for each speaking task of whether or not a subject was speaking.

The median coefficient of correlation for total percent of disfluency computed by each observer for the subjects were .9669 for the oral reading task, .8352 for the conversation task, and .8531 for the monologue speaking task. These coefficients are significant at the .01 level of confidence, thus indicating significant reliability of the TID Test on an inter-observer agreement basis.

The second measure of presenting reliability, coefficients of correlation for total percent of disfluency computed by each observer for each subject, reveals a higher level of agreement than does the first measure, coefficients of agreement among observers for each speaking task. However, the correlational procedure does not provide information as to the extent to which observers agree on specific intervals. Even if two observers computed identical total percent of disfluencies, they could have recorded disfluent behavior during different 10-second intervals. Coefficients of agreement indicate directly the degree to which judges agree for each ten-second interval.

A significant aspect of the present study was the indication that reliability was higher on the oral reading task than on the conversation or the monologue speaking tasks. Perhaps the observer was more able to anticipate what the subject would say next, from contextual cues on each reading card. This might have enabled the observer to concentrate more on listening for a disfluency than on listening to the context of the utterances.

It appears that the brief, but explicit training session was sufficient to produce reliability of the Time Interval Disfluency Test. However, it is possible that a more extensive training procedure could produce even higher coefficients of agreement. Changes in the training sessions might include longer practice samples (4 minutes instead of 2) from more than three subjects, or raising the criterion of agreement to a master score sheet to 100% for each observer before actual recording sessions were initiated.

In reference to the procedures utilized in this research, some disagreements among observers may have been a function of one or more observers failing to hear the bleep tone, which signals moving on to the next 10-second interval, or an observer becoming confused as to which 10-second interval he should be recording in. This would cause an observer to mark intervals from that point on incorrectly. A possible solution to alleviate this problem would be to use progressively higher pitched tones over a one minute period; thus the bleep indicating the first 10-second interval would be a 500 Hz tone, the second bleep a 1000 Hz tone, and so on. Another suggestion would involve utilizing video taped speech samples, with a clock in the background. The bleep could be timed to coincide with ten-second intervals on the clock, thus giving observers a visual and audio stimulus to signal that the observer move on to the next 10-second interval. This would eliminate confusion as to which 10-second interval the observer should be recording in.

Implications for Application and Further Research

The results of this study indicate that the Time Interval Disfluency Test is reliable on an inter-observer basis. This procedure involves listening to four minute samples each of monologue, conversation and oral reading speech samples, and recording one mark on a score sheet within a ten second interval block, if a disfluency was heard during that time. An observer requires approximately fifteen minutes to complete this procedure. This includes computing percent of time that the subject is speaking, and the percent of disfluency. Therefore it would appear that this is a most efficient, time-saving procedure. Recall that Johnson (1961) stated that the Iowa Speech Disfluency Test takes several hours to obtain transcriptions and identifications of the disfluencies in an individual speech sample. Experience of this author indicates that the disfluency measures suggested by Young (1961) and Sander (1961) take from two to three hours to complete for individual subjects.

Rating scales appear to be a fairly fast method of determining the severity of stuttering. However, Allyon and Azrin (1968) have suggested that distortion and error can result from interpretive characterization of performance, utilizing rating scales. Establishment of intra- and inter-examiner reliability is more difficult when rating scales are used as the measuring instrument. Multiple observers are necessary to obtain a reliability ratio, and the observer is susceptible to applying his own personal biases to the rating of the behavior. Allyon and Azrin suggest describing the behavior in specific terms that require a minimum of interpretation. In this way, the observers would avoid reading their interpretation into the data by forcing them to attend to and record only in terms of the physical aspect of the response.

The Time Interval Disfluency Test appears to meet this criterion. Disfluencies are described in specific terms so that the observer attends to only the physical aspect of the response. In this way, much, if not all of the observer's subjectivity is eliminated.

From the standpoint of therapy, the findings in this study tend to indicate that the TID test would be a fast, efficient method of initially evaluating a stutterer, and that it is also applicable to assessing progress after a period of therapy. Comparisons could be made for each speaking task of the initial and subsequent testing.

The TID test is a more efficient and less time-consuming procedure of measuring disfluencies than measures suggested by Johnson (1961), Young (1961) and Sander (1961), and it is as reliable as these procedures. It would seem to be a useful clinical tool for clinicians interested in a quick, efficient method of assessing a client's percent of disfluent intervals.

Future research could provide a means to increase reliability of the TID test procedure by utilizing a longer training session, or by raising the criterion of observer agreement to above 80% during the training session. Subsequent research could also be directed to modifying the procedures utilized in this research to produce higher coefficients of agreement.

REFERENCES

1. Allyon, T. and Azrin, N., The Token Economy, New York: Appleton-Century Crofts, 1968, p. 37-39.

2. Bailey, J. S., Wolf, M. M., and Philips, E., Home-based Reinforcement and the Modification of Pre-delinquent Classroom Behavior, Journal of Applied Behavior Analysis, 1970, 3, 223-233.

3. Bijou, S. W., Peterson, R. F., and Ault, M. H., A Method to Integrate Descriptive and Experimental Field Studies at the Level of Data and Empirical Concepts, Journal of Applied Behavior Analysis, 1968, 1, 175-191.

4. Broden, M., Bruce, C., Mitchell, M., Carter, V., and Hall, R., Effects of Teacher Attention on Attending Behavior of Two Boys at Adjacent Desks, Journal of Applied Behavior Analysis, 1970, 3, 199-203.

5. Cooper, M. L., Thomson, C. L., and Baer, D. M., The Experimental Modification of Teacher Attending Behavior, Journal of Applied Behavior Analysis, 1970, 3, 153-157.

6. Craig, H. B. and Holland, A. L., Reinforcement of Visual Attending in Classrooms for Deaf Children, Journal of Applied Behavior Analysis, 1970, 3, 97-109.

7. Cullinan, W. L., Prather, E. M., and Williams, N. E., Comparison of Procedures for Scaling Severity of Stuttering, Journal of Speech and Hearing Research, 1963, 6, 187-194.

8. Johnson, W., Measurements of Oral Reading and Speaking Rate and Disfluency of College-Age Male and Female Stutterers and Non-stutterers, Journal of Speech and Hearing Disorders, Monograph Supplement 7, 1961, 21-30.

9. Johnson, W., Darley, F. L., and Spriestersbach, D. C., Diagnostic Methods in Speech Pathology, Harper and Row, New York, 1963.

10. Johnston, M. S., Kelley, C., Harris, F. R., and Wolf, M. M., An Application of Reinforcement Principles to Development of Motor Skills of a Young Child, Child Development, 1966, Volume 37, 379-387.

11. Lewis, P. and Sherman, D., Measuring Severity of Stuttering, Journal of Speech and Hearing Disorders, XVI (1951), 320-326.

12. Nordquist, V. M. and Wahler, R. G., Naturalistic Treatment of an Autistic Child, Journal of Applied Behavior Analysis, 1973, 6, 79-87.

13. Riley, G. D., A Stuttering Severity Instrument for Children and Adults, Journal of Speech and Hearing Research, 37, 1972, 314-321.

14. Sander, E. K., Reliability of the Iowa Speech Disfluency Test, Journal of Speech and Hearing Disorders, Monograph Supplement 7, 1961, 21-30.

15. Schmidt, G. W. and Ulrich, R. E., Effects of Group Contingent Events Upon Classroom Noise, Journal of Applied Behavior Analysis, 1969, 2, 171-179.

16. Shumak, C., A Speech Situation Rating Sheet for Stutterers, In W. Johnson (ed.) Stuttering in Children and Adults, Minneapolis: University of Minneapolis Press, 1955.

17. Solomon, R. W. and Wahler, R. G., Peer Reinforcement Control of Classroom Problem Behavior, Journal of Applied Behavior Analysis, 6, 49-56.

18. Tuthill, C. E., A Quantitative Study of Extensional Meaning with Special Reference to Stuttering, Speech Monograph 13, p. 81-98, 1946.

19. Wertheim, E. S., A New Approach to Classification and Measurement of Stuttering, Journal of Speech and Hearing Research, 37, 1972, 242-251.

20. Young, M., Predicting Severity of Stuttering, Journal of Speech and Hearing Disorders, Monograph Supplement 7, 1961, 31-54.

APPENDICES

APPENDIX A

CONVERSATION

What types of leisure activities do you engage in?

Have you ever participated in any organizations?

How active are you in political affairs?

How would you describe your personality characteristics?

- - - - - -

Where are you living on campus? Do you like it?

How relevant do you feel your education was for you?

Do you feel that coming to KSU (KU) has affected your attitudes in any way?

What would you have liked to change about your schooling so far?

What field of study are you in?

After you obtain your degree are you planning on graduate study?

- - - - - -

Do you have any brothers and sisters?

Could you tell me about them?

How would you describe your relationship with your parents?

- - - - - -

What places have you traveled to?

Where would you like to go if you had the opportunity? Why?

- - - - - -

What types of things really irritate you about people or situations?

What particular traits do you admire in people?

If there was one thing you could change about yourself what would it be?

What do you feel your best qualities are?

What person have you most admired? Why?

APPENDIX B

RECORDING SHEET FOR THE TIME INTERVAL DISFLUENCY TEST

Disfluency						
Subject Speaking						
Examiner Speaking						

RELIABILITY OF THE TIME INTERVAL DISFLUENCY TEST
ON AN INTER-OBSERVER AGREEMENT BASIS

by

CAROL BENNETT BUETZER

B. S., Kansas State University, 1972

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF ARTS

Department of Speech

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1973

The purpose of this research is to ascertain the reliability of the Time Interval Disfluency Test on an inter-observer agreement basis.

Tape recorded speech samples were obtained from nine adult male stutterers ranging in stuttering severity from mild to severe. The Time Interval Disfluency Test involved the eliciting and the recording of a four minute sample of speech in each of three speaking situations: oral reading, conversation, and monologue, with a bleep recorded at 10-second intervals. Eleven observers analyzed the speech samples. An observer marked within a 10-second interval on a recording sheet of four minute segments divided into 10-second intervals, if a disfluency was heard during that time period. The observer indicated on the score sheet if the subject was speaking and if he was disfluent during a 10-second interval. The disfluency index was expressed as a percentage by dividing the number of ten-second intervals in which the subject was disfluent, by the total number of ten-second intervals in which speech was observed.

Observer reliability was presented by two methods: Coefficients of agreement of each observer to all other observers in terms of agreement whether a 10-second interval was marked or unmarked, as a function of disfluency occurrence; and coefficients of correlation of the total disfluency percent computed for each of the speaking tasks from each subject by each of the eleven observers.

Results of this study indicated that median coefficients of agreement for each speaking task: 88 percent for the oral reading task, and 79 percent agreement for both the conversation and monologue speaking tasks, were reliable, given Bijou's criteria of reliability. Median coefficients of correlation for each speaking task were significant at the .01 level of confidence. Thus both measures utilized indicated sufficient reliability of the Time Interval Disfluency Test on an inter-observer agreement basis.

The TID test is a more efficient and less time-consuming procedure of measuring disfluencies than measures suggested by Johnson (1961), Young (1961) and Sander (1961), and it is as reliable as these procedures. It would seem to be a useful clinical tool for clinicians interested in a quick, efficient method of assessing a client's percent of disfluency.

Future research could provide a means to increase reliability of the TID test procedure by utilizing a longer training session, or by raising the criterion of observer agreement to above 80% during the training session.