12-1997

A Comparison of Listener and Speaker Perception of Stuttering Events

Anne Jocelyn Schagen
Portland State University

THESIS APPROVAL

The abstract and thesis of Anne Jocelyn Schagen for the Master of Arts in Speech Communication: Speech and Hearing Science were presented November 7, 1997, and accepted by the thesis committee and the department.

COMMITTEE APPROVALS:

[redacted]

John A. Tetnowski, Chair

[redacted]

Mary E. Gordon-Brannan

[redacted]

G. Tucker Childs
Representative of the Office of Graduate Studies

DEPARTMENT APPROVAL:

[redacted]

Stephen A. Kosokoff, Chair
Department of Speech Communication

* *

ACCEPTED FOR PORTLAND STATE UNIVERSITY BY THE LIBRARY

by [redacted] on Dec. 23, 1997

ABSTRACT

An abstract of the thesis of Anne Jocelyn Schagen for the Master of Arts in Speech Communication: Speech and Hearing Science presented November 7, 1997.

Title: A Comparison of Listener and Speaker Perception of Stuttering Events.

Stuttering identification, measurement, research, and treatment have for many years had their basis in listener judgment of stuttering, but the covert aspects of stuttering are not behaviorally observable, and inter-rater reliability has repeatedly been shown to be low. Perkins (1990) has emphasized the importance of consulting the speaker for the most reliable perspective on stuttering identification. The question raised in this study is whether there is a significant correlation between stuttering identification based upon internal perception by a speaker who stutters, and identification based upon external perception of listeners, with points of inter-rater disagreement removed.

Six adult males, aged 18 to 47, who stutter with at least moderate severity read 25 sentences aloud and marked perceived points of stuttering as they occurred. Eight speech-language pathology graduate students listened to the same audiotaped samples and marked points where they perceived stuttering. Points where at least 7 listeners agreed that stuttering was or was not present were compared to speaker perceptions. Findings were analyzed using Cohen's kappa (Cohen, 1960), a correlation measure which controls for chance agreement.

Results showed a kappa correlation of .276 which was significant at the p = .001 level. While this correlation is highly significant, it is representative of very poor

agreement, kappas of greater than .70 being acceptably high to show strong agreement. Few instances of stuttering actually occurred for 5 out of 6 speakers, so agreement was based largely on fluent speech. When analyzed separately by speaker, kappas ranged from .176 to .887, but could not be calculated for 3 speakers as there were no instances where stuttering was perceived by 7 out of 8 listeners. Out of a possible 2,040 points of agreement, 70 were not analyzed due to listener disagreement.

These results suggest that, while speakers and listeners generally agree in their perceptions of fluent speech, agreement between speakers and listeners regarding stuttered speech is low. If we take speaker perception to be the standard of reliability, this study suggests that our accepted methods of stuttering identification and measurement in research and treatment assessment, baseline measurement, tracking, and measurement of progress are highly suspect.

A COMPARISON OF LISTENER AND SPEAKER PERCEPTION OF STUTTERING EVENTS

by

ANNE JOCELYN SCHAGEN

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF ARTS
in
SPEECH COMMUNICATION:
SPEECH AND HEARING SCIENCE

Portland State University
1997

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION AND STATEMENT OF PURPOSE

Introduction

Stuttering is a disorder that has defied precise definition and measurement. Young (1975) cited a familiar cliché, "everybody knows what stuttering is except the experts" (p. 530). While the matter of defining stuttering may seem academic and immaterial, it is in fact a matter with significant clinical implications. Accurate identification of incipient stuttering is a challenging and important clinical question which is as yet unanswered. Any information that helps identify stuttering as distinct from normal disfluency is helpful in this quest, which hopefully will one day lead us to treat emerging stuttering as early as possible.

In clinical settings, assessment is usually accomplished by a frequency count, a severity rating, or both, by a listener (Kully & Boberg, 1988). Listener evaluation is also the foundation for baseline measurement and tracking of progress in treatment, in spite of the fact that inter-rater agreement rates have been shown to be consistently low (Curlee, 1981; Perkins, 1990; Young, 1975), and to vary depending upon the means of subdividing the text, and whether a total count, a point-by-point analysis, or a time interval analysis is performed. Furthermore, it is generally accepted that there are both overt and covert aspects to stuttering, but the perceptions of listeners cannot be expected to identify the covert elements. Behavioral assessment and treatment are useful in managing the overt aspects, but the covert aspects go unaddressed.

Previous research into identification of stuttering has focused on identification based upon listener perception. Comparisons of total counts of stuttering during a given speech sample have been reasonably reliable (Kully & Boberg, 1988), but research by Young (1975) and Curlee (1981) into point-by-point

assessment by listeners has revealed consistent and significant lack of agreement. Cordes and Ingham (1995) have achieved higher agreement with time interval assessment, but it has been at the expense of specificity. If point-by-point observer agreement of moments of stuttering is so inconsistent (Curlee, 1981; Kully & Boberg, 1988; Moore & Perkins, 1990; Young, 1975), how do we know when stuttering has truly occurred? Little research has been devoted to whether or not the people who are being studied so carefully agree with these external evaluations of their speech (Borden, 1990, Perkins, 1990). Only a few studies (Ingham & Cordes, 1997; Kelly & Conture, 1988; Moore & Perkins, 1990) have looked at a comparison between the evaluation by listeners and the self-evaluation by the person who stutters.

Some researchers believe that stuttering is more than a behavioral disorder (Perkins, 1990), and that subjects' self-evaluations of their stuttering is important. It is possible that the only truly reliable judgments come from the persons who stutter themselves (Perkins, 1990). If a self-measure is taken as the accurate representation of stuttering, to what degree do observers' measurements correlate?

This study looked at agreement between listeners and subjects who stutter. If theory, research, and treatment are all based on listener judgment, and listener judgment does not reflect what individuals who stutter perceive in themselves, then the basis for theory, research, and treatment is questionable. This research sought to examine the correlation between point-by-point judgments of stuttering made by the person who stutters at the time of speaking versus the judgments of stuttering made by independent listeners when listening to the audiotaped recording of a person who stutters.

Statement of Purpose

The purpose of this study was to determine the degree of agreement between the self-perception of stuttering by speakers and the identification of stuttering by listeners. This was accomplished by comparing the subjective point-by-point ratings by speech-language pathology graduate students to the marked point-by-point moments of stuttering as identified by the speaking subjects themselves.

The research questions were: (a) what is the correlation between a speaker's self-perception at the moment of stuttering during an oral reading task and the highly agreed upon joint perceptual judgment by listeners of stuttered and non-stuttered points in the same oral reading sample, and (b) is the correlation significant? The hypothesis was that there would not be a significant correlation between self-perception of speakers and perception by listeners. The null hypothesis was that a significant correlation would occur between speaker and listener perceptions of stuttered points.

CHAPTER II

REVIEW OF THE LITERATURE

In this chapter, several definitions of stuttering are reviewed, followed by a brief history of methods of identifying and measuring stuttering. Problems with past methods are also addressed, as is the rationale for the current study.

Definition of Stuttering

Identification, measurement, evaluation, and treatment of stuttering are related to a fundamental concept or definition of stuttering. Yet every author who has attempted to define stuttering has composed a unique definition based upon personal beliefs and experience. Many have declined to define stuttering, operating under the assumption that its nature is generally agreed upon, if not understood. Others make reference to the inadequacy of prior definitions without offering an alternative. As theories of the nature of stuttering have changed over the years, so too have proposed definitions.

According to Bloodstein (1993) and Wingate (1988), prior to the 20th century, references to stuttering tended to be speculative of various causative factors. Van Riper (1982) offered several examples of early definitions which were either overly broad, very restrictive, or incomplete, and which did not gain wide acceptance.

Wendell Johnson's (1946) semantogenic theory of stuttering was popular for many years. He believed that stuttering is an evaluational disorder. "It is what results when normal nonfluency is evaluated as something to be feared and avoided; it is, outwardly, what the stutterer does in an attempt to avoid nonfluency" (Johnson, 1946, p. 452). As research has provided increasing evidence for a neuromotor basis for stuttering, Johnson's semantogenic theory has been largely discredited, though its

popularity among the general public, and even among many practicing speech-language pathologists, has persisted.

In 1941, Van Riper (as cited in Hahn, 1956) stated that stuttering is the result of a weakness in the nervous system involving coordination and timing, and that psychosocial problems are the result, rather than the cause of the speech disorder. In his 1982 book, Van Riper devoted an entire chapter to discussing attempts at definition, and offered the following: "Stuttering occurs when the forward flow of speech is interrupted by a motorically disrupted sound, syllable, or word or by the speaker's reactions thereto" (p. 15). Van Riper's view remained remarkably consistent over the years, and corresponds to the current thinking that underlies research into stuttering as a neuromotor disorder.

During the 1960s and 1970s, psychological theories dominated the stuttering literature. An example is Sheehan's (1970) theory of stuttering: "Basically, stuttering is not a speech disorder but a conflict revolving around self and role, an identity problem" (p. 4). While it is acknowledged that stress can influence stuttering, psychological conflict as the cause of stuttering has also been discredited.

With the advent of behaviorism came the possibility of quantifying and measuring stuttering behavior, and the tendency to define everything in terms of conditioned learning. Wingate's (1964) definition is the one most commonly accepted and used clinically today:

> The term stuttering means:
>
> 1. (a) Disruption in the fluency of verbal expression, which is (b) characterized by involuntary, audible or silent, repetitions or prolongations in the utterance of short speech elements, namely: sounds, syllables, and words of one syllable. These disruptions (c) usually occur frequently or are marked in

character and (d) are not readily controllable.

2. Sometimes the disruptions are (e) accompanied by accessory activities involving the speech apparatus, related or unrelated body structures, or stereotyped speech utterances. These activities give the appearance of being speech-related struggle.

3. Also, there are not infrequently (f) indications or report of the presence of an emotional state, ranging from a general condition of "excitement" or "tension" to more specific emotions of a negative nature such as fear, embarrassment, irritation, or the like. (g) The immediate source of stuttering is some incoordination expressed in the peripheral speech mechanism; the ultimate cause is presently unknown and may be complex or compound (p. 488).

Wingate's definition is descriptive of both speech and non-speech characteristics associated with stuttering, acknowledges difficulty in controlling the behaviors described, and includes an acknowledgment of external or internal factors which may be significant.

In his 1990 article, Perkins discussed the development of various theories of the nature, cause, measurement, and clinical management of stuttering. He offered his own production-oriented definition, which encompassed several elements from prior theories, and also included the recommendation that stuttering should be identified based on the speaker's perception, rather than on the listener's perception. This was an innovative approach, as recent definitions had focused primarily on listener perception as the definitive means of identifying stuttering. Perkins (1990) wrote, "stuttering is the involuntary disruption of a continuing attempt to produce a spoken utterance" (p. 376). His revised definition appeared in Perkins, Kent, and Curlee (1991): "Stuttering

is disruption of speech experienced by the speaker as loss of control" (p. 734). Although these definitions were accepted with some degree of enthusiasm, their validity and reliability remained untested.

Bloodstein (1990) suggested three separate definitions of stuttering, any of which could be used, depending on one's use or preference. The first is based on the listener or observer perception; the second is the descriptive, dictionary-type definition; and the third is the one proposed by Perkins, the perception by the person who stutters.

The relationship between Bloodstein's (1990) first and third definitions has been only minimally explored. Although it is likely that clinicians frequently discuss instances of stuttering with their clients, it is also likely that in the final analysis, the clinician's perception and evaluation will prevail. In research, the more quantifiable listener perception method has been almost universally employed, although there is a problematic lack of inter-rater agreement, which will be discussed later.

More recently, Bloodstein (1995) offered the following definition of stuttering: "Whatever is perceived as stuttering by a reliable observer who has relatively good agreement with others" (p. 10). He stated further that:

> If we want to be guided by a more "objective" definition we must not ask questions about "stuttering," but about repetitions, prolongations, broken words, speech rate, and the like, and must be content with answers that are not about stuttering, but about repetitions, prolongations, and so forth (p.10).

In other words, if we accept that definitions of stuttering vary, we must be cautious in our use of the word, and are safer referring to the exact behaviors of which we wish to speak, or risk misunderstanding.

Cordes and Ingham (1995) echoed the need for consensus in the field regarding the theory, and/or definition of stuttering. They noted that existing definitions in many cases contradict one another. Yet they too persisted, until recently, in focusing on listener perception in identification of stuttering.

Stuttering Identification and Measurement

Stuttering vs. Not Stuttering

Stuttering is an apparently universal phenomenon, as references to it persist throughout recorded history (Van Riper, 1982). The earliest and most basic form of stuttering identification is the determination of whether or not a person stutters (Cordes & Ingham, 1994). Williams and Kent (1958) and Boehmler (1958) investigated listener identification of stuttering in short segments of disfluent speech. They compared listener perceptions to the severity and the specific types of disfluencies present in the samples judged, and challenged the assumption that stuttered speech is easily distinguished from normal disfluency. Both found that sound and syllable repetitions were the most likely to be labeled as stuttering by listeners, but that inter-rater reliability was low in general, and neither study investigated self-evaluation by speakers.

MacDonald and Martin (1973) and Curlee (1981) conducted studies investigating whether or not there is a distinction between stuttering and nonstuttered disfluency. Their experiments were conducted using exclusively listener perception as a means for evaluating speech samples. While both studies demonstrated that stuttering and normal disfluency are not clearly distinguishable from one another by listeners, they also revealed low inter- and intra-rater reliability among judges. Neither of these studies included any self-evaluation component.

Total Count

Prior to the mid 1970s, comparisons of inter-rater judgments of stuttering behavior were based on a total count of stuttering during a given passage. An early such study was conducted by Tuthill (1946). Tuthill's series of experiments revealed low inter-rater reliability. Cordes and Ingham (1994) pointed out that an impression of adequate reliability of total count analysis has nevertheless persisted. In 1975, Young devised a system to analyze stuttering identification based on a point-by-point analysis, and devised an index of agreement. The index is a number between 0.0 and 1.0, with 1.0 representing perfect agreement. Young refined this system in 1977. This formula created data with much greater specificity, and also brought into sharp focus the problems of low inter-rater reliability of judgments of stuttering behavior, especially when compared on a point-by-point basis. In Young's 1975 study, mean agreement index was 0.52 (with a range of 0.21 to 0.83) for one analysis, and 0.50 (with a range of 0.18 to 0.77) for the second. Total count agreements for these analyses were 0.92 and 0.93 respectively (Young, 1975). This discrepancy highlights the misleading nature of total stuttering counts when compared to point-by-point analyses.

Behavioral Measurement

Bloodstein (1995) summarized five ways of measuring stuttering: frequency of stuttering, mean duration of stutterings, frequency of specified disfluency types, speech rate, and severity rating. Under the category of specific disfluency types are eight features suggested by Johnson (1959): interjections of syllables, sounds, words, or phrases; repetition of sounds or syllables; repetition of words; repetition of phrases; revisions; incomplete phrases; broken words; and prolonged sounds. These categories all represent behavioral manifestations of stuttering.

Ham (1989) surveyed recognized professionals in the field of stuttering research and treatment on their personal approaches to the evaluation and measurement of stuttering. He surveyed areas including stuttering spasm elements, stuttering avoidance behaviors, criteria in counting instances of stuttering, and speech modes measured. There was a fair amount of disagreement among these professionals. A great majority of professionals labeled certain features, namely broken-word (two or more syllables of a multisyllable word) repetitions, syllable repetitions, sound repetitions, prolongations, and stoppages, as stuttering. Specific avoidance behaviors did not yield consistent responses. Ham reported that a small group preferred not to count instances at all because they identify stuttering as a "self-defined event" (p. 241) for which external measurement is meaningless. He reported a dramatic instance of the difference between the internal experience of stuttering and external observation, noting that he had treated a woman for 20 weeks, during which time she did not stutter at all based on external observation, but spent the entire period in a constant state of anxiety over the possibility that she might stutter. Ham concluded by stating that definition, measurement, perception, and self-perception of stuttering are all in disarray.

Tuthill (1940), in an attempt to obtain an "extensional definition" (p. 189) of stuttering, showed that individuals who stuttered perceptually identified different speech elements than did nonstuttering speech clinicians or lay individuals when listening to a recording of the stuttered speech of others and identifying points where they believed stuttering to have occurred. In this study, no comparison to self-evaluation of the speaker was made. Although these judgments were all based on listener evaluation, they do reveal clearly that differences in the perspective of the

listener affect judgment, and more specifically, that people who stutter judge stuttering differently even as listeners.

Time Interval Analysis

Cordes, Ingham, Frank, and Ingham (1992), concerned with the lack of inter- and intra-rater agreement, devised a different means of evaluating stuttered speech. They subdivided speech into timed units of various intervals, ranging from 0.2 to 7.0 seconds each, and addressed the question of whether or not stuttering had occurred during these time intervals, looking for the smallest time interval that yielded high reliability among judges. Agreement was 50% or less for under 2.0 seconds. The possibility of chance agreement increased over 3.0 to 4.0 seconds. An interval of 3.0 seconds yielded approximately 60% agreement. An interval of 5.0 seconds was chosen as the shortest interval to yield high agreement (83% to 93%) without a high chance agreement factor. In other words, for high agreement among listeners to take place, a relatively long interval of time must be utilized. Considering the number of syllables that can be uttered within a 5-second interval (approximately 15-20), the utility of this technique is questionable.

Cordes and Ingham (1996) attempted to increase agreement further by training listeners with a comparison to the evaluation of a particular speech sample by persons judged to be experts in stuttering. Agreement was improved in this way, but a 5-second time interval was still required. Again, the measure was a listener perception task. The subjects were not consulted for their opinions of when stuttering had occurred.

Acoustic Temporal Measurement

In order to investigate the relationship between internal and external perception, Kelly and Conture (1988) designed a study examining the acoustic

characteristics of authentic stuttering as compared to simulated stuttering by the same subjects. In their study, the stuttered speech of adults was compared by acoustic measurement of six temporal features to the repeated imitations of the same stuttered words. Self-perception by the subjects of the controllability of both stuttering and imitations was also surveyed. The purpose was to find a relationship between an external measurement of stuttering and perception of stuttering by a person who stutters. No acoustic differences between authentic and simulated stuttering were found by acoustic temporal measurement, corroborating Moore and Perkins's 1990 claim that true stuttering cannot be reliably detected by external perception. Kelly and Conture (1988) concluded by suggesting that people who stutter do perceive their level of control over their stuttering, and that external behaviors associated with self-perceived stuttering merit further investigation, as differences not identified in this study may exist.

Self-Analysis

In 1990, Perkins raised the question of whether or not listener perception is a valid means of identifying stuttering. He focused on the speaker's experience of a loss of control as the primary means of identifying stuttering, as opposed to listener perception. To investigate further the validity of acoustic information as representative of authentic stuttering, Moore and Perkins (1990) conducted an experiment in which a person who stuttered identified instances of stuttering as she experienced them, then replicated acoustically these instances without actually losing control or feeling that she was stuttering on the replications. In this 1990 single-subject study by Moore and Perkins, it was asserted that acoustic and behavioral manifestations of stuttering represent only part of the total experience of stuttering. The subject produced phrases, signaling instances of authentic stuttering, then

simulated the same stuttering events. She attempted to distinguish auditorily between the two recordings at one-hour, one-day, and four-day delays. Independent nonprofessional listeners made the same judgment. As was hypothesized, the subject's accuracy of identification of authentic vs. simulated stuttering deteriorated from 100% at the time of occurrence to 54% at four days, compared to 57% by the listeners. As her memory of the experience decayed, so did her ability to distinguish authentic from simulated stuttering. She was unable to auditorily identify authentic stuttering, even in herself.

The following validation procedure was used to support the assumption that the subject was able to detect her own stuttering reliably at the time of occurrence, and then simulate it. The subject was asked to read a passage aloud and signal when she felt that she stuttered. The experimenter would immediately signal to her to complete the utterance. The subject did not in any case continue speaking following a signal to do so, and it was accepted that this was due to a genuine inability to continue. On this basis, the subject's self-evaluations of having stuttered were considered to be valid. After completing a sentence containing stuttering, the subject was asked to replicate the sentence acoustically, with a simulation of her stuttering. As she did so, the experimenter signaled her to continue speaking during each simulated stutter. As she did immediately continue in 31 out of 33 instances (the other two having become actual stutters), the validity of the simulations was accepted. While many studies comparing agreement among raters have been conducted (Curlee, 1981; Kully & Boberg, 1988; Young, 1975), this has been, until recently, the only one published that specifically compared evaluation by a person who stutters to the evaluation by listeners.

The Moore and Perkins (1990) study was, however, a single-subject study, which necessarily limits the implications of the results. Furthermore, the approach was still based on an external measure: the comparison of a speaker's experience to her auditory perception of the experience of stuttering. The strength of this study was its demonstration of the lack of perceptual factors which might distinguish genuine from simulated stuttering. This study took the first step toward demonstrating the difference between the speaker's experience and a listener's perception. Moore and Perkins did not, however, compare point-by-point agreement between the speaker and the listeners. The present study compared point-by-point agreement between a panel of independent listeners and speaking subjects who stutter.

Ingham and Cordes (1997) recently published the results of a study comparing observer judgments and self-judgments of stuttering. Their study included self-judgments under many conditions by 15 adults who stutter, including on-line judgments, off-line judgments of the same speech samples recorded on video, and the speech of one another on video, as well as judgments by 10 authorities on stuttering. The interval examined in all these conditions was the 5-second interval as developed by Cordes, Ingham, Frank, and Ingham (1992). There were differences among speakers in self-agreement rates when comparing off-line judgment tasks to on-line tasks, though speakers were given many opportunities to assure that they were content with their judgments. Only 1 out of 5 speakers made judgments of his own speech that were consistent across on-line and off-line tasks. Ingham and Cordes (1997) suggested that perhaps some speakers are too involved in the act of speaking to identify stuttering reliably as it occurs, though they are able consistently to be satisfied with their observations while watching a video of themselves. These results challenge Perkins's notion that it is the on-line perception of loss of control that must be

considered the most valid means of identification of the stuttering moment. This phenomenon was not consistent across speakers, emphasizing the variability of stuttering from one speaker to another.

In general, authorities on stuttering demonstrated better agreement with speakers than did other persons who stuttered, but overall there were substantial inconsistencies across judgment conditions and across judges, be they speakers on- or off-line, or listeners. Although some intervals generated good agreement, this study did not, overall, demonstrate high agreement between speakers and listeners of stuttered points.

Inter-rater Agreement

Many studies have been conducted investigating agreement between and within listeners (Boehmler, 1958; Curlee, 1981; Kully & Boberg, 1988; MacDonald & Martin, 1973; Perkins, 1990; Young 1975). Rates of agreement have been consistently low, calling into question the validity of using listener perception as an evaluative tool.

Kully and Boberg (1988) studied interclinic agreement of identification of stuttered syllables, and found significant discrepancies from one treatment center to another. They suggested that the high inter-rater agreement reported in previous studies might have been due to observers having trained together or having had more explicit instructions. Yet when Young conducted his 1975 inter-rater reliability studies, he found that manipulating such variables as instructions to raters and definitions of stuttering to be identified had no significant effect on results. Similarly, Curlee (1981) manipulated variables such as graduate/undergraduate student raters, speech/reading samples, presence/absence of a working definition of stuttering, and order of samples rated, and found, again, no significant effect on inter-rater agreement.

In all these studies, agreement was in the range of 50%-60% or lower when evaluated on a point-by-point basis.

Cordes (1994) stressed the importance of carefully establishing reliability in research, and of detailing methods used in establishing reliability. She concluded that data obtained by human observation will likely always be somewhat unreliable, due simply to the uniqueness of each human observer.

Cordes and Ingham (1995) conducted a study of inter-rater agreement using a time interval of 5.0 seconds rather than a word-by-word comparison. Higher levels of agreement (83% to 93%) were achieved in their study. When one considers, however, that between 15 and 20 syllables are normally spoken during a 5-second period of relatively fluent speech, the question arises as to how much information is really gained in this way, other than the convenience of adapting the task of listener perception to produce better agreement.

Critiques of Past Methods

During much of the 20th century, stuttering was thought to be the result of poor psychological adjustment. While it is acknowledged that stuttering is influenced by psychological components, theories and treatments of stuttering as primarily a disorder of psychological origin have been largely discredited (Bloodstein, 1995).

Conture (1982) addressed the use of behavior modification techniques in identifying and treating stuttering. He cautioned against complete confidence in modification of a behavior that is not completely understood. While we can describe speech behavior that is identified as stuttering, and can share this information with the client, the process that underlies the behavior is not understood. Therefore, exactly what we are attempting to modify is at this time unclear (Conture, 1982).

As many researchers have determined, listener perception as a means of identifying and measuring stuttering is a less than perfect solution, given poor reliability. Cordes' and Ingham's (1994) time interval measurement, while yielding higher reliability, does so at the expense of specificity.

The objection to reliance on self-perception of stuttering (Bloodstein, 1990; Kelly & Conture, 1988) is that it is difficult, if not impossible, to measure or scientifically quantify. Kelly and Conture (1988) suggested that if self-reports of stuttering were found to correlate well with external measures, they might be given more credence. This view, however, still considers the external observation to be the standard against which the self-reports are to be measured. In other words, the self-report would not be considered valid or acceptable unless it corresponded to external perception. Yet, as Cordes (1994) pointed out, external observation is also subjective.

Research Paradigms

Young (1994) discussed the evolution of stuttering research paradigms. According to Young, the first basic research paradigm was based on perceptions by stutterers. This paradigm was discarded in favor of the behavioral model, wherein attention was focused on contingent environmental consequences. Research has turned most recently toward investigations of neuromotor or other physiological differences between those who stutter and those who do not. Advances in genetic and medical research and technology continue to open new avenues of research. A growing body of evidence points toward a genetic predisposition to stutter. While research in this direction is both fascinating and promising, it is not at this point very relevant to treatment, and does not generally consider behavioral manifestations of stuttering, just as the behavioral paradigm did not take into consideration the perceptions of the speaker who stuttered.

As separate as these avenues of research appear to be, they have in common their goal of a better understanding of stuttering. If barriers between these theoretical paradigms were ignored, a more holistic view of this complex disorder might yield better insights. This investigation into the compatibility (or lack thereof) of the personal perception of the speaker who stutters with the perception of stuttering behavior by listeners was proposed in the spirit of a holistic view. In a culture that views stuttering as a disorder, self-perception of stuttering must be viewed in the context of behavioral manifestations and the reactions of listeners. The behavioral model still governs treatment, as behavioral manifestations of stuttering can be manipulated by techniques such as reinforcement, reward, and punishment, but strict behaviorism by definition excludes internal cognitive factors. If, as Perkins (1990) contends, a self-perceived loss of control is the element that separates stuttering from other disfluency, then an internal cognitive process is imperative to the identification of stuttering.

Acknowledging, then, that self-perception of a loss of control is possible, and acknowledging that stuttering cannot be reliably detected by external measurement or perception alone, the current study proposed to compare the self-perception of subjects who stutter to the perception of speech-language pathology graduate student listeners, in order to determine to what degree the two perceptions correspond. The research questions proposed by this study are: (a) what is the correlation between a speaker's self-perception at the moment of stuttering during an oral reading task and the highly agreed upon joint perceptual judgment by listeners of stuttered and non-stuttered points in the same oral reading sample, and (b) is the correlation significant?

CHAPTER III

METHOD

Six persons over the age of 14 who stutter read aloud a list of 25 sentences, and marked on-line point-by-point stuttering as it occurred. This reading was audiotaped for later analysis. Eight speech-language pathology graduate students independently listened to the tape of each subject, and marked point-by-point perceived moments of stuttering. A comparison was then made between each subject's self-evaluation and the joint evaluation by the listeners. Only points where at least 7 out of 8 listeners agreed that stuttering did or did not occur were considered in final analysis, although judgments for all points were recorded.

Subjects

Speakers

The speaking subjects (speakers) were 6 people over the age of 14 who stutter. Speakers evidenced normal articulation and language abilities, as determined by clinical records and by the professional judgment of the researcher during casual conversation. All were free of hearing impairments, as evidenced by pure tone audiometric screening in both ears. Screening was conducted at 20 dB HL in a sound treated room at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. Speakers were free of cognitive and/or physical disability, and were able to read text aloud. Speakers evidenced stuttering of at least moderate severity, as demonstrated by the Stuttering Severity Instrument for Children and Adults, Third Edition (SSI-3) (Riley, 1994) and/or self report. An informed consent form was signed by each speaker prior to participation in the study (see Appendix A). Speaker age, gender, stuttering severity and percentage of stuttering during the reading task are summarized in Table 1.

Table 1

Breakdown by Speaker of Gender, Age, and Stuttering Severity

Speaker	Gender	Age	Stuttering Severity
S1	M	18	Moderate
S2	M	47	Moderate
S3	M	28	Moderate
S4	M	39	Moderate
S5	M	29	Moderate
S6	M	30	Severe

Listeners

The listening subjects (listeners) were 8 Portland State University graduate students in the Speech and Hearing Sciences department. All either had completed or were currently enrolled in a graduate course in stuttering. None had a personal history of stuttering. All were free of hearing impairments, as evidenced by pure tone audiometric screening at 20 dB HL in both ears. Screening was conducted in a sound treated room at 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. An informed consent form was signed by each listener prior to participation in the study (see Appendix B).

Text

The text read by all speakers consisted of 25 sentences generated by the researcher. All phonemes in the English language were represented at least once in the entire text (see Appendix C). The order of presentation of sentences was randomly varied among subjects. Each sentence was printed twice on a single sheet (see Appendix D). The text of the lower sentence on each page was broken down by means of slash marks at the beginning of each word and at the end of each word, thus

dividing each sentence into words and spaces between words, including a space prior to the first word of the sentence. Each of these words and spaces was termed a point. The 25 sentences including words and spaces before words equal a total of 2052 potential points of stuttering (342 for each of 6 speakers). Each point was assigned a number on a master copy of the text.

Speaker Perception Task

Speakers scheduled individual times with the researcher. Speakers were given a packet containing the 25 sentences, one sentence printed twice per page as explained above. Each speaker was instructed to read the top sentence on the page aloud and to mark a red X on the lower sentence at each point between slashes where stuttering occurred. See Appendix E for Instructions to Speakers.

Listener Perception Task

Listeners scheduled individual times with the researcher. Individual listeners were instructed to listen to the audiotaped reading of each speaker and to mark a red X between slashes where stuttering was perceived. Copies of the sentences were provided in the same order they were read by speakers. Listeners were permitted to replay any segments they chose. See Appendix F for Instructions to Listeners.

A definition of stuttering was not offered, nor was a distinction between stuttering and normal disfluency. Each listener judged and marked points based on personal perception. Listeners and speakers were discouraged from discussing their perceptions with one another until all data had been collected.

Instrumentation

Recordings were made in a professional sound booth with an Audio-Technica Condenser Lo-Z unidirectional microphone, with a mouth-to-microphone distance of

20 cm. Maxell DM60 Digital Audio Tape was used in a Sony Digital Audio Recorder, model PCM-2300.

Listeners marked sentence lists while listening to the digital audio tape on the same Sony recorder through a Phillips FA950 stereo amplifier and JBL speakers. Output levels were set by individual listeners at a comfortable loudness level.

Recording Technique

Speakers were assigned a number (S1 through S6), which was marked on each page of text. Each speaker individually read aloud from the text, which was already marked with slashes separating words and spaces between words. The researcher was present and identified the subject number onto the tape and audiotape-recorded the sample as it was read. The subject marked on the text each moment of stuttering as it occurred. This on-line marking by subjects was intended to assure that it was the internal perception of stuttering that was recorded.

Individual speakers' markings of stuttered points were recorded onto a form with a column of cells, each representing a point. A point were no stuttering was perceived was assigned the value of 0. A point marked as stuttered was assigned the value of 1. Adjacent were columns of cells to contain entries of 0 or 1 for the perceptions marked by listeners L1 through L8.

Analysis Technique

The responses of both speakers and listeners were plotted onto the matrix as described above, with potential points of stuttering listed vertically on the left. Only points where at least seven out of eight listeners agreed that stuttering did or did not occur were considered. The elimination of instances where listeners disagreed removed the inter-rater disagreement problem. Responses by listeners were summed for each point. A sum of 0, 1, 7, or 8 indicated agreement of seven out of eight

listeners (0 = unanimous absence of stuttering, 1 = 7 out of 8 agreed on absence of stuttering, 7 = 7 out of 8 agreed on presence of stuttering, and 8 = unanimous presence of stuttering). These points were compared to the 0 or 1 entered for the speaker's self-evaluation and agreement was analyzed. Points where the listeners' sum was between 2 and 6 were disregarded in analysis.

Each of the qualifying evaluation points was compared to the self-evaluation by each speaker of whether stuttering had occurred or not. Analysis was made using the SPSS software program and Cohen's kappa (Cohen, 1960), a correlation analysis tool which yields a measure of reliability that controls for the likelihood of chance agreement.

Cohen's kappa reflects the level of agreement likely to occur by chance and removes it from the observed agreement, yielding a more accurate measure than a simple correlation based on percentages can (Bakeman & Gottman, 1986; Cohen, 1960; Kraemer, 1982). Kappa is expressed as: $K = (Po - Pc) / (1 - Pc)$. It yields a figure between 0 and 1, (the kappa), 0 representing agreement at the level expected for chance agreement only, and 1 representing perfect agreement. Negative values of kappa are also possible, indicating less than chance agreement (Cohen, 1960).

CHAPTER IV

RESULTS

The kappa value for this study was .276, with a level of significance of p = .001, indicating highly significant agreement between speakers' and listeners' perception of stuttered points, and affirming the null hypothesis. Nevertheless, although the agreement demonstrated was highly significant, it was not strong agreement, as a value of kappa closer to 1.0 would have demonstrated, and is of little practical importance.

A total of 1,970 points out of 2,052 (96%) generated agreement among at least 7 out of 8 listeners, and were compared to the self-evaluations of the speakers. This leaves 70 points where listeners disagreed with one another and 12 points which were not analyzed, as Speaker #1 omitted one sentence. Out of this 1,970 points, listeners agreed on 1,953 points of no perceived stuttering and 17 points of perceived stuttering.

Of the 1970 points analyzed, speakers generated self-evaluations of 1,910 points where no stuttering occurred and 60 points where stuttering did occur. Speakers and listeners agreed at 1,904 points that no stuttering had occurred, and speakers and listeners agreed at 11 points that stuttering was present. Speakers and listeners disagreed on a total of 55 points. At 49 of these, speakers identified stuttering, but 7 out of 8 listeners failed to perceive it, and at 6 points, 7 out of 8 listeners perceived stuttering when the speaker did not.

It should be noted that there were no instances where all eight listeners agreed that stuttering had occurred, though there were 1,891 instances where the absence of stuttering was unanimously agreed upon among listeners. Of the 70 points not analyzed due to listener disagreement, speakers perceived stuttering in 25, or 35.7%.

Examination of results by individual speaker yields greater detail and contributes to an understanding of the basis for the .276 result for the entire study. Table 2 summarizes a breakdown by speaker of total points out of a possible 342 (330 for S1, due to the omitted sentence) marked as stuttered.

Table 2

Total and Percentage of Stuttered Points Marked by Speakers

Speaker	Stuttered Points	Total Points	% Stuttered Points
S1	13	330	3.9
S2	13	342	3.8
S3	5	342	1.5
S4	4	342	1.2
S5	3	342	.9
S6	47	342	13.7

Table 3 breaks down total points self-perceived by each speaker as not stuttered (NS) and stuttered (S), and compares them to the joint perceptions of listeners, expressed as sums. Sums represent a total of listeners' 0 (not stuttered) and 1 (stuttered) entries for each individual point. Sum 0 represents a total of 8 entries of 0 for a given point, Sum 1 represents an entry of 1 by 1 listener (7 out of 8 marked 0, therefore this is a point of agreement), Sum 5 represents entries of 1 by 5 listeners, and so forth. Points included in Sum 2 through Sum 6 were not considered in the kappa analysis as they were considered disagreed by listeners.

For example, Speaker S1 marked 317 points where no stuttering was felt. Of these, 301 were unanimously agreed by listeners as not stuttered (Sum 0), 5 were agreed by 7 out of 8 listeners as not stuttered (Sum 1), and a total of 11 were judged as

Table 3

Comparison of Speaker Perceptions of Non-Stuttered (NS) and Stuttered (S) Points to Joint Listener Perceptions Expressed as Sums

			Distribution of Sums of Listener Perceptions							
Speaker		Speaker Total	Sum 0	Sum 1	Sum 2	Sum 3	Sum 4	Sum 5	Sum 6	Sum 7
S1	NS	317	301[a]	5[a]	2	2		4	3	
	S	13	7[b]	4[b]	2					
S2	NS	329	317[a]	9[a]	2			1		
	S	13	2[b]		1		1	2	5	2[a]
S3	NS	337	324[a]	7[a]	2	2			1	1[b]
	S	5							1	4[a]
S4	NS	338	336[a]		1	1				
	S	4	4[b]							
S5	NS	339	321[a]	11[a]	6	1				
	S	3	1[b]		1			1		
S6	NS	295	257[a]	16[a]	8	1	2	2	4	5[b]
	S	47	21[b]	10[b]	6		1	2	2	5[a]
Totals		2040	1891	62	31	7	4	12	16	17

[a]Represents agreement

[b]Represents disagreement

stuttered by 2 to 6 listeners (Sum 2 to Sum 6), and were therefore not included in the analysis. There were no points where 7 or 8 listeners perceived stuttering for S1. S1 marked 13 points as stuttered. At 7 of those points, 8 listeners marked no stuttering (Sum 0), at 4 points, 7 listeners marked no stuttering (Sum 1), and 2 points were not considered, due to listener disagreement. In no instance did 7 or 8 listeners perceive stuttering.

Individual kappas were calculated for speakers S2, S3, and S6. Separate kappas could not be calculated for individual speakers S1, S4, and S5 because there were no instances where seven out of eight listeners identified stuttering, so the analysis grid could not be completed. The individual kappa for speaker S2 is .664, the kappa for S3 is .887, and the kappa for S6 is .176.

Speaker 2 identified 13 instances of stuttering, 2 of which were also identified as stuttered by 7 out of 8 listeners, 2 of which were not identified by any listeners as stuttered, and 9 of which were not considered due to listener disagreement. This leaves 329 points identified as not stuttered by S2, and agreed as not stuttered by listeners, and only 3 points disregarded due to listener disagreement. The kappa for S2 is .664, $\underline{p} = .001$.

Speaker 3 identified 5 points of stuttering, 4 of which were also identified as stuttered by 7 out of 8 listeners, and 1 of which was not considered due to listener disagreement (6 out of 8 in this case). This leaves 331 points identified as not stuttered by S3, and agreed as not stuttered by listeners, 1 point identified as not stuttered by S3 but identified as stuttered by listeners, and 5 points disregarded due to listener disagreement. The kappa for S3 is .887, with $\underline{p} = .001$, the highest value calculated for any speaker in this study.

Speaker 6 identified 47 instances of stuttering and 295 non-stuttered points. Of the 47 stuttered points identified by S6, only 5 were identified as stuttered by 7 out of 8 listeners. Thirty-one were identified as not stuttered by listeners, and 11 were disregarded due to listener disagreement. Of the 295 points not identified as stuttered by S6, 273 were agreed as non-stuttered by listeners, 5 were identified as stuttered, and 17 yielded listener disagreement. The kappa for S6 is .176, $p = .001$, which is quite low despite its significance.

CHAPTER V

DISCUSSION

Although significant agreement beyond that expected by chance was confirmed in this study, the value of Cohen's kappa, .276, was not high. Fleiss (1981) suggested that values of kappa below .40 represent poor agreement beyond chance, values above .75 represent excellent agreement, and values in between indicate fair-to-good agreement. Bakeman and Gottman (1986) suggested that, although low values of kappa can indeed be significant, kappas of less than .70 should be viewed with caution. Sample size is also a factor. Samples used as examples by Bakeman and Gottman were in the realm of 100, whereas this study has a sample size of 1970, which accounts for the extremely high level of significance of the results.

It is important to note that this task appeared to generate very little stuttering for most speakers, with 47 self-perceived points noted by Speaker 6 being the highest number out of 342 possible points. In light of this, the kappa of .276 reflected mainly agreement of non-stuttered points. It is also important to consider the 70 points not considered due to interrater disagreement, where between 2 and 6 listeners perceived stuttering to have occurred. Out of 85 total points marked by speakers as stuttered, 25 of these, or 29%, were among the 70 unanalyzed points.

There is a large range of total points identified as stuttered by listeners, ranging from 30 for L5 to 86 for L2. Listeners L5 and L6 consistently marked fewer points than did other listeners. Total counts of points marked as stuttered per listener are summarized in Table 4, with mean total counts compared to self-evaluation total counts by speaker. For this calculation, all points, regardless of listener disagreement, were considered.

There were differences among both speakers and listeners in terms of whether stuttering was perceived to be before a word or on the word. Listener 4 almost

Table 4

Total Points Individual Listeners Marked as Stuttered With Comparison of Average to Total Points Marked as Stuttered by Speakers

	Listener									
Speaker	L1	L2	L3	L4	L5	L6	L7	L8	Mean	Self
S1	10	10	7	6	6	2	15	5	7.60	13
S2	11	13	11	11	2	2	17	11	9.75	13
S3	7	10	7	8	9	6	9	8	8.00	5
S4	1	1	0	1	0	0	1	1	0.60	4
S5	3	11	3	3	2	0	10	1	4.00	3
S6	21	41	20	29	11	21	29	23	24.38	47
Total	53	86	48	58	30	31	81	49		

exclusively marked points before words. Speaker 6 often marked both the point before a word and the subsequent word also, a pattern seldom replicated by listeners. The distinction between points before words and on words probably reduced agreement where stuttering did occur, as some listeners nearly always marked before a word, while others marked on the word. Combining them, however, would have halved possible points of agreement, thereby potentially reducing overall agreement, and certainly reducing significance of results. As these points were considered separately, agreement was not established in cases where perceived stutters were split between on-word and between-word loci of stuttering.

Several examples of this occurred with S6. In some of these cases, such as at points 9 and 10 for S6, 2 points were not included in analysis due to listener disagreement. In this case, S6 marked the point before the word as stuttered, but the word as not stuttered. Two listeners marked the point before the word, and the other 6 marked on the word. These points, if combined, would have been considered as unanimously agreed with the speaker to be stuttered. At other points, such as 81 and 82 for S6, the result of combining points would be slightly different. S6 marked, in this case, both the word and the space before the word as stuttered. Seven listeners marked only the word as stuttered, so this was considered a point of agreement. The eighth listener marked the space before the word only, so this was considered a point of disagreement. Had the points been considered together, unanimous agreement that stuttering had occurred would again have been achieved. At points 155 and 156 for S6, two points of disagreement were recorded, as S6 marked only the point before the word, and 7 listeners marked only on the word.

Overall, for S6, agreement would have increased significantly had the separation of words and spaces between words not been made. As the study was constructed, only 5 out of 47 self-perceived stutters were agreed by 7 out of 8 listeners. It is recommended that any replication of this study separate the text at the end of each word, and consider spaces between words together with the subsequent word. To maintain a large sample size, more speakers or more sentences could be added to compensate. It is not felt that expanding the interval beyond the word level would be advantageous in terms of increasing agreement. It was clear, especially with S6, that dividing the text into word and between-word points created a high number of disagreements that would have been avoided had the text been divided at the end of each word only, but that extending the intervals beyond this would not have increased

agreement significantly. It would, rather, have served to create agreement based on a much less specific judgment, leaving the true level of agreement somewhat ambiguous.

Considering the widely different agreement rates across speakers, the widely different total counts of stuttering perceived by listeners despite similar training, and the low incidence of stuttering throughout this study, results should be viewed with caution. Perhaps the clearest conclusion which can be drawn is that listeners agree with one another and with speakers when no stuttering is present. Nevertheless, of the 4 stuttered points identified by S4, not one was identified by any listener as stuttered, suggesting that, for this speaker, behavioral data are grossly unreliable in detection of stuttering. Similarly, S1's 13 stuttered points all went undetected by a majority of listeners, as did S5's 3 stuttered points.

CHAPTER VI

CONCLUSION

This study showed that, in general, speakers and listeners agree on perception of stuttered points to a degree that is highly unlikely to have occurred by chance. The level of this agreement, however, is alarmingly low, and suggests that the use of a strictly behavioral definition should be approached with caution.

Stuttering is manifested differently in different speakers, as has been demonstrated in the different kappas generated by different speakers when analyzed individually. Stuttering is, furthermore, perceived differently by different listeners, as is demonstrated in the widely different total counts of stuttering throughout the study when tallied by individual listener. The low frequency of stuttering perceived in this study by both speakers and listeners appears to have been responsible for the observed agreement between speakers and listeners. This conclusion would, however, be more reassuring to both clinicians and people who stutter if the level of agreement were higher, and if it were based more on stuttering than the lack thereof. The low frequency of stuttering in this study (based on both speaker and listener perception), coupled with the number of points disregarded due to inter-observer disagreement, may have combined to create a false impression of agreement.

It would be interesting to observe results from a replication of this study with speakers who stuttered more during the reading task. Although the speakers selected for this study were definitely people who stutter, their stuttering was not, in most cases, manifested strongly in the reading task required for this study. A replication of this study would yield results with stronger implications for the field of speech-language pathology if the percentage of stuttered words were greater than 10% at the very least. Speaker 6, who experienced more instances of stuttering during this task

(13.7%) than did Speakers 1-5 combined, generated the lowest kappa of any speaker analyzed individually, bringing the kappa for the entire study down. Speakers who stuttered very little generated much higher agreement, but this agreement was based on generally fluent speech. It is unfortunate that 3 out of 6 speakers were not analyzed individually due to a lack of listener agreed points of stuttering, so a trend cannot be reliably identified showing decreasing agreement with increasing stuttering, though it is suspected that such a trend would become evident.

Clinical Implications

This study suggests that there is a relationship between behavioral manifestations of stuttering and internal perception of stuttering by speakers. Considering, however, the many variables inherent both in individual speakers who stutter and in listeners and the training and perceptions they bring to treatment, the relationship may be a weak one.

It is clear from this study that some people who stutter, such as S3, do so in a way that is easily observed by listeners, and that generates fairly reliable perceptual judgments, using self-perception as the standard for judgment validity. It is also clear, however, that other people, such as S1, S2, S4, and S5 (67% of the speakers in this study), do not stutter in a way that is easily observed by listeners. The analysis of S6, perhaps the most revealing, dramatically highlights the differences in agreement rates that may be obtained based on the interval of measurement used.

Treatment is designed following assessment of stuttering severity. Stuttering severity is most commonly determined by a listener-judged frequency count and measurement of duration of stuttering events. The Stuttering Severity Instrument-3 (Riley, 1994) is the most common norm-referenced assessment instrument, and includes these two measures taken during a speaking task and a reading task (for those

who can read), as well as an evaluation of physical concomitants. This assessment is based exclusively on clinician perception and judgment. Results are widely accepted and assumed to be valid. It is clear from the lack of strong agreement demonstrated in this study, however, that these commonly accepted severity ratings may be highly questionable in the majority of clients who stutter. Other diagnostic tools also are based exclusively on listener judgment

Similarly, both focus of treatment and measurement of progress in treatment become suspect in light of the current findings. If the clinician is not perceiving hesitations that are perceived by the client, such as would be likely for S4, none of whose self-perceived stuttered points were identified by any of 8 listeners, the clinician's judgment, generally accepted as valid, is flawed.

This study has clinical implications that apply most strongly to treatment of adults. Small children are often unaware of their own disfluency, and treatment is often indirect or designed to reshape fluent speech rather than to focus on the moment of stuttering. Clients naturally differ in their level of self-awareness, and this must be taken into consideration in treatment. A discussion at the outset of treatment between the clinician and the client establishing the levels of input by each in the identification of stuttered points is recommended in order to tailor treatment to the needs of each individual client. Further, personal perceptions of the nature of stuttering should be shared by both client and clinician, and an understanding of the working definition to be used in assessment and treatment should be established. A task such as this one, wherein a clinician and a client make independent judgments of stuttered points and then review those perceptions together, would be valuable in identifying the relationship between behavioral manifestations of stuttering and covert manifestations of the disorder, which clearly vary greatly from client to client. In this situation, it

would be important to identify the interval of measurement when comparing client to clinician judgments.

Research Implications

Implications for research on stuttering are many. It has been repeatedly established that agreement among listeners as to what is and what is not stuttered is unreliable at best. When the element of listener disagreement was removed in this study, a comparison between highly agreed points of stuttering or no stuttering and a speaker's self-perception of the same points became possible. The outcome suggests that, although there is a relationship of agreement, the level of agreement is very low for most speakers, and the conclusion is suggested that there is very poor agreement between listeners and speakers on a point-by-point basis of stuttered speech.

If a task similar to this one were performed with each research subject prior to data collection in a behavioral stuttering research project, subjects could perhaps be eliminated if they stuttered in a way that was obscure to listeners. Listeners could also be trained carefully to increase inter-rater agreement. These procedures could increase the validity of a study, but the results would have weaker implications, as only a certain group of people who stutter would be included, those whose stuttering is manifested strongly in observable behaviors. Clinical implications would also be weakened, as clinicians would not have access to the same specific training. Data gained in this manner would not accurately reflect the full spectrum of the complex disorder we call stuttering.

A replication of this study is recommended, with a number of changes suggested by the current results. As the agreement achieved here was based primarily on non-stuttered speech, and the focus of comparison was intended to be stuttered speech, finding speaking subjects with a greater severity of stuttering is recommended.

Alternatively, as speech condition strongly affects stuttering severity, a spontaneous speech sample, as opposed to a reading task, could generate more relevant data in terms of application to daily speaking situations. Videotaping, rather than audiotaping, could provide listeners with more information to assist them in their judgment. Having the speakers judge their own samples on videotape, as Ingham and Cordes (1997) did, would provide a very interesting opportunity for comparison between self-perception on-line and self-perception of external factors, as well as comparison between listener perception and self-perception in both conditions.

This study did not offer any training to listeners, as it had not been previously found to affect judgment (Curlee, 1981). Cordes and Ingham (1996) did, however, increase agreement with listener training, suggesting that listener training can maximize agreement. Further, a discussion was not held with either speakers or listeners regarding a distinction between normal non-stuttered disfluency and stuttering. Such a discussion could also influence results by raising awareness among both speakers and listeners that not all disfluent speech is stuttering.

As the review of literature demonstrates, definitions of stuttering abound. It is likely that each speaker and each listener has a personal working definition of stuttering. This should be discussed with each subject, and for the purpose of research, a single definition should perhaps be agreed upon. The definition suggested by Wingate (1964) is recommended as it is the one most commonly in clinical use today, and has the advantage of being both specific as to behavioral characteristics and broad with respect to potential emotional factors and lack of control by the speaker.

Increasing agreement for the sake of increasing agreement is not a productive pursuit. What is important is recognizing the limitations of listener-based evaluation of stuttered points. There are clearly components of stuttering that defy observation,

and the proportion of these components to observable components varies considerably from one speaker to the next. Research studies with more than one speaking subject tend to level the field of stuttering manifestations and may give data that are representative of no one. Since stuttering is manifested differently in different speakers, perhaps conducting research that is based on groups of subjects whose stuttering is similarly manifested would yield results more meaningful to specific people who stutter.

As Perkins (1990) has suggested, past research into stuttering must be viewed with skepticism. Not including the person who stutters in the equation of our working definition of stuttering risks drawing conclusions that are not based in the full reality of the disorder, and committing an injustice to those we are trying to help.

References

Bakeman, R., & Gottman, J. M. (1986) Observing interaction: An introduction to sequential analysis. Cambridge; New York: Cambridge University Press.

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

Bloodstein, O. (1993). Stuttering: The search for a cause and cure. Nedham Heights, MA: Allyn and Bacon.

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

Boehmler, R. M. (1958). Listener responses to nonfluencies. Journal of Speech and Hearing Research, 1, 132-141.

Borden, G. J. (1990). Subtyping adult stutterers for research purposes. ASHA reports, 18, 58-62.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, No. 1, 37-46.

Conture, E. G. (1982). Stuttering. Englewood Cliffs, NJ: Prentice-Hall, Inc.

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

Cordes, A. K., & Ingham, R. J. (1994). 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.

Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley & Sons.

Hahn, E. F. (1956). Stuttering: Significant theories and therapies (2nd ed.). Stanford, CA: Stanford University Press.

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

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

Johnson, W. (1946). People in quandaries: The semantics of personal adjustment. New York: Harper & Row.

Johnson, W. (1959). The onset of stuttering: Research findings and implications. Minneapolis: University of Minnesota Press.

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

Kraemer, H. C. (1982). Kappa Coefficient. Encyclopedia of Statistical Sciences, 4, 352-354.

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.

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

Perkins, W. H., Kent, R. D., & Curlee, R. F. (1991). A theory of neuropsycholinguistic function in stuttering. Journal of Speech and Hearing Research, 34, 734-752.

Riley, G. D. (1994). Stuttering severity instrument for children and adults (3rd. ed.). Austin: Pro-ed.

Sheehan, J. G. (1970). Stuttering research and therapy. New York, Evanston, & London: Harper and Row.

Tuthill, C. (1940). A quantitative study of extensional meaning with special reference to stuttering. Journal of Speech Disorders, 5, 189-191.

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

Van Riper, C. (1982). The nature of stuttering (2nd ed.). Engelwood Cliffs, NJ: Prentice-Hall.

Williams, D. E., & Kent, L. R. (1958). Listener evaluations of speech interruptions. Journal of Speech and Hearing Research, 1, 124-131.

Wingate, M. E. (1964). A standard definition of stuttering. Journal of Speech and Hearing Disorders, 29, 484-489.

Wingate, M. E. (1988). The structure of stuttering: A psycholinguistic analysis. New York: Springer-Verlag.

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

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

Young, M. A. (1994). Evaluating differences between stuttering and nonstuttering speakers: The group difference design. Journal of Speech and Hearing Research, 37, 522-534.

Appendix A

INFORMED CONSENT BY SPEAKING SUBJECTS

I, _____, agree to take part in this experimental research project conducted by Anne Schagen on a comparison of speaker and listener perception of stuttered points in oral reading.

I understand that to take part as a speaker in this study, I must be diagnosed as a person who stutters, with a severity of moderate or higher, based on results of the Stuttering Severity Instrument for Children and Adults, Third Edition, be at least 14 years of age, and be free of cognitive and/or physical disability.

I understand that the study involves undergoing hearing screening, and oral reading of 25 sentences, which will be recorded. As I read aloud, I will be simultaneously marking on the page points where I feel I stutter. I understand that I will be alone in a sound treated booth during the reading.

I understand that during my participation in this study, I will be spending approximately one half hour in a small room without windows. I understand that I will be free to discontinue participation in the study at any time.

Anne Schagen has told me that the purpose of this study is to explore the correlation between the identification of stuttering by the subjective perception of listeners and the self-perception of stuttering by speakers.

I may not receive any direct benefit from taking part in this study, but the study may help to increase knowledge that may help others in the future.

Anne Schagen has offered to answer any questions I have about the study and what I am expected to do.

She has promised that all information I give will be kept confidential to the extent permitted by law, and that the names of all people in the study will be kept confidential. My responses will be recorded using a subject number only.

I understand that I do not have to take part in this study and may withdraw at any time, and that this will not affect any course grade or hurt my relationship with Portland State University.

I understand that if I have concerns or questions about this study, I may contact the Chair of the Human Subjects Research Review Committee, Research and Sponsored Projects, 105 Neuberger Hall, Portland State University, 503/725-3417.

I have read and understand the above information and agree to take part in this study.

Date:_____ Signature:_____

Date:_____ Signature of Parent or Guardian:_____

Date:_____ Signature of Witness:_____

Appendix B

INFORMED CONSENT BY LISTENING SUBJECTS

I, _____, agree to take part in this experimental research project conducted by Anne Schagen on a comparison of speaker and listener perception of stuttered points in oral reading.

I understand that to take part as a listener in this study, I must be a graduate student in the Speech and Hearing Sciences department of Portland State University, and either have completed or be enrolled currently in a graduate level course in stuttering. I do not have a personal history of stuttering.

I understand that the study involves undergoing hearing screening, and listening to the tape-recorded reading by six speaking subjects of 25 sentences. I understand that I will have a written copy of the text before me, and will be marking points on the page where I perceive stuttering to have occurred. I understand that I may listen to each recorded sentence as many times as I wish.

I understand that during my participation in this study, I will be spending approximately one and a half hours in a small room without windows. I understand that I will be free to discontinue participation in the study at any time.

Anne Schagen has told me that the purpose of this study is to explore the correlation between the identification of stuttering by the subjective perception of listeners and the self-perception of stuttering by speakers.

I may not receive any direct benefit from taking part in this study, but the study may help to increase knowledge that may help others in the future.

Anne Schagen has offered to answer any questions I have about the study and what I am expected to do.

She has promised that all information I give will be kept confidential to the extent permitted by law, and that the names of all people in the study will be kept confidential. My responses will be recorded using a subject number only.

I understand that I do not have to take part in this study and may withdraw at any time, and that this will not affect any course grade or hurt my relationship with Portland State University.

I understand that if I have concerns or questions about this study, I may contact the Chair of the Human Subjects Research Review Committee, Research and Sponsored Projects, 105 Neuberger Hall, Portland State University, 503/725-3417.

I have read and understand the above information and agree to take part in this study.

Date:_____ Signature:_____

Date:_____ Signature of Witness:_____

Appendix C

Sentences

1. She changed the sheets on her mother's bed.
2. They say it rains an awful lot in Portland.
3. Debbie was down in the dumps.
4. What time did you say you were leaving?
5. The jam landed on the yellow napkin.
6. Remember to measure the ceiling height.
7. Children can be very silly.
8. Zinc is said to be good for colds.
9. What's new?
10. I'd like a cheeseburger, please.
11. Don't underestimate the power of suggestion.
12. Potato pancakes are good with applesauce.
13. Roll out the red carpet.
14. Shall we go fishing with Sam Smith?
15. Do I have to clean up my room?
16. All twenty trees will have to come down.
17. Paris is quite different from the French provinces.
18. In India, the afternoons are hot.
19. Do you prefer the beige blouse or the brown one?
20. Does she want me to help her?
21. In that case, I want to go home.
22. He flew to the top of the Eiffel Tower.
23. The camping trip was fun.

24. He's burning the candle at both ends.

25. What long whiskers your cat has.

Appendix D

Don't underestimate the power of suggestion.

/ /Don't/ /underestimate/ /the/ /power/ /of/ /suggestion.

Speaker Number_____

Listener Number_____

Appendix E

INSTRUCTIONS TO SPEAKERS
THESIS PROJECT
Anne J. Schagen

Thank you for agreeing to participate as a speaker in my thesis project. Following this page of instructions are 25 separate sheets, each one printed with a single sentence. The sentence is printed twice, one above the other. The sentence on the bottom is marked by slashes before the first word and between all following words. Please read the sentence aloud one time, and as you read it, mark a red X on the lower sentence at any point where you feel you stutter. You may mark on a word or between words. Please be sure and make the marks AS YOU READ. Do not go back and mark points after you have finished reading. After you complete one sentence, you may turn the page and go on to the next. Continue until you have read all 25 sentences. Do you have any questions?

Appendix F

INSTRUCTIONS TO LISTENERS
THESIS PROJECT
Anne J. Schagen

Thank you for agreeing to participate as a listener in my thesis project. You have been given six original Digital Audio Tapes (DATs), each one holding the recording of a person who stutters reading 25 sentences. You will be listening to them with a copy of the sentences before you, and marking points where you perceive stuttering. Following this page of instructions are six packets of 25 separate unmarked sheets, each one printed with a single sentence. Each sentence is printed twice, one above the other. The sentence on the bottom is marked by slashes before the first word and between all following words.

Please listen to each sentence, and mark a red X on the lower sentence at any point where you perceive stuttering. You may mark on a word or between words. You may listen to any part more than once.

Each speaker reads the same 25 sentences, but the order was varied randomly. The order in which you will be listening to speakers has also been randomized among listeners, so please try to keep tapes and packets in the order presented, as each packet of sentences is in the correct order for that speaker.

If you have any trouble with the equipment, or if there is an error in sentence order, please bring the key with you and find me, probably in the lab. When you have finished, please return the tapes, the file with your sentences, and the key to me or to Rebecca in the office. Do you have any questions?