MEASURING SPEECH NATURALNESS OF CHILDREN WHO DO AND DO NOT STUTTER: THE EFFECT OF TRAINING AND SPEAKER GROUP ON SPEECH NATURALNESS RATINGS AND AGREEMENT SCORES WHEN MEASURED BY INEXPERIENCED LISTENERS

by

ROBIN LYNNE EDGE

(Under the Direction of Anne Bothe Marcotte)

ABSTRACT

The two studies presented in this dissertation investigated the speech naturalness of children who do and do not stutter and the effect of rater training and stuttering severity on speech naturalness. Study One evaluated three groups of judges' speech naturalness ratings of children who stutter. Judges' rater agreement was also reported. The effect of training using a modified version of the naturalness portion of the *Stuttering Measurement System (SMS)* on rater agreement was also investigated. Results of Study One showed statistically significant improvement in the training group's rater agreement when compared to the two control groups.

Study Two was a follow-up replication study that used a larger sample size and added normally speaking children. The effect of stuttering severity (normally speaking children versus children who stutter at mild, moderate, and severe levels) and *SMS* speech naturalness training on speech naturalness ratings and agreement ratings was investigated. Unlike Study One, the *SMS* training was not found to change listeners' agreement ratings to a statistically significant level. Stuttering severity was found to affect both speech naturalness ratings and rater agreement

with normal speakers rated the most natural (lowest naturalness ratings) and children who stutter at a severe level rated the least natural (highest naturalness ratings). Agreement was higher for normal speakers and children who stutter at a severe level than children who stuttered at mild and moderate levels.

These two studies provide initial data from inexperienced listeners using Martin et al.'s (1984) 9-point speech naturalness scale to rate the speech naturalness of children who do and do not stutter and their agreement levels when rating speech naturalness. The effectiveness of the *SMS* training program is inconclusive at this time as data supporting its effectiveness were reported in Study One, but it was not shown to be effective in Study Two. Possible explanations for this as well as future research addressing identified issues are discussed. This follow up research is needed before this measure can be recommended for use with children.

INDEX WORDS: Speech naturalness, children who stutter, interrater agreement, intrarater agreement, 9-point scale, equal appearing intervals, stuttering severity, *Stuttering Measurement System*

# MEASURING SPEECH NATURALNESS OF CHILDREN WHO DO AND DO NOT STUTTER: THE EFFECT OF TRAINING AND SPEAKER GROUP ON SPEECH NATURALNESS RATINGS AND AGREEMENT SCORES WHEN MEASURED BY INEXPERIENCED LISTENERS

by

ROBIN LYNNE EDGE

B.S.Ed., University of Georgia, 1998

M.Ed., Valdosta State University, 2000

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

# MEASURING SPEECH NATURALNESS OF CHILDREN WHO DO AND DO NOT STUTTER: THE EFFECT OF TRAINING AND SPEAKER GROUP ON SPEECH NATURALNESS RATINGS AND AGREEMENT SCORES WHEN MEASURED BY INEXPERIENCED LISTENERS

by

ROBIN LYNNE EDGE

Major Professor: Anne Bothe Marcotte

Committee: Patrick Finn
Duska Franic
Janis Ingham
Yolanda Keller-Bell

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

DEDICATION

To the two most important people in my life, my Mama and Daddy. Without you, none of this would be. I owe y'all everything. I love you both more than you will ever know.

ACKNOWLEDGEMENTS

I can not believe it's time to write the acknowledgement section….that means it is the end of this incredibly long journey. It's not been an easy one (I doubt it ever is) and although I've learned a great deal of academic lessons from this experience, I've learned even more LIFE lessons, and I've found strength I never knew I had. For THAT I am eternally grateful; for the experience, for the lessons, for all the people that have helped me find my inner Wonder Woman, and for the new life in which I am about to embark.

There are SO many people to thank that helped me complete my Ph.D. First and foremost, I want to thank God. There were many times when I wasn't sure I'd make it, not just through the Ph.D. experience, but in life, and He always carried me through and showed me the way. Always, without fail.

I am forever grateful to my Mama and Daddy not only for teaching me to believe in God, but also for believing I could do this, no matter how tough it got. I am especially indebted to my Mama for cheering louder than anyone throughout this marathon. She NEVER doubted, not once, even when I did, and that was a major reason I always kept on going even when I didn't want to. She is THE strongest woman I've ever known, and I thank her for raising me to be stubborn, and for teaching me to NEVER give up. I've never seen one thing whip my Mama yet, and I guess at least a little of that has rubbed off on me. For that I am eternally grateful. I learned from the master, Mama. I only hope I can one day be half the woman you are!

My sister Rhonda, who was there to help me pick up the pieces when they fell apart, has always been a rock (a worrisome rock, but a rock none the less) for me to turn to. I would not

have survived this past year without you. THANK YOU for picking up the slack and standing beside me through the good and not so good times!!! (And y'all don't worry about me in Mississippi, I WILL BE FINE!!!!)

To my babies, Ripley, Buddy, and Phoebe for always being by my side (or in my lap) and being my constant companions (and guards) through the insanely long hours. Thank you also for the kisses, cuddles, and never ending love. Thank you most especially for making me quit when I needed a break even though I didn't want to take one. Y'all often knew better than I did that my body and mind just couldn't take anymore.

To my committee, Drs. Finn, Franic, Ingham, Keller-Bell, and especially Dr. Marcotte. Thank you for your patience, for second (and third!) chances, and for not saying "good enough" even when I wanted you to!! Y'all stuck by me through this lengthy process without complaint and taught me how to be a scientist. I appreciate all of your hard work!!!! Thank you for sticking with me Anne. I know there were times when you thought "no way in the world she'll make it," and I can never thank you enough for not giving up on me. YOU modeled how to be an INSANELY brilliant researcher and writer and I take the skills I learned from you and will use them the rest of my life. Thank you will never be enough!!!

THANK YOU to two very important people (and one donkey!) in my life that have had HUGE parts in my success, Bill Delaune and David Bothe. Bill, when we met I was floundering and on the verge of giving up and you helped me see the way again. I'm not sure this would have happened without your help and encouragement (and your car rides and grass cutting and lawn mower, and bug catcher, and donkey). Your statistics expertise is second to none in my book and I will be forever grateful not only for your stats help, but more importantly for your friendship and advice (and Robby!!). Many times just seeing you and the animals for a social

call was all I needed to clear my head and get back to work. David, you have seen me through family tragedy, academic and personal difficulties, and you showed me how to channel my inner tick, and THAT has gotten me through more than you know!!! (Not to mention the MILLIONS of knots you've worked out of my HEAD, neck, and back over the years!) Both of you helped me see parts of myself I couldn't see and it means more to me than either of you can imagine.

Dr. Bess Taylor, Jill Hollis (and Chuckles!), and my Redista (and Mama Redista). Words can not describe what your MANY chats, encouragement, prayers, and FUN have meant to me (and just letting me fuss when that was what I needed). Y'all have been beside me EVERY step of the way without wavering and y'all will never know how much it means to me. I love each of you to pieces!!

To my academic family, Drs. Lisa Hammett Price, Jessica Richardson, and Jason Davidow. Thank you for sharing pieces of this journey with me. I will always cherish our time together. THANK YOU for showing me how to be a stellar scientist and for modeling successful academic careers for me to emulate (you too Dr. Taylor)!!!

Thanks to the students in the CMSD department at UGA who participated in these studies and who, through their insightful questions and comments, kept me excited about these projects when I was growing tired. Dr. DeChicchis, thank you for being so kind as to allow me to use your students as participants!! THANK YOU will never be enough!!! This would not have happened without you Dr. D!!

To all of my family and friends that encouraged me, sent up prayers for me, and believed in me when I had trouble believing in myself, (my sister and brother, Brooke, Stephanie, Christy, Sheila, the Smathers clan….and the MANY, MANY more of you I don't have room to name) THANK YOU!!!! I love you all more than you will ever know!!

Finally, to my sweet, beautiful, and charming great niece Gracie Faye, thank you for making me smile during the dark days. You will never know what you mean to me. I look forward to watching you grow and take over the world!! Robin The Great adores you with all of her heart!!!!

I leave you with part of a song that's helped me through the tough times "(S)he's one of those who knows that life is just a leap of faith. Spread your arms and hold your breath and always trust your cape...**(S)he did not know (s)he could not fly, so (s)he did**." ~Guy Clark.

Here's to continuing to trust my cape for the rest of my days…

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# LITERATURE REVIEW AND STATEMENT OF THE PROBLEM

## Introduction

Speech naturalness measures have been advocated as a basic part of stuttering treatment outcome assessment for many years especially as some treatments became directed by fluency inducing conditions (Bloodstein, 1995; Conture & Guitar, 1993; Costello, 1983; Curlee, 1993; Ingham & Cordes, 1997; Ingham & Costello, 1985; Ingham & Riley, 1998). Although the research base supporting this recommendation is considerable with adults who stutter, there is much less research investigating the speech naturalness of children who stutter. Researchers seem to assume that results from the adult studies are also applicable to children, but research needs to be adequately designed and conducted to determine if speech naturalness is a construct that can be measured reliably in children who stutter. This research is necessary so the suggestion to use speech naturalness as a part of stuttering outcome assessment with children (e.g., Ingham & Riley) can be based on sound scientific evidence rather than on inferences made from the adult speech naturalness literature.

To begin to address this issue, the following sections will review the measurement of speech naturalness in the speech of people who stutter with a focus on children when such research is available. In order to understand the importance of reliably measuring speech naturalness, one must understand the broader historical context of the concept from which the current research has emerged including issues with defining speech naturalness. This review will begin this introductory chapter (Chapter 1). After a discussion of speech naturalness outside of

communication disorders, current research involving speech naturalness in other areas in speech and language as well as in stuttering will be reviewed including issues affecting its measurement, mostly from the relatively narrow field of speech-language pathology (with most related specifically to stuttering), to address judges' reliability and ways for judges to improve their reliability. The research specific to the measurement of speech naturalness in stuttering was used to develop the two studies presented in this dissertation. Study One, presented in Chapters 2 and 3, assessed rater agreement and the effect of training on rater agreement when listeners rated speech naturalness in children who stuttered. Its results were used as the basis for developing Study Two, which evaluated speech naturalness ratings, rater agreement, and judgment training in the speech of children who do and do not stutter and is presented in Chapters 4 and 5. Finally, Chapter 6 discusses the findings of both studies and presents related future research needs stemming from the two studies presented in this document.

## Speech Quality

Speech naturalness is one element of speech quality, which is a concept intended to quantify the extent to which a given speech sample resembles normal sounding speech (Onslow & Ingham, 1987). In stuttering research, specifically since the 1950s (Bloodstein, 1950), speech quality measurements, including speech naturalness, have become more and more valuable because therapies that use unnatural sounding speech patterns have become increasingly popular (Bloodstein, 1995; Ingham, 1984; Van Riper, 1973). Besides speech naturalness, speech quality has been measured in several other ways, including having listeners rate speech samples based on how "normal" they sounded while using delayed auditory feedback (DAF; Jones & Azrin, 1969) and after treatment with either DAF or other rate control treatments (Perkins, Rudas, Johnson, Michael, & Curlee, 1974). Speech quality measurements have also been used to

determine if listeners could distinguish the fluent speech samples of people who stutter from the speech of normally fluent speakers, both with the people who stutter in treatment (Few & Lingwall, 1972), post-treatment (Ingham & Packman, 1978; Runyan & Adams, 1978; 1979), and when no treatment is reported as being administered (Krikorian & Runyan, 1983; Wendahl & Cole, 1961; Young, 1964). Another way speech quality has been measured is by comparing the acoustic aspects of the speech of people who stutter and fluent speakers (Max & Gracco, 2005; Metz, Onufrak, & Ogburn, 1979; Prosek & Runyan, 1982; Ramig, 1984). Currently, the most commonly used speech quality measurement in stuttering requires listeners to rate speech naturalness using a 9-point equal appearing interval scale (Martin, Haroldson, & Triden, 1984). This scale was chosen over the previous methods for use in the studies presented in these pages as it has a larger body of research including reliability analyses when compared to the other methods. This scale, and the research using it, will be discussed in more detail below after a discussion of the definition of speech naturalness.

## Speech Naturalness Definition

The speech naturalness concept has appeared in professional literature for almost 60 years, beginning with W.M. Parrish's "The Concept of 'Naturalness'" (Parrish, 1951). Parrish discussed speech naturalness by asking how teachers could teach public speakers to speak naturally. In his essay, Parrish discussed what defined naturalness including specifically whether naturalness was determined by nature, grammar, or speaker. He concluded that naturalness is a person's normal means of expression, and also determined that naturalness is not what feels natural to the speaker, but is instead what sounds natural to the listener. This idea is still active in stuttering; recent speech naturalness research often focuses on what listeners perceive as natural rather than what speakers feel is natural (Mackey, Finn, & Ingham, 1997; Martin et al.,

1984; Van Borsel & Eeckhout, 2008). Although this is the approach typically used in speech naturalness research in stuttering, the speaker's view of his or her own speech naturalness has also been measured to determine how he or she perceives the naturalness of his or her own speech (Craig et al., 1996; Finn & Ingham, 1994; Hearne, Packman, Onslow, & O'Brian, 2008; Ingham, Ingham, Onslow, & Finn, 1989).

Parrish's (1951) ideas were empirically tested in 1966 when Nichols performed the first published attempt to reliably measure speech naturalness by investigating the idea that listeners determine what is natural. To do this, he assessed listeners' naturalness ratings of spoken and written sentences by having college students listen to 50 nine-word sentences, 25 that contained frequently spoken English words and 25 constructed from infrequently used English words. The listeners rated how natural each sentence sounded when read by two different speakers and when they read the sentences silently. Naturalness was rated on a 9-point scale in which "1" represented *high naturalness* and "9" represented *low naturalness*. Nichols found a significant difference between the naturalness of the two types of sentences; mean naturalness of frequently used words was 4.17 and mean naturalness of infrequently used words was 5.07. Listeners' reliability was found to be low for individual raters ($r$=. 12-.20) but relatively high for the group ($r$=. 74-.84). The two presentations of the sentences, written versus oral, were not significantly different, implying that the raters did not rate the naturalness of the sentences based on their presentation style. This study supported Parrish's view of listener determined naturalness and was a prequel to a line of naturalness research conducted almost 20 years later by Martin et al. (1984). The latter study also used listeners' ratings of speech naturalness and replicated Nichols' reliability findings of lower individual reliability when compared to group reliability.

Although there is a general consensus that the listener (or reader), not the speaker, determines speech naturalness, as is common in other stuttering variables (see Packman & Onslow, 1998; Teesson, Packman, & Onslow, 2003), what naturalness *consists* of, or its operational definition, is not readily agreed upon. Speech naturalness has been described within the stuttering literature as a multidimensional construct requiring rate, breathstream management, prosody, and self-confidence (Perkins, 1973). Because of this multidimensionality, most speech naturalness researchers do not specifically define the concept but instead opt to have listeners use their own internal standards as to what sounds natural to them (Ingham & Packman, 1978; Ingham & Riley, 1998; Logan, Roberts, Pretto, & Morey, 2002; Martin et al., 1984; Schiavetti & Metz, 1997).

Studies that did provide listeners with an external definition of naturalness defined the concept in various ways. In early speech naturalness research in stuttering, Jones and Azrin's (1969) definition of naturalness included:

> Sometimes this person stutters and sometimes his speech sounds unnatural because he spaces his words in an unusual manner…Here is a good question to ask yourself as you make each judgment: 'If I had just heard this person for the first time reading to an audience, would his speech sound unnatural to me?' (p.224)

In other speech naturalness research in stuttering, examples of good and bad naturalness were given to parents and children in their instructions:

> Base your opinion on the degree of variation in your tone when speaking, the variation in loudness, speed, and appropriate pause/phrasing structure. For example, very poor speech naturalness might be very monotonous or continuous drone, very slow or fast speed, no variation in loudness and very choppy type speech. A very good rating of

> naturalness might be pleasant variation in tone, very acceptable variations on loudness of speech, normal speed and a very appropriate flow of speech. (Craig et al., 1996, p. 815; Hancock et al., 1998, p. 1244)

Natural speech has also been defined to listeners "in terms of intonation, voice quality, rate, rhythm, or intensity adjustments" when measuring the speech naturalness of speakers with dysarthria (Yorkston, Hammen, Beukelman, & Traynor, 1990, p. 551) and more generally as "typical speech you would expect to hear in any given situation" when measuring the construct in speakers with cleft palate (Benoit, Munson, Thurmes, Cordero, Baylis, & Moller, 2008, n.p.). In all of these studies, the authors presented no explanation or data supporting the use of the speech naturalness definition chosen or the components included in the definitions.

When listeners are not given a definition of speech naturalness, it is difficult to accurately quantify what criteria they are using to evaluate the concept. Although the global speech naturalness rating commonly used in stuttering (Martin et al., 1984; see below) may be useful in determining if a person who stutters has met his or her goal of speaking naturally, it does not give the person feedback as to what makes speech sound natural versus unnatural. Franken and colleagues tried to address this problem by developing an instrument to obtain a more "detailed and comprehensive description of the quality of the speech of stutterers before and after therapy" (Franken, Boves, Peters, & Webster, 1992, p. 225). They attempted to determine what specific attributes made speech sound natural (or unnatural) by using 14 7-point bipolar scales adapted from Fagel, van Herpt, and Boves' (1983) voice quality instrument.

In the initial study reported in a textbook chapter, Franken (1987) used Fagel et al.'s (1983) 14 scales with the only change being an added 15$^{th}$ scale, *unnatural – natural*. Fagel et al.'s measure was developed to evaluate seven dimensions of voice quality: melodiousness,

articulation quality, voice quality, pitch level, speaking rate, evaluation, and potency. Franken wanted to determine if the speech naturalness of people who stutter fell into one of these seven categories or if it constituted an additional factor. As seen in Appendix A, factor analysis of the naïve listeners' assessments of the speech samples revealed a 5-factor solution with the first factor, "speech rate and general quality," showing high loadings of the scales *slow – quick*, *dragging – brisk*, *unnatural – natural*, *unpleasant – pleasant*, and *ugly – beautiful,* and accounting for 64.2% of the variance in the ratings. The second factor was an "articulation quality factor contaminated with general evaluation" (p. 288) that had high loadings on the *slovenly – polished*, *broad – cultured*, *ugly – beautiful* and *pleasant – unpleasant* scales and accounted for 15.9% of the variance in the data. The last three factors accounted for much less variance: the "voice dynamics" factor accounted for only 7.7%, "voice pitch and static voice quality" factor accounted for 7.4%, and "potency factor" accounted for only 4.8% of the sample's total variance. These data support the idea that speech naturalness is a multi-dimensional concept, but they do not successfully delineate which specific speech factors account for changes in speech naturalness. The author's conclusion was that Fagel et al.'s scales adequately described the general aspects of speech quality but lacked the specificity needed to describe the speech of people who stutter before and after therapy. Franken and her colleagues investigated this result in their future studies as they adapted the instrument to, in their estimation, better fit stuttered speech.

This adaptation included replacing four scales with words Franken and her colleagues felt were more descriptive of stuttered speech (see Appendix A). *Ugly – beautiful* was replaced with *tense – relaxed*, *husky – not husky* with *weak accentuation – strong accentuation*, *broad – cultured* with *slurred – precise*, *dull – clear* with *fluent – halting*, and *dragging – brisk* with

*unnatural – natural* (Franken et al., 1992; Franken, Boves, Peters, & Webster, 1995). Although the authors' purpose was to change the instrument to one more befitting of stuttered speech in an effort to define speech naturalness, they fell short of this goal as they do not provide evidence to justify why the new scales were added and if they do better define the qualities of stuttered speech. Because of this major limitation, the results of Franken et al.'s studies should be interpreted with caution.

Franken et al. (1992) had naïve listeners rate speech samples from 32 people who stuttered before, immediately after, and 6 months after fluency shaping stuttering therapy along with the speech samples from 20 matched people who did not stutter. As shown in Appendix A, discriminant analysis of these data extracted three "canonical discriminant functions": the "distorted speech dimension", with high correlates of the scales *halting – fluent*, *tense – relaxed*, and *unnatural – natural*; the "dynamics/prosody dimension" with high correlates of the scales *monotonous – melodious*, *flat – expressive*, and *weak accentuation – strong accentuation*; and a "voice dimension" with high correlations of the scales *soft – loud* and *low pitch – high pitch* (Franken et al., 1992). The cumulative percentage of variance explained by these three functions was 55.8%, 95.4%, and 100%, respectively (Franken et al., 1992). Using these three discriminant functions, 80.2% of the speech samples could be correctly classified as pretreatment, posttreatment, or follow up for people who stuttered or as people who did not stutter. Although group differences were found between people who stuttered and people who did not stutter, no difference was found between the naturalness scores for the pre- and posttherapy samples of the people who stuttered.

It appears that the same speech samples were used in the 1992 and 1995 publications because speaker and speech sample characteristics, as well as the mean and standard deviation

for speech rate and syllables spoken per second, are the same (Franken et al., 1992; 1995). The only "difference" between the data in the two studies is that the percentage of syllables stuttered is presented as mean and standard deviation in the 1992 paper and as median and range in the 1995 paper. The authors do not address if the same speech samples were used in both studies. Twenty-four listeners in each study judged the first speech sample tape and 22 listeners in the 1992 paper and 20 in the 1995 paper judged the second tape. Both studies describe these listeners as first semester students in logopedics. Based on these descriptions, it is not known if these were the same students with two dropped between the two studies or if an entirely different set of listeners was used. Because of this ambiguity in the description of the subjects in the two studies, one cannot know if the same data were used in both studies. If these are indeed the same data, the 1995 study is a reassessment of the 1992 data rather than a replication study that would provide more data using the instrument thus enlarging the research base.

As shown in Appendix A, factor analysis was again used (Franken et al., 1995) with three factors found: "voice dynamics," "articulation quality," and "pitch" factor, explaining 57.4, 21.2, and 7.9 percent of the variance, respectively. The "voice dynamics" factor had high loadings of *weak – powerful*, *weak accentuation – strong accentuation*, *flat – expressive*, *monotonous – melodious*, *soft – loud*, and *slow – quick*. "Articulation quality" showed high loadings with *fluent – halting*, *tense – relaxed*, *slurred – precise*, *slovenly – polished*, *unpleasant – pleasant*, and *unnatural – natural*. The "pitch" factor only showed high loadings with *shrill – deep* and *low pitch – high pitch*. The authors noted that *unnatural – natural* and *unpleasant – pleasant* had divided loadings between the first two factors. From this, they concluded that "this shows that naturalness and pleasantness of speech samples are at least two-dimensional concepts-concepts that relate to other, more technical aspects of the speech quality in

complicated ways" (Franken et al., 1995, p. 284). They also conclude that unlike in the previous study, characteristics that determine the naturalness ratings "appear" to be different pretreatment and posttreatment. They found that at pretreatment "articulation quality", "voice dynamics", and "pitch" explained the most variance at 43%, 24.5%, and 11.9%, respectively. For the posttreatment condition, the "dynamics" and "pronunciation" scales accounted for 48.5% of variance with the "pitch" factor accounting for 19.3%. Their line of research provides evidence supporting speech quality as a multidimensional concept, but does not provide a specific external speech naturalness definition, as was their intent.

Although this research supports speech naturalness as a multi-dimensional concept, as mentioned above, the results must be interpreted with caution due to the studies' limitations. The most obvious flaw in this body of research is the lack of explanation regarding the scales chosen for the adapted instrument. The only brief explanation the authors present about how they determined which replacement scales were added was, "we included four scales that seem especially relevant for the evaluation of stuttered speech before and after a fluency shaping therapy" (Franken et al., 1992, p. 229). The authors' explanation of why scales were removed was:

> It can be argued that most of these scales pertain to aspects of the speech that are not supposed to be affected by the therapy; the scale Dragging <–> Brisk, that seemed relevant for (treated) stuttered speech, was removed because its wording suggested an attitudinal interpretation of the samples under evaluation. (Franken et al., 1992, p. 229)

As mentioned above, this explanation is inadequate, and the lack of documentation for the standards deleted from and added to their speech quality instrument results in a psychometrically unsound measure. This research base is also limited in that only two peer-reviewed papers could

be identified using this 14-point bipolar scale. As discussed above, even though they were published as two separate studies, both appear to use the same data with the only differences being the presentation of the stuttering frequency data and the statistical procedure used (factor versus discriminant analysis).

As shown by the above review, no agreed upon formal definition of speech naturalness currently exists. Although attempts have been made to identify what specific variables define speech naturalness, no sound research providing concrete information defining the concept has been conducted. Because of this, rather than providing raters with an external definition of speech naturalness, the currently used and accepted paradigm for measuring the concept is to allow raters to use their internal definition or to provide example videos. The studies presented in this dissertation, therefore, investigated this commonly used paradigm (raters using their internal definition of the concept) via rater agreement data and then trained listeners to rate speech naturalness using audiovisual exemplars to investigate the effect of training on rater agreement. In particular, these projects assessed the currently used internal definition standard by collecting intrajudge and interjudge agreement data for two large groups of listeners on two rating occasions. These initial agreement results, prior to any training or exposure to training samples, can be used to determine the need for an external definition of speech naturalness, because "high levels of interobserver agreement provide some indication that operational definitions of target communication behaviors are adequate. Levels of interobserver agreement that fall below acceptable limits may provide an indication that operational definitions are unacceptable and must be refined" (Kearns, 1990, p. 80). In the current studies "operational definition" is interpreted as the listener's internal definition of speech naturalness. Interrater

agreement, along with other reliability options, will be discussed further after a discussion of speech naturalness research in stuttering.

## Measurement of Speech Naturalness in Stuttering

### *The Speech Naturalness Scale*

Martin et al. (1984) published the first account in stuttering of the now commonly used 9-point naturalness instrument. As briefly mentioned above, their equal appearing intervals scale requires listeners to rate the speech naturalness of speakers based on their innate sense of what sounds "natural." They described the scale as follows (a copy of the scale can be found in Appendix B):

> Each rating scale was a horizontal line approximately 6 inches long (15.24 cm). Vertical lines were placed at each end of the horizontal line, and seven additional vertical lines were spaced evenly between the ends. The vertical lines were numbered 1-9. Above the "1" was typed "highly natural", and above the "9" was typed "highly unnatural." (p. 54)

This naturalness instrument has equal appearing intervals, which requires raters to assign numbers to stimuli that correspond to "a linear partition of the continuum" (Schiavetti, Sacco, Metz, & Sitler, 1983, p. 568). This is accomplished by having the listener divide the continuum of interest (in this case speech naturalness) into categories of equal width, and then assign a category number to each stimulus (in this case speech sample; Young, 1969a). Descriptors may be given for each interval of the scale, or they may only be given for the first and last points as in Martin et al.'s (1984) speech naturalness scale.

An alternative scaling procedure is direct magnitude estimation. Unlike the equal appearing interval-scaling procedure, direct magnitude estimation does not constrain listeners to fit their ratings into the "linear partition" of the speech naturalness scale with fixed minimum and

maximum values (Schiavetti, 1992). This method allows observers to rate each stimulus with a number that is proportional to the perceived ratio between the stimulus and some standard stimulus (i.e., half as severe or twice as severe; Young, 1969a). A standard or model sample with a preassigned speech naturalness rating may or may not be provided to the listeners for comparison when using direct magnitude estimation. If a standard sample is not provided, the listeners are asked to compare all of the stimuli to the first stimulus they rated.

In an effort to determine the construct validity of Martin et al.'s (1984) speech naturalness scale in stuttering, Metz, Schiavetti, and Sacco (1990) used Stevens' (1975) method of determining whether a continuum is prothetic or metathetic. Stevens' method has listeners calculate the speech naturalness of the same speech samples, using both direct magnitude estimation and interval scaling procedures using speech samples that represent the breadth of speech naturalness (Metz et al). To represent the breadth of speech naturalness, Metz et al. had listeners rate both stuttered and nonstuttered speech using both scaling procedures in an effort to determine if speech naturalness is a prothetic or metathetic continuum.

> A prothetic continuum is an additive, *quantitative* continuum that is best scaled with direct magnitude estimation (DME) because observers cannot subdivide a prothetic continuum into equal intervals. A metathetic continuum is a substitutive, *qualitative* continuum that can be scaled with either DME or equal-appearing interval scaling procedures. (Metz et al., 1990, p. 516)

To determine which type of continuum a concept is, Stevens' (1975) method plots the arithmetic means of the interval scale values as a function of the geometric means of direct magnitude estimation for all of the samples. If the relationship between the two is linear, the continuum is

metathetic; but if the interval scale values form a downward bowed curve when plotted as a function of the direct magnitude estimates, then the continuum is prothetic (Stevens).

Both Metz et al. (1990) and Schiavetti, Martin, Haroldson, and Metz (1994) found speech naturalness to be a metathetic continuum that could be scaled by either direct magnitude estimation or equal appearing intervals. Both assessed the reliability of each scaling method, with Metz et al. finding group and individual reliability to be slightly higher for equal appearing intervals. Schiavetti et al. found a bigger difference between the two as they found group reliability to be slightly higher for the equal appearing intervals, but found individual reliability to be much higher for the equal appearing interval scaling procedure than for direct magnitude estimation. Based on these findings, the authors of both of these studies advocated the continued use of the current equal appearing interval scale (Martin et al., 1984). This recommendation was based not only on their studies' results, but also because of the reasonable body of literature that already exists using equal appearing intervals, as well as the adequate group reliability results found throughout this literature (see below). Another reason they advocated for the continued use of the equal appearing intervals scale was the scale's availability and familiarity to most users due to the widespread clinical use of interval scaling not only in communication disorders but also in common real-world situations often encountered by most individuals (i.e., often heard at the doctors office, "on a scale of 1-10 how bad does it hurt?"). An additional rationale for the continued use of equal appearing intervals is that direct magnitude estimation is more cumbersome to use because it requires either a standard speech passage with a preassigned value or "different modulus equalization techniques to remove listener variance" (Schiavetti et al., p. 28). It also requires relatively large numbers of stimuli when compared to the equal appearing interval scaling procedure (Schiavetti et al.), therefore, equal appearing intervals is preferred to

direct magnitude estimation when measuring speech naturalness in stuttering. Based on the above review, the recommendation to use Martin et al.'s equal appearing intervals scale when measuring speech naturalness was followed in the studies presented here. The use of this scale in stuttering speech naturalness research is discussed further below.

*The Initial Study*

Martin et al. (1984) initially used the 9-point naturalness scale because they thought speech quality measurements in stuttering were not scaled or quantified in a meaningful way and were, therefore, not suitable for use in a clinical setting. In order to quantify normal sounding speech, they believed the measure must be scalable, have adequate reliability, and have construct validity (or at the very least produce scores that could readily distinguish between obviously different types of speech quality). With this framework in mind, they developed what they found to be a more clinically useful method of assessing speech quality. Their first study using this new measure had 30 undergraduate raters listen to 1-min speech samples from people who stuttered, people who did not stutter, and people who stuttered speaking while under the influence of DAF. The listeners were asked to score the naturalness of each speech sample on a 9-point equal appearing intervals rating scale with "1" signifying *highly natural* and "9" *highly unnatural*. They found that listeners rated the speech of the people who stuttered, both with and without DAF, as significantly more unnatural than the speech of the speakers who did not stutter.

Interrater reliability was assessed using intraclass correlations (ICCs; see below for a more in depth discussion of ICCs). The listeners were found to be reliable as a group (stuttered samples and DAF .98, nonstuttered samples .75) but unreliable as individual raters (stuttered samples .74, DAF .57, and nonstuttered samples .10; Martin et al., 1984). The authors used .80 as their standard and as can be seen, for individual raters, only the stuttered samples approached

their criterion. As discussed above, this finding replicated earlier reliability findings in an area outside communication disorders (Nichols, 1966). Martin et al. also assessed rater agreement in their initial study. They found interrater agreement to be 74% for the speech samples of people who stutter, 77% for people who stutter using DAF, and 75% for normal adult speakers. Listeners' intrarater agreement was higher at 90% for people who stutter, 85% for people who stutter using DAF, and 89% for normal speakers. Although Martin et al. found relatively high agreement ratings, it is unknown whether this same level of agreement would be found when listeners rate the speech naturalness of children who stutter, an issue investigated in the two studies presented in subsequent chapters.

*The Measurement of Speech Naturalness in Stuttering Using Martin et al.'s 9-Point Scale*

Since their initial study, Martin et al.'s (1984) rating scale has been used not only to distinguish between the speech of people who stutter and people who do not stutter (Ingham, Gow, & Costello, 1985; Metz et al., 1990; Van Borsel & Eeckhout, 2008), but also to detect changes in the speech of people who stutter before and after treatment (Gow & Ingham, 1992; Ingham, Moglia, Frank, Ingham, & Cordes, 1997; Ingham & Onslow, 1985; Ingham, Warner, Byrd, & Cotton, 2006; Tasko, McClean, & Runyan, 2007), and to compare the speech of people who stutter at various severity levels (Runyan, Bell, & Prosek, 1990; Stuart & Kalinowski, 2004). This scale has also been used to quantify how natural speech *felt* to people who stutter (Finn & Ingham, 1994) and as a treatment variable to determine if people who stutter could alter their speech naturalness based on a clinician's naturalness rating fed back to them in real time (Ingham et al., 1989; Ingham, Martin, Haroldson, Onslow, & Leney, 1985). Although the scale has been used in other ways, the majority of the speech naturalness research in stuttering has focused on comparing the speech of people who stutter to the speech of people who do not

stutter. All of the evidence to date suggests that the speech of people who stutter is less natural than the speech of people who do not stutter (see Appendix C).

*Effect of speech sample presentation on speech naturalness ratings.* The effect of speech sample presentation on speech naturalness ratings has been studied by assessing how different audiovisual methods of speech sample presentation and various speech sample lengths affect speech naturalness ratings. Martin and Haroldson (1992) investigated whether speech naturalness ratings varied according to the presentation method of the speech samples, audio only versus audiovisual. Twenty-four inexperienced listeners rated the speech naturalness of 20 1-min speech samples (using Martin et al.'s 1984 scale) from 10 people who stuttered and 10 gender-, race-, and age-matched fluent speakers. Listeners rated these 20 speech samples twice; once while watching the video of the speech samples and once while only hearing the audio portion of the speech samples. Each listener saw and heard 10 speech samples and only listened to 10 different speech samples during each rating session with the order randomized between participants. The assigned order was switched for the second rating that occurred at least 2 weeks after the initial session. Between the two rating sessions each rater judged the speech naturalness of all 20 speech samples audiovisually and only auditorally. No samples were seen or heard twice during the same rating session. An analysis of variance statistical procedure (ANOVA) for mean naturalness values was conducted in which sample type (stutterer versus nonstutterer) was a between factor and sample stimulus (audio or audiovisual) was a within factor. Martin and Haroldson found significant main effects for sample stimulus and sample type as well as a significant interaction effect for sample type by sample stimulus. A Newman-Keuls analysis of pairs of individual cell means revealed that, for both sample types, the mean naturalness values were higher for people who stuttered than for people who did not stutter. The

audiovisual samples from the people who stuttered resulted in significantly higher naturalness ratings than the audio only samples, but no significant difference was found between the naturalness ratings for the two types of samples for people who did not stutter. Although the mean difference between the naturalness ratings of the two different types of speech samples from the group who stuttered was significant, the difference was only 0.76 scale points with a range from 0.21 to 1.51 points. When compared to the difference in the speech samples of people who stutter, the nonsignificant difference between the two types of presentations for the normal speech samples suggests that there are salient secondary characteristics or other unknown defining features in the speech samples of people who stutter that are not discernible when using audio only samples.

This finding of video presentation affecting research results has also been seen in other areas of stuttering research including a higher stuttering frequency from videos of preschool children who stuttered (Rousseau, Onslow, Packman, & Jones, 2008) and more favorable personality traits from live versus audio only presentation (Wenker, Wegener, & Hart, 1996). This finding was not always the case as others have not found a difference between the two presentation methods when having listeners identify stuttering frequency (Ingham, Cordes, & Finn, 1993) or severity (Williams, Wark, & Minifie, 1963). Video presentation has also been found to make a difference in areas outside of stuttering such as listeners' ratings of higher intelligibility from video versus audio presentation, both for speakers with dysarthria (Garcia & Cannito, 1996; Hustad & Cahill, 2003, for the severely dysarthric speaker only) and speakers with bilateral facial paralysis (Keintz, 2007). Male to female transgender speakers were found to have a more feminine voice after voice treatment on video when compared to audio only conditions (Baker & Pickering, 2010) and laryngectomy patients were judged more favorably

using a visual analogue scale from video versus audio only speech sample presentation (Evitts et al., 2010).

Speech sample presentation was also addressed by investigating the effect of speech sample duration on speech naturalness ratings (using Martin et al.'s 1984 scale) made by experienced and inexperienced raters (Onslow, Adams, & Ingham, 1992). Sixty judges, 30 experienced and 30 inexperienced, were placed in six subgroups each consisting of 10 experienced and 10 inexperienced listeners. Each subgroup rated the same speech samples of people who stutter at one of three different time intervals: 15 s, 30 s, and 60 s. The stutterfree monologue speech samples used in this study were taken from 10 adults and adolescents who stuttered who were enrolled in a residential prolonged speech treatment program. The main dependent variable for this study was the reliability of speech naturalness ratings (discussed in greater detail below) when different lengths of speech samples were used. Agreement for this study was defined as the percentage of speech samples that were within +/- 1 naturalness scale value of each other upon rerating (intrarater agreement) and when compared to other judges' ratings (interrater agreement). (A more in depth discussion of rater agreement is presented later in this chapter.) For intrarater agreement, sophisticated listeners had the highest agreement for the 30-s duration (75.2%); followed by 60 s (72.3%), and 15 s (65.6%); but the inexperienced listeners had the highest agreement for the 60-s duration (72.4%), followed by 30 s (69.1%), and 15 s (68.7%). For interrater agreement levels, 30 s was the least agreed upon by both groups (50.1% and 48.7% for sophisticated and inexperienced listeners, respectively), with the 60-s duration having the highest agreement (61.8% and 59.2% for sophisticated and inexperienced listeners, respectively).

Other researchers have found higher rater agreement for both trained and untrained listeners using 30-s speech samples (see Appendices D and E). Ten trained listeners had an interrater agreement level of 79% (Packman, Onslow, & van Doorn, 1994), and 2 trained listeners had interrater agreement levels ranging from 62.5% to 97.5% (Ingham et al., 1997) using 30-s speech samples. It should be noted that the higher levels in the Ingham et al. study may have been affected by the small number of raters as the likelihood of 2 raters agreeing is higher than the likelihood of 10 raters agreeing with each other (Cordes, 1994). Interrater agreement in studies using 60-s speech samples ranged from 32% (Kalinowski, Noble, Armson, & Stuart, 1994) to 94% (Martin et al., 1984) with considerable overlap between the 30 s and 60 s durations. A more in-depth discussion of reliability in speech naturalness measurement follows.

## Reliability of Speech Naturalness Ratings

The basic concept of reliability is characterized as consistency (Huck, 2000). As discussed by Cordes (1994), there are two ways to think about the concept with the most basic being that reliability refers to the general trustworthiness of data. It refers to the dependability or reproducibility of data, or the extent to which any measurement procedure yields the same results upon repeated trials (Carmines & Zeller, 1979). In other words, reliability focuses on whether the data would be reproduced if they were recollected under the same conditions (Field, 2009). The second definition of reliability explains the concept as a coefficient that is the ratio of true score variance to observed score variance, or the proportion of variance in the observed scores that can be attributed to the variance of the true score (Crocker & Algina, 1986). This second definition of reliability as a mathematical concept can be thought of as a subtype of the more general reliability described above (Cordes, 1994). Reliability has been commonly measured

throughout speech naturalness literature in stuttering in two ways: using reliability coefficients and observer agreement scores.

The terms "reliability" and "agreement" are often used interchangeably in research literature and are frequently referred to jointly as "reliability" (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Statistically, however, there is a clear distinction between the two terms (Cordes, 1994; Crocker & Algina, 1986). Rater agreement is defined as "the extent to which observers agree in their scoring of behavior" (Kazdin, 1982, p. 48). It is the degree of correspondence among the ratings or counts of the same behavior or event assigned by different judges or the same judge on two separate occasions (Cordes; Kazdin). It may be expected that this correspondence between the two ratings be exact or meet a preestablished criterion (Kearns, 1990; Tinsley & Weiss, 1975). The criterion commonly seen in speech naturalness literature in stuttering is for the two scores being compared to be within +/- 1 scale value of each other (Mackey et al., 1997; Martin et al., 1984). Rater agreement does not address reliability in the strict psychometric sense of the word involving true, observed, and error score variance, but does address whether differences in measurement can be attributed to differences among observers or among multiple observations of one observer (Cordes). Even though agreement scores do not directly address variance, high agreement is evidence that changes in the observed score are due to actual changes in the true score. In other words, high agreement provides evidence that changes in the observed score are due to changes in the subject's actual behavior rather than error variance or inconsistency on the raters' part (Cordes). In a more basic sense, agreement determines if the dependent variable can be measured with consistency.

It is important to note that assessing agreement is not the same thing as assessing accuracy. Accuracy is whether or not the observer's data reflect the actual behavior being

observed (or the client's actual performance) or is the correspondence between an observer's scores and some established criterion (Barlow & Hersen, 1984; Cordes, 1994; Kearns, 1990) "with the presumed ideal being the case where the criterion is known, absolutely and unquestionably, to be thoroughly and completely 'correct'" (Cordes & Ingham, 1999, p. 864). Unfortunately, "correct" ratings of the speech naturalness of children or adults who stutter do not exist, with the possible exception of the 24 samples included in the *SMS* training program (Ingham, Bakker, Moglia, & Kilgo, 1999; Ingham, Ingham, Moglia, & Kilgo, 2010). The standards (criterion naturalness ratings) for the 24 *SMS* speech naturalness training samples were developed using the average of the naturalness ratings of 20 reliable naïve judges (20 out of 27 undergraduate students whose intrarater agreement was within +/- 1 of their initial rating; Ingham & Ingham, 2010). Only 6 of the 24 speech samples used in the *SMS* program are from children. This small number of samples makes the investigation of accuracy, in the above sense, when measuring the speech naturalness of children who stutter, difficult at the present time.

An acceptable level of agreement is one in which the researcher is confident that the observers are sufficiently consistent in their ratings, that the behaviors are adequately defined, and that the measure will be responsive to changes in a client's performance over time (Kazdin, 1982). Specifically in the case of speech naturalness measurement, because no external definition is typically given, the question becomes whether being instructed to use one's own internal definition of speech naturalness provides a "definition" of naturalness that results in the reliable use of the speech naturalness scale. Listeners' agreement ratings in the two studies presented in this document provide data investigating this notion. No certain value can be set as to what is acceptable agreement that will extensively apply to all research although acceptable agreement levels range from 70% to 90%, with 80% identified as a "traditional" lower limit

(Baer, Wolf, & Risley, 1987; Kazdin). Some researchers used 90% agreement as their standard (Costello & Hurst, 1981; James, 1981a; 1981b; 1983), because the higher the rater agreement, the more confident one can be that changes in measurement ratings are due to changes in the behavior of interest, rather than changes in the raters' judgments. Although the 80% lower limit traditionally used seems reasonable, "predetermined criteria cannot supersede agreement requirements determined by experimental questions" (Ingham, Cordes, & Gow, 1993, p. 505).

Whereas agreement deals with the replicability of individual scores, reliability, when used in the traditional sense of a correlation coefficient, deals with the replicability of ranked scores. The reliability coefficient, typically calculated via a correlation coefficient, is designed to measure the extent to which two sets of observations covary. One problem with this method is that it is possible to obtain a correlation coefficient of "substantial magnitude" from data in which the two observers fail to agree on a single observation (Lewis, 1994, p. 271). This failure to agree can occur when one observer consistently produces inflated numbers on a measure relative to another observer, or relative to their original ratings on the measure.

Two main problems with using reliability correlations to address the reliability of a measure become evident. One is that the procedure provides no information about the replicability of one observation because the reliability coefficient is estimated by correlating sets of scores rather than individual scores (Cordes, 1994; Kearns, 1990). Because of this, the procedure is not preferred when the investigator is interested in the reliability of individual trials, as is the case in the studies presented in the pages that follow. Secondly, the correlation may be high even in the absence of agreement between two sets of scores because, as discussed above, correlations provide no information regarding the similarity of absolute scores (Cordes). For example, the ratings 1, 2, 3, 4, 5, and 21, 22, 23, 24, 25 correlate perfectly, even though they are

extremely different ratings. Accordingly, this method is not preferred when a researcher is interested in the similarity of the absolute scores of two different observers (or one observer on two separate occasions); therefore, correlation coefficients are better suited for use in classical test theory rather than in observational research.

Although providing evidence of acceptable reliability is crucial in any research study, the methods used to estimate or summarize reliability or agreement are even more pertinent when studying subjective judgments made by a rater (such as speech naturalness data recorded using the 9-point equal appearing interval rating scale), because the measurements are not based on an externally defined concept or readily observable overt behavior(s). (See above for a discussion of defining speech naturalness.) The majority of speech naturalness research in stuttering that used Martin et al.'s (1984) scale did report some type of reliability scores (all those this author is aware of are listed in Appendix F), with less than 25% of the studies reporting no reliability data for speech naturalness at all (Block, Onslow, Packman, Gray, & Dacakis, 2005; Hearne et al., 2008; Ingham, Gow, & Costello, 1985; O'Brien, Cream, Onslow, & Packman, 2001; O'Brien, Packman, Onslow, & O'Brien, 2003; Van Borsel & Eeckhout, 2008). In stuttering research, the reliability of speech naturalness ratings was most commonly assessed using intraclass correlation coefficients and agreement scores. The most commonly reported reliability calculation in speech naturalness research in stuttering was interrater agreement, which was reported in 15 studies (see Appendix D). Intraclass correlation coefficients were reported in 14 studies (see Appendix G), and intrarater agreement was reported in 13 studies (Appendix E). Both interrater agreement and intraclass correlation coefficients were reported in 7 studies (Hewat, Onslow, Packman, & O'Brian, 2006; Kalinowski et al., 1994; Martin & Haroldson, 1992; Martin et al., 1984; O'Brian, Packman, Onslow, Cream, O'Brian, & Bastock, 2003; Onslow, Adams, & Ingham, 1992;

Packman et al., 1994). Studies using each type of reliability assessment in speech naturalness research in stuttering will be discussed further below.

*Intraclass Correlation Coefficient*

The intraclass correlation coefficient (ICC) indicates the degree of correspondence between ratings provided by multiple judges (Lahey, Downey, & Saal, 1983). It is a measure of the proportion of the total variance in a measurement that is due to the variation of the stimuli being rated (Tinsley & Weiss, 1975), and it can be calculated as the variance of interest divided by the sum of the variance of interest plus the error variance (Shrout & Fleiss, 1979). This method indicates how well judges agreed that the speech samples had the same *relationship* to each other, but does not necessarily indicate how well judges agreed in terms of absolute scale values (i.e., two judges may have agreed that samples one and three were 2 scale values apart, but one person may have seen the change as 1 to 3 and the other as 7 to 9). As discussed above, this method correlates the mean speech naturalness of each rater in the group with "the average individual rater for each group" (Martin et al., 1984, p. 56). ICCs from .4 to .75 are considered to be fair, with an ICC greater than .75 considered excellent (Fleiss, 1986). If a study's goal is to assess the reliability or consistency of individual raters, this method is not ideal.

As mentioned above, ICCs have been used to assess reliability in studies of speech naturalness ratings in stuttering (see Appendix G). In these studies, the ICC would be interpreted as the proportion of the total variance in speech naturalness ratings that is due to the variance in the speech samples (or speakers) being rated. ICCs are said to be used in stuttering research over other methods of reliability because participants do not have to repeat ratings, which is thought "to avoid error problems associated with [the] memory of raters" (Metz et al., 1990, p. 519). Another way error problems may be avoided is to separate the two rating occasions rather than

having listeners rate the samples twice in the same session. Although ICCs are commonly used, as discussed above, the information they provide does not allow for conclusions to be drawn about the agreement levels of individual raters or the replicability of one observation because they deal with the relative ordering of ratings rather than absolute values (Tinsley & Weiss, 1975).

The seminal study of speech naturalness in stuttering by Martin and colleagues (1984) began the use of ICCs to calculate the reliability of speech naturalness ratings, and the technique continues to be used in current speech naturalness research in stuttering (e.g., Armson & Kiefte, 2008). The limited body of research reporting both group and individual ICC values suggests that reliability for group ICC ratings is higher than for individuals (see Appendix G). Individual ICCs were not found to be sufficiently reliable (in the sense of accounting for score variance) for routine clinical or research use in the majority of studies that reported individual ICCs (Martin & Haroldson, 1992; Martin et al., 1984; Metz et al., 1990; O'Brian, Packman et al., 2003; Onslow, Adams et al., 1992; Stuart & Kalinowski, 2004; Stuart, Kalinowski, Rastatter, Saltuklaroglu, & Dayalu, 2004; Stuart, Kalinowski, Saltuklaroglu, & Guntupalli, 2006). As seen in Appendix G, less than 80% of a rater's individual variance in naturalness scores can be attributed to speakers as ICC ratings which range from .78 (Martin & Haroldson, 1992) to .34 (Onslow, Adams et al., 1992), show.

Another issue related to ICCs is the speaker group or groups from which the ICCs are calculated. Whereas Martin et al. (1984) used only stuttering speakers in their calculation of this form of reliability, other studies reported ICC values for people who stuttered and people who did not stutter combined (Hewat et al., 2006; Metz et al., 1990; O'Brian, Packman, Onslow, & O'Brian, 2003; Schiavetti et al., 1994; Stuart & Kalinowski, 2004; Stuart et al., 2006). This

decreases the sensitivity of the measure by affecting who the results can be generalized to, and the combination of the two speaker groups makes it impossible to determine which specific speaker group accounts for how much of the variance in speech naturalness. Because the group is less homogeneous when both people who stutter and people who do not stutter are included, the reliability coefficient for this group is expected to be inflated when compared to a more homogeneous group. This makes sense because it is easier to identify speech naturalness differences between more extreme speech samples (i.e., normal speakers versus persons with severe stuttering) and thereby the reliability of these ratings would be higher than it would when raters are asked to identify differences in the speech naturalness of more similar speech samples (i.e., various speech samples of mild and moderate stuttering only) and thereby the reliability of these ratings would likely be lower. Assuming the speech of people who stutter sounds less natural to listeners than the speech of people who do not stutter (as has been consistently found in speech naturalness research in stuttering and is reported in Appendix C), higher reliability coefficients would be expected for a combined group of people who do and do not stutter than if the group consisted of all people who stutter (or people who do not stutter) of equal size (assuming random error variance is constant for the two groups), as classical test theory asserts (Crocker & Algina, 1986).

An additional basic issue with the use of ICCs in speech naturalness research in stuttering is the researchers' tendency not to report which type of ICC procedure they used. This is important because the error variance definition depends on the reliability method chosen (Morrow & Jackson, 1993). Six types of ICC procedures exist (McGraw & Wong, 1996; Shrout & Fleiss, 1979) each of which reports reliability for either a single rater or a group of raters using one of three possible statistical models (Huck, 2000). Some studies did not report the type of

ICC used, the exceptions being Hewat et al. (2006), O'Brien, Packman, Onslow, Cream et al. (2003), Onslow, Adams, & Ingham (1992), Packman et al. (1994), Stuart and Kalinowski (2004), Stuart et al., (2004), Stuart et al., (2006), and Tasko et al. (2007). Small sample sizes were another reliability problem when ICCs were used. The smaller the sample size the more the results should be interpreted with caution due to the potential variability from sample to sample.

According to Kazdin (1982), the evaluation of observer agreement is central to the collection of observational data for three primary reasons. First, assessment of any construct is only useful if it can be measured with some consistency. If counts differ based on who is counting, then it would be difficult to impossible to identify the actual behavior of interest. A second reason agreement between raters is necessary, is to minimize the impact of an individual rater's biases. A listener may record the behaviors differently over time or across samples based on his or her unstable perceptions rather than the actual behaviors. Finally, agreement is vital because it reflects whether the behavior being measured is well defined. Well defined constructs typically have higher agreement levels between raters. As discussed earlier, this study will investigate the listeners' use of their internal definition of speech naturalness via rater agreement. Two types of agreement scores are commonly used as a measure of reliability: interrater and intrarater. Both of these methods have been commonly used in stuttering speech naturalness literature (see Appendices D and E) and are discussed below.

*Interrater Agreement*

As mentioned above, interrater agreement calculation is an alternative to ICCs. One method of assessing interrater agreement when using rating scales compares each rater 's observation to every other rater's observation in the group. This judgment consistency can be calculated using difference scores by subtracting the score of one rater from the corresponding

scores of each of the other raters in his or her group. Once these difference scores are derived, agreement levels can be reported as the percentage of rater comparisons that were within +/- 0, 1, 2, 3, 4, 5, 6, 7, or 8 rating scale points as is often done in the speech naturalness literature in stuttering. The fewer scale points between the listeners' two scores, the better they are said to agree. Traditionally in speech naturalness research in stuttering, the standard criterion for agreement scores is that the ratings be within +/- 1 scale value of each other or better (Finn, Ingham, Ambrose, & Yairi, 1997; Martin et al., 1984); therefore, interrater agreement data are typically summarized as the percentage of rater comparisons between two raters that were within +/- 1 scale value of each other or better.

The interrater agreement of speech naturalness ratings of the speech samples from people who stuttered, as rated by listeners who did not stutter, were derived using the difference score procedure described above and ranged from 97.5% (for 1 participant in the study; Ingham et al., 1997) to 32% (Kalinowski et al., 1994; see Appendix D). Martin et al. (1984), in the original naturalness study, found 74% of raters were within +/- 1 scale value of each other. Ingham et al.'s other 3 subjects had agreement levels of 87.5%, 80% and 62.5%, while values of 80% and 79% (Martin & Haroldson, 1992; Packman et al., 1994) as well as 62% (Onslow, Costa, Andrews, Harrison, & Packman, 1996) and 59% (Mackey et al., 1997) were also found. When the speech naturalness of people who stuttered at different severity levels was assessed before and after treatment, Kalinowski et al. found higher agreement levels for mild versus severe stutterers (67% pretreatment and 59% posttreatment, mild; 60% pretreatment and 32% posttreatment, severe) and for pretreatment versus posttreatment naturalness ratings (67% and 60% mild versus 60% and 32% severe for pretreatment versus posttreatment, respectively). Because 80% is considered the traditional lower limit of acceptable agreement (Kazdin, 1982),

one can see that the majority of these studies do not meet this criterion (see Appendix D). Even when using the more lenient standard of 70%, raters would still not meet the agreement level for one of Ingham et al.'s (1997) participants or in Kalinowski et al.'s (1994), Mackey et al.'s, or Onslow et al.'s studies.

The interrater agreement of speech naturalness ratings of the speech samples from people who did not stutter, as rated by listeners who did not stutter (in studies where they were also rating the speech of people who stutter), were also derived using the difference score procedure described above and ranged from 56% for people who did not stutter but had a dialectical difference (Mackey et al., 1997) to 81% (Martin & Haroldson, 1992 using audiovisual samples). The lowest agreement rating for people who did not stutter without dialectical differences was 75% (Martin et al., 1984). Studies that reported the interrater agreement of listeners' speech naturalness ratings of people who stutter and people who do not stutter combined ranged from 64.7% (Finn, 1997; in this case it was people who used to stutter and people who did not stutter) to 87.2% (Ingham, Warner, et al., 2006). As can be seen, the interrater agreement for the adults who stuttered had the largest range (32-97.5), with the range of normally speaking adults much smaller (56-81) and the combined groups having the smallest range (64.7-87.2). This supports the classical test theory assumption discussed above that the more homogeneous the samples rated the lower the reliability (Crocker & Algina, 1986).

There is also a possibility that the agreement levels between listeners were affected by how many listeners were used to rate speech naturalness. Higher reliability is expected from studies that use a higher number of observers (Cordes, 1994; Crocker & Algina, 1986). As explained by Cordes, this principle suggests that it is more likely that any 10 observers will reproduce the mean or total score of any other 10 observers, than for 1 observer to reproduce a

score by another observer. Most of the studies assessing the interrater agreement of the speech samples of people who stutter used 24 or more raters, with only one study using 10 (Packman et al., 1994) and two studies using 2 (Ingham et al. 1997; Onslow et al., 1996). Overall, the studies that had more listeners rate speech naturalness tended to have higher agreement levels, but even so, only one showed levels of 80% or higher (Martin & Haroldson, 1992). Although the majority of speech naturalness research with adults reported interrater agreement ratings less than 80% (see Appendix D), it is not known if the interrater agreement for speech naturalness ratings in children who stutter will correspond to these adult values.

*Intrarater Agreement*

Whereas interrater agreement deals with the consistency raters have with other raters, intrarater agreement assesses the consistency of individual raters with themselves, across time (Maxwell & Satake, 2006). This consistency measure is also calculated in speech naturalness research in stuttering using difference scores. These scores are derived by subtracting each rater's score for each speech sample on rating occasion one to their corresponding score for that same speech sample on another rating occasion. As is the case with interrater agreement, the intrarater agreement criterion that is typically used in speech naturalness research in stuttering is that the rater's two scores be within +/- 1 scale value of each other or better. This information was typically presented as the percentage of raters whose two scores met this criterion, although one study presented the information as 67% of the listeners had at least 75% of their ratings within 1 scale value of each other (Onslow, Hayes, Hutchins, & Newman, 1992; Appendix E).

As Appendix E shows, the studies reporting the intrarater agreement levels (the percentage of raters whose ratings on two occasions varied by +/- 1 scale value or less) for speech samples of people who stutter ranged from 47% (Martin & Haroldson, 1992) to 94.3%

(Onslow et al., 1996). The 47% was a study that used audio only speech samples, which may have had an effect on agreement ratings (Martin & Haroldson). As discussed above, video samples seem to provide information regarding the secondary characteristics or other aspects of the speech of people who stutter that may help with observational/rating agreements. This conclusion is supported by Martin and Haroldson's study in which raters who watched the audiovisual speech samples had a significant increase in agreement when compared to the same listeners' ratings using the audio only samples. Other intrarater agreement levels reported for the speech samples of people who stutter were 76% (Mackey et al., 1997) and 90% (Martin et al., 1984). When looking at the intrarater agreement of mild and severe stutterers pre- and posttherapy, the agreement ratings were comparable. Mild pretreatment and severe posttreatment were the highest at 89%, while mild posttreatment was 86% and severe pretreatment was the lowest at 83% (Kalinowski et al. 1994).

The intrarater agreement of speech naturalness ratings of the speech samples from people who did not stutter, as rated by listeners who did not stutter (in studies where they were also rating the speech of people who stutter), were also derived using the difference score procedure described above and ranged from 80% to 98% (see Appendix E). (Both were Mackey et al., 1997 and the 80% was for people who did not stutter but exhibited a dialectical difference in their speech.) The lowest value for normal speakers was 89% (Martin et al., 1984). Studies that reported the intrarater agreement of listeners' speech naturalness ratings for speech samples from people who stutter and people who do not stutter combined ranged from 70% (Stuart & Kalinowski, 2004 combining normal speakers, people who stutter, and people who stutter under DAF) to 100% (O'Brian, Onslow, Cream, & Packman, 2003). Studies assessing intrarater agreement of people who stutter only again had the largest range (47-94.3), with normal speakers

having the smallest range (80-98) and the groups combined in between (70-100). Unlike the findings for interrater agreement, here the combined group did not have the smallest range, but this could be influenced by the fact that both of the extremes for the combined group came from studies that had raters rerate the samples during the same session (70% Stuart & Kalinowski; 100% O'Brian, Onslow et al.) versus all of the other studies that had listeners rerate the samples at least a week after the initial ratings. It should also be noted the O'Brian and colleagues only had listeners rerate 16.6% of the samples from people who stutter and 5.5% from the normal speakers. As discussed above, the fewer the ratings, the higher the likelihood for increased agreement scores.

Overall, raters had better intrarater agreement than interrater agreement for speech naturalness. This trend has also been seen in research having listeners identify stuttering events in adults' speech (Cordes & Ingham, 1994b; 1995; 1999; Ingham, Cordes, & Gow, 1993), but the two have also been found to be relatively equal when listeners identify stuttering using binary judgment procedures (Bothe, 2008; Cordes, 2000). Bothe's study is noteworthy, as it was the only one that evaluated the speech of children who stuttered. When having expert judges measure the presence or absence of stuttering in 5-s speech samples, she found intrarater agreement levels of 89.8% to 97.3% and interrater agreement levels of 87% to 89%. These are within the range of acceptable agreement levels for the identification of stuttering in adult speech samples. As the agreement values when identifying stuttering are comparable for adults and children, a needed area of research is to investigate if this trend holds true for agreement levels of speech naturalness ratings for adults and children.

*The effect of time intervals between ratings on intrarater agreement.* Timing of reratings is a crucial factor to consider when addressing intrarater agreement as it may have an impact on

the internal validity of a study. In speech naturalness studies in stuttering, reratings were reported to occur either twice during the same session (Coughlin-Woods, Lehman, & Cooke, 2005; Martin & Haroldson, 1992; O'Brian, Onslow et al., 2003; Onslow, Haynes et al., 1992; Stuart & Kalinowski, 2004) or 1 week (Finn, 1997; Finn et al., 1997; Kalinowski et al., 1994; Onslow et al., 1996), 1 to 3 weeks (Finn & Ingham, 1994; Mackey et al., 1997; Martin et al., 1984), or 3 weeks (Onslow, Adams, & Ingham, 1992) after the initial rating. The intrarater agreement of the studies having listeners rerate samples in the same session ranged from 47% (Martin & Haroldson) to 100% (O'Brian, Onslow et al.). It is necessary to note that the agreement values reported by O'Brien, Onslow et al. were ratings of the speech samples of people who stuttered and people who did not stutter combined which may have inflated the reliability estimate, as discussed above. This study cannot be reasonably compared to studies reporting the agreement of speech samples from people who stutter only because the speech samples were taken from two distinct populations. The highest intrarater agreement in a study that reported the agreement of speech naturalness in the speech samples of people who stuttered was 62% (Martin & Haroldson). Listeners with 47% agreement rated audio samples and were 15% lower than the lowest audiovisual speech samples at 62% (Martin & Haroldson). The lack of a break between rating sessions could introduce a practice effect, meaning the intrarater agreement may have been affected by the listeners' familiarity, practice, or learning due to the short time between the repeated exposure to the same stimulus (Maxwell & Satake, 2006).

Studies less likely to suffer from practice effects separated rating sessions by 1 to 3 weeks. The studies that used rerating times of 1 week had intrarater agreement values from 83% (Kalinowski et al., 1994) to 94.3% (Onslow et al., 1996). Both this study and O'Brian, Onslow et al.'s (2003) study cited above used only 2 or 3 raters, which they have argued may have a

negative effect on reliability scores (O'Brien, Packman, Onslow, & O'Brien, 2003). Using more judges does not eliminate disagreements among judges, but it does decrease the likelihood that "spurious experimental conclusions will be drawn from them" (Cordes, 1994, p. 273). Both of the other studies that had a 1-week break between rating sessions reported intrarater agreement scores of 83% or higher and used at least 10 raters (Finn, 1997; Kalinowski et al.). Two studies allowed 1 to 3 weeks between their ratings and both had intrarater agreement levels of 76% (Mackey et al., 1997; Onslow, Hayes et al., 1992). Like O'Brien, Packman, Onslow, Cream et al. (2003), Onslow, Hayes et al. reported agreement scores for people who stuttered and people who did not stutter together; therefore, these results could only be generalized to this combined group. The study allowing the longest delay between ratings (3 weeks) reported intrarater agreement ratings between 65.6% and 75.2% (Onslow, Adams et al., 1992).

The limited body of research reviewed above suggests that ratings separated by 1 week had the strongest intrarater agreement because the listeners in all three studies agreed 83% of the time or more. When the literature reporting intrarater agreement in the speech naturalness of adults is compared, reratings during the same session versus one week, the values overlap. The same session ratings start lower than ratings 1 week or more (46.7% versus 65.6%) and have a larger range than ratings separated by at least a week (46.7%-100% versus 65.6%-94.3%, respectively). Comparisons between the two time frames should be interpreted with caution because no one study compared the two rerating times; and the studies reported vary in the number of raters used, number and type of speech samples used, and the speech sample lengths used (see Appendix E).

*The effect of number of samples rerated on intrarater agreement.* The number of samples researchers chose to have their listeners repeat has been shown to affect intrarater

agreement levels. Oftentimes, researchers choose to only have a portion of their data rerated rather than using the more rigorous method of reliability analysis in which all of the data are rerated. Rerating only a portion of the data may result in reliability data that are based on a potentially unrepresentative data sample (Cordes, 1994). In other words, the same reliability values may not have been found had all of the data been rerated or had a different portion of the data been rerated. To combat this issue, it is recommended that the entire task be repeated on a separate occasion (Onslow & Ingham, 1987).

Speech naturalness studies in stuttering either had listeners rerate 100% of their samples or very few samples. The majority of the studies followed Onslow and Ingham's (1987) stringent rerate criteria and had listeners rerate all of the samples (Finn, 1997; Finn et al., 1997; Finn & Ingham, 1994; Kalinowski et al, 1994; Mackey et al., 1997; Martin et al, 1984; Onslow, Adams et al, 1992; Stuart & Kalinowski, 2004), with all except Kalinowski et al. also separating the two sessions by at least 1 week. The number of samples rerated was not always reported (Onslow et al., 1996). The remaining studies had listeners rerate 25% of the samples or less (Martin & Haroldson, 1992; O'Brien, Onslow et al., 2003; Onslow, Hayes et al., 1992 study one) with the lowest being 14.2% (Onslow, Hayes et al., study one). Studies with higher rerating percentages are likely to have higher reliability than studies that had listeners rerate only a few speech samples because reliability is increased by the number of observations that contribute to the reliability score (Cordes, 1994; Crocker & Algina, 1986).

In the speech naturalness measurement literature with adults who stutter, this was true. Studies that had listeners rerate 100% of their data had intrarater agreement scores ranging from 65.6% (Onslow, Adams et al., 1992) to 90% (Martin et al., 1984), while studies reporting rerate percentages of 25% or less ranged from 47% to 62% (audio and audiovisual from Martin &

Haroldson, 1992, respectively). One study did report intrarater agreement higher than 62%, but reported it differently than the other studies. In the above referenced studies, the authors referred to the percentage of raters that had rerating scores within one scale value or less of their initial score. Onslow, Hayes et al. (1992) had listeners rerate only 22.2% of the original samples and from this, reported that 67% of their listeners had 75% or more of their ratings within +/- 1 scale value of the original ratings. This criterion (75% of the ratings within +/- 1 scale value) was used as a standard the listeners had to meet to have their data included in the study. Only 10 of their 15 listeners met this criterion.

*The Effects of Listener Experience on Rater Reliability*

As discussed above, the most commonly used paradigm in speech naturalness research is to have the listener assess the speaker's speech naturalness. The "listener" is anyone with whom the speaker comes into contact, the majority of whom are inexperienced, untrained listeners. It has been convincingly argued (and the current author would agree) that inexperienced listeners are more representative of the majority of the people with whom a person who stutters will converse and interact and who will ultimately judge the quality and acceptability of the speech of the person who stutters (Runyan & Adams, 1979). Because the reactions and impressions of inexperienced listeners comprise the majority of interactions for a person who stutters, some have argued that their reactions and impressions should be of prime concern (Coughlin-Woods et al., 2005; Runyan & Adams). Franken et al. (1995) state this premise very succinctly:

> In our experiment "naïve" judges were used because we feel that their judgments should be the reference. What counts most for a stutterer when he or she assesses the result of a treatment are the reactions and opinions of the persons with whom he or she must communicate in normal daily life-not the judgments of experts. (p. 282)

It can also be argued that trained speech-language pathologists may judge the speech of people who stutter too stringently because trained therapists may notice idiosyncrasies that may be undetectable to the average, untrained listener. Data supporting this idea were reported when the speech naturalness of recovered children who stuttered was measured by trained speech-language pathologists and inexperienced listeners. Trained speech-language pathologists rated the speech naturalness of recovered children who formerly stuttered 3.71, whereas inexperienced listeners' ratings were 1.77 (Finn et al., 1997). Therefore, from a speech quality or speech naturalness standpoint, untrained listener ratings appear to be more beneficial than speech-language pathologists because these are the people with whom people who stutter have the majority of contact. Both Curlee (1993) and Onslow et al., (1992) assert that the most valid measures of stuttering disability (in this case speech naturalness) are based on perceptual judgments of reliable observers; therefore, data needed to be collected to determine if naïve, inexperienced non-speech language pathologist listeners could reliably measure speech naturalness in the speech of children and if training affects their reliability ratings.

The effect of listener experience has not been widely studied in stuttering and results have been mixed in other areas of stuttering measurement regarding the effect experience with stuttering has on raters. Experienced listeners have been found to have higher reliability and agreement scores than inexperienced listeners when identifying stuttering events (Cordes, Ingham, Frank, & Ingham, 1992), but experience has also been shown not to make a significant difference in the measurement of other stuttering tasks (Cordes & Ingham, 1994a; Curlee, 1993). Like stuttering in general, the effect of listeners' experience on speech naturalness ratings has not been widely studied, with only one study directly comparing the reliability of "sophisticated" and "unsophisticated" listeners' speech naturalness ratings. Onslow, Adams, & Ingham, 1992

studied sophisticated listeners who were speech-language pathology students in their final 6 months of training, and inexperienced listeners who were nonstudents and students in their first year of an undergraduate program who had not taken a speech-language disorder course or clinical practicum. The authors did not address the inexperienced listeners' personal exposure to stuttering; therefore, the two groups could be exactly alike in terms of their exposure to/familiarity with people who stutter. Sixty subjects, 30 inexperienced and 30 sophisticated, participated in the study. Inter- and intrarater agreement scores showed sophisticated listeners performed no better in terms of the percentage of raters within 1 rating scale value of themselves or each other than the inexperienced listeners.

Inexperienced listeners were the most commonly used raters in speech naturalness research in stuttering, with half of the studies providing some sort of information stating that the listeners did not have any experience or coursework with stuttering (Block et al., 2005; Finn et al., 1997; Kalinowski et al., 1994; Martin & Haroldson, 1992; O'Brian, Packman, Onslow, Cream et al., 2003; Onslow, Adams, & Ingham, 1992; Onslow, Hayes et al., 1992; Schiavetti et al., 1994; Stuart & Kalinowski, 2004; Stuart et al., 2004; 2006) and half not addressing whether the inexperienced listeners had any experience with stuttering (Armson & Kiefte, 2008; Coughlin-Woods et al., 2005; Hewat et al., 2006; Ingham et al., 1985; Ingham, Sato, Finn, & Belknap, 2001; Mackey et al., 1997; Martin et al., 1984; Metz et al., 1990; O'Brian, Onslow et al., 2003; O'Brian Packman, Onslow, O'Brian, 2003; Van Borsel & Eeckhout, 2008). Interrater agreement ranged from a low of 32% (Kalinowski et al., 1994) to a high of 94% (Martin et al., 1984). Intrarater agreement ranged from 46.7% (Martin & Haroldson, 1992) to 100% (O'Brien, Onslow et al., 2003) with ICC values ranging from .34 (Onslow, Adams, & Ingham, 1992) to .99 (Kalinowski et al.) for group ratings. Although smaller in number, the studies that did use

sophisticated listeners (Finn, 1997; Finn & Ingham, 1994; Finn et al., 1997; Ingham et al., 1997; 2006; Ingham & Riley, 1998; Metz et al., 1990; Onslow, Adams, & Ingham, 1992; Onslow et al., 1996; Packman et al., 1994; Runyan et al., 1990; Tasko et al., 2007; Van Borsel & Eeckhout, 2008) had an interrater agreement range of 50.1% (Onslow, Adams et al.) to 79% (Packman et al., 1994), intrarater agreement of 65.6% (Onslow, Adams, & Ingham) to 94.3% (Onslow et al., 1996), and an ICC range of .43 to .64 (both of which were Onslow, Adams, & Ingham). It is important to note that of the studies listed above that used experienced listeners, only four used one or two raters to collect the speech naturalness data (Ingham et al., 1997; 2006; Ingham & Riley, 1998; Onslow et al., 1996). Experienced listeners do show a smaller range of performance, but there is considerable overlap between the two groups of listeners on all three reliability metrics and that, combined with Onslow, Adams, and Ingham's and Finn et al.'s results, does not provide significant evidence of a major difference related to experience. Although experience is valuable for many reasons, it does not seem apparent that it necessarily increases agreement levels as assessed not only in stuttering (Cordes & Ingham, 1995), but also in voice disorders (Gerratt, Kreiman, Antonanzas-Barrosa, & Berke, 1993; Kreiman, Gerratt, & Precoda, 1990; Kreiman, Gerratt, Precoda, & Berke, 1992), and dysarthria (Zyski & Weisiger, 1987). The above review, combined with the framework that listeners in a speaker's environment judge what is natural, is the basis for untrained listeners being used in the studies presented in these pages.

The literature review above highlights some of the reliability issues regarding speech naturalness ratings using speech samples of people who stutter. What is apparent from this review is that the reliability of speech naturalness measurement in stuttering is currently not at an acceptable level for use with individual listeners and often does not meet the traditional

reliability standard of 80%. Because listeners' agreement levels, especially interrater agreement levels, do not meet the standard criterion used in speech naturalness of 80%, a method to increase listener reliability warrants investigation. One way to increase listeners' reliability ratings when measuring the speech naturalness of people who stutter is to train them to measure the concept. A discussion of training throughout communication disorders leads to a discussion of training specifically in stuttering below.

### The Effect on Reliability Scores of Training Judges

Throughout various areas of communication disorders, including stuttering frequency measurement (Cordes & Ingham, 1999; Costello & Hurst, 1981; Ingham, Cordes, & Gow, 1993; Yaruss, 1998), active, structured training has been advocated, sometimes using external reference samples rather than solely depending on judges' internal standard (Gooch, Hardin-Jones, Chapman, Trost-Cardamone, & Sussman, 2001; Keuning, Wieneke, & Dejonckere, 1999; Young, 1969b) or relying on repeated exposure to the same stimuli to train raters (Young, 1969a). Training programs have been reported throughout communication disorders including for the measurement of stuttering frequency, but detailed explanations enabling replication of the training programs are not always given. Articulation is one such area that regularly reports training listeners, but the training is repeatedly reported as a passing reference with exceedingly little supporting details given. For example, authors have reported that raters completed "a series of four training sessions designed to increase the average intergroup agreement" (Schissel & Flournoy, 1978, p. 460) or "four hours of training, which consisted primarily of practice in scoring taped and live articulation tests" (Siegel, 1962, p. 31). Other examples are that listeners were presented with speech samples exhibiting the end points and midpoint of a subjective severity scale (Gordon-Brannan & Hodson, 2000) or were given descriptions of the

"physiological and articulatory characteristics of compensatory articulations" and were given examples of each (Gooch et al., 2001, p. 66). Training reported in this manner makes replication extremely difficult.

Although this lack of detail in reporting training procedures is the more common reporting method, detailed training programs in communication disorders have been developed. One example is the specialized audiotape training program developed by Shriberg and his colleagues that is readily available for purchase (Shriberg & Kent, 1995). This training technique employs training tapes that are scored by "experts" (Shriberg, 1972), and it has been used to train listeners in phonetic transcription research by using a consensus method to determine the accuracy of a transcription (Shriberg, Kwiatkowski, & Hoffman, 1984). The availability of this technique makes its use in future studies of speech sound production transcription agreement more easily accomplished. Similar training studies have been conducted in stuttering measurement and are discussed below.

*Training in the Measurement of Stuttering Outcome Variables*

Specific training to measure variables in stuttering has been recommended throughout the past three decades (Cordes & Ingham, 1994a; Costello & Hurst, 1981; Ingham, 1993; Ingham & Cordes, 1997; Ingham, Cordes, & Finn, 1993; Yaruss, 1998). To date, the majority of the research on the effect of training on stuttering measurement has been conducted in the identification of the presence of stuttering in speech samples divided into 5-s intervals. Although no group research has been conducted to assess methods that might improve judges' agreement levels when measuring speech naturalness in the speech of people who stutter, one study reported the beneficial effects of a specific training program with two clinicians' measurements of stuttering frequency, speech rate, and speech naturalness (Ingham & Riley, 1998). Ingham

and Riley successfully trained these raters using a precursor to the training program used in the current studies (Fowler & Ingham, 1987). Training research in stuttering is discussed in greater detail below.

Costello and Hurst (1981) published the first report of an increase in interjudge agreement for stuttering identification after judges were trained to identify stuttering. Prior to conducting their research, they trained two independent judges to achieve 90% event-by-event interjudge agreement for the different types of stuttering events they recorded during their study. At the completion of the experiment, mean interjudge agreement between each independent judge and the experimenter was 92.42% and 91.63%. This was the first study to suggest that not only could raters achieve adequate interrater agreement levels, but they could also be trained to achieve these levels. Speech naturalness was not measured in this study. The agreement levels reported were for the identification of the presence or absence of stuttering. The *Stuttering Measurement System (SMS;* Ingham & Ingham, 1987; Ingham et al., 1999; Ingham et al., 2010), a formal training program designed to train clinicians to reliably identify stutters, syllables for speech rate calculations, and to rate speech naturalness in real time, was developed from this line of research.

Even though no group research has been conducted on the reliability of the measurement of speech naturalness, a standardized method has been developed to train judges to rate the construct. This speech naturalness training has been developed and made available to the public as part of the larger *SMS* program (Ingham & Ingham, 1987; Ingham et al., 1999; 2010). The *SMS* is designed to train raters simultaneously to identify syllables spoken, occasions of stuttering, and to rate speech naturalness, all in real time. Because this study's focus is speech naturalness, only that portion of the *SMS* training program was used. The *SMS* trains listeners to

rate speech naturalness by using the speech samples of 3 nonstuttering speakers, 3 stuttering speakers and 10 speakers with "varied naturalness" (Ingham & Ingham, 2010, p. 78). Four of these samples are 2 min in length and the other 12 are 1 min in length. Naturalness is rated every 60 s with the naturalness ratings averaged automatically by the computer program for the longer speech samples. Five of the 16 speakers in the *SMS* training program are children. According to the *SMS Training Manual* and as discussed above, in the "Reliability of Speech Naturalness Ratings" section, the naturalness standards for this program were determined by having 27 naïve undergraduate students rate the speech naturalness of all of the speech samples included in the speech naturalness training portion of the *SMS*, each divided into 60-s segments. Each judge independently rated all of the speech samples twice, 1 to 3 weeks apart, using the 9-point rating scale and the same instructions given to judges in the initial speech naturalness study (Martin et al., 1984). "Only ratings from judges whose first and second ratings were reliable (within +/- 1 rating unit) were included in the final calculations (20 of the 27 judges; 10 males and 10 females"; Ingham & Ingham, 2010, p. 76). These standards represent the +/- 1 scale value range around the averaged ratings of these 20 reliable raters for each speech sample (Ingham & Ingham, 2010).

As part of this program, raters listen to the first speech sample and rate the speech naturalness of the sample using the Martin et al. (1984) 9-point scale. After they rate the sample the first time, the naturalness standard (as previously determined-see above) for that sample is given to the listener in writing. If the listener's rating is not within the target naturalness range, he or she is told to "listen to that sample a second time keeping in mind the naturalness rating assigned to that sample on the data sheet. This should serve as an exemplar of that particular rating scale value" (Ingham & Ingham, 2010, p. 76). Once the listener's speech naturalness

rating is within the target range for the first speech sample, the listener is to continue to the next sample and repeat this process until all 16 speech samples are completed and the speech naturalness ratings for all of them are within the target range presented in the program. The *SMS* program also includes an eight sample criterion test to assess listeners' speech naturalness ratings posttraining. Each of these eight samples, seven adults and one child, are 3 min in length. As in the training portion of the program, speech naturalness is rated every 60 s and the three 60-s samples are averaged automatically by the computer program for each sample's final naturalness rating.

While to date a formal training study using the *SMS* has not been published, the program has been used to train raters to rate the speech of adults who stutter as part of larger studies (Fox, Ingham, Ingham, Zamarripa, Xiong, & Lancaster, 2000; Ingham, Fox, Ingham, Xiong, Zamarripa, Hardies et al., 2004; Ingham, Kilgo, Ingham, Moglia, Belknap, & Sanchez, 2001). As mentioned above, one study (Ingham & Riley, 1998) successfully trained two clinicians using a previous version of the *SMS* called the *Stuttering Treatment Rating Recorder* software (Fowler & Ingham, 1987) to rate preschool children's speech. During training, the trainee had to obtain total stutter counts that were within +/- 5% of the standardized data for all 48 samples in the training module. After successful completion of the training program, each trainee had to achieve self-agreement of 80% or higher before collecting data in the study (Ingham & Riley). After meeting these criteria, interrater agreement for the speech naturalness of 2 preschool children who stuttered was reported as 73% for subject 1 (who did not receive any treatment) and 90% during baseline, 96% during treatment, 74% during withdrawal, for a combined average of 85% agreement for subject 2. Naturalness agreement was defined as the two observers' naturalness ratings being within 1 rating unit of each other.

Ingham and colleagues built on the training research base discussed above by developing a standardized stuttering measurement system entitled *Stuttering Measurement Assessment and Training (SMAAT;* Ingham, Cordes, Kilgo, & Moglia, 1998). Unlike the *SMS*, the *SMAAT* uses time interval measurement to measure the presence (or absence) of stuttering. Time interval measurement differs from traditional stuttering frequency measurement because it measures the occurrence of a target behavior by recording whether or not the behavior occurs during predetermined time intervals using real time stuttering event judgments (Ingham, Cordes, & Finn, 1993; Ingham, Cordes, & Gow, 1993). It was introduced to stuttering research by Ingham, Cordes, and Gow and as the more current version by Cordes and Ingham (1994a) to improve the reliability of stuttering judgments. This training has been shown to increase rater agreement (Cordes & Ingham, 1996; 1999; Ingham, Cordes, & Gow) and also to increase rater accuracy in regard to individual moments of stuttering (Cordes & Ingham, 1994b; 1999; Ingham, Cordes, & Gow). These increases were shown to generalize as accuracy and agreement increased not only for the speakers' samples used as training samples but also for speech samples of speakers not used in training (Cordes & Ingham, 1996; 1999; Ingham, Cordes, & Gow). Stuttered intervals tended to have higher agreement levels than nonstuttered intervals or intervals that were disagreed upon by the stuttering experts who helped develop the program (Cordes & Ingham, 1996). When evaluating each study, the extent of judges' improvement in accuracy and/or agreement is typically small, but when looking at all of the studies together, small but consistent improvements have been demonstrated with several groups of judges in several studies (Cordes & Ingham, 1994b; 1996; 1999; Ingham, Cordes, & Gow). This series of studies provides evidence that judges can be trained to produce accurate and reliable real time judgments of the presence or absence of stuttering in intervals of spontaneous speech of adults who stutter.

Even though all of the studies using the *SMAAT* program reviewed above used adult speech samples, one study has modified the *SMAAT* for children by using speech samples of Icelandic preschool children who stuttered (Einarsdottir & Ingham, 2008). The authors had 20 preschool teachers from Iceland, separated into two groups, judge the presence or absence of stuttering in 9 prejudged, 5-s intervals from connected Icelandic speech on two separate occasions, 2 to 3 weeks apart. An interesting finding was that both groups of teachers had accuracy ratings of 80% or higher in identifying stuttering on the first occasion (before any training). Although pretraining accuracy was at least 80%, the authors still trained one of the teacher groups using an Icelandic version of the *SMAAT* in an effort to determine its effectiveness. They found training to be associated with significantly higher accuracy ratings for the trained group than for the control group on the second rating occasion (control group 81.3% on occasion 1 and 82.4% on occasion 2, whereas the experimental group had 84.2% on occasion 1 to 88.6% on occasion 2). Although the authors found a statistically significant difference between the two groups on occasion two, the merit of this finding can be questioned because both groups started with accuracy levels greater than 80%. Because of the high starting accuracy levels, the need to train the teachers to identify stuttering is uncertain.

As discussed above, the research literature does not provide a study assessing rater agreement of the speech naturalness of children who stutter using a large group of listeners or with children who currently stutter: therefore, based on this literature review, it is not known if listeners' agreement levels when rating the speech naturalness of children who stutter will be low enough to justify the investigation of a training program. This issue was addressed in Study One to provide agreement levels when using the 1-9 scale to rate the speech naturalness of children who stutter. Based on the results of Study One, listener training appeared warranted (these

results are presented in Chapter 3), and this conclusion was investigated in Study Two under more controlled conditions than were used in Study One to provide further data on the topic. The speech naturalness portion of the *SMS* program was also investigated in both studies.

Because the research reported using the *SMS* program is limited, it is not known if this program can adequately train listeners to reliably rate the speech naturalness of children who stutter. Data are needed to determine if this already developed program can adequately train listeners to assess the speech naturalness of people who stutter, as was found by Ingham and Riley (1998), or if another training system needs to be developed to train listeners. Because the standardized *SMS* training program is already readily available via the internet (http://www.coe.uga.edu/csse/sms/), the research presented in this document provides evidence regarding the program's ability to train raters to reliably measure speech naturalness by investigating its effects on listeners' agreement levels when rating the speech naturalness of children who do and do not stutter.

One issue warranting comment is the difference between the speech samples included in the *SMS* training program and criterion test and the experimental stimuli used in the studies presented in the remainder of this document. Although training judges with a wide range of stimuli from various speech samples and types of speakers should produce greater generalization when compared to training with a more restricted range of intervals, it remains to be seen if the *SMS* provided sufficient exemplars of children to generalize to the experimental speech samples of children who stutter (Cordes, 1993; Stokes & Baer, 1977; Stokes & Osnes, 1989). As mentioned above, the naturalness portion and criterion test of the *SMS* uses speech samples from 18 adult speakers and 6 speech samples from children. Whether or not listeners could generalize speech naturalness rating skills obtained from a training program in which the majority of the

speech samples are from adult speakers, to the speech of children, deserved investigation and therefore, was a key issue investigated in Study One. Results from both studies will be presented in subsequent chapters after a discussion of the speech naturalness literature pertaining to children.

## Speech Naturalness Measurement in Children

Although advocated by numerous authors (Conture & Guitar, 1993; Ingham & Riley, 1998; Onslow, Costa, & Rue, 1990), there has been limited research involving the speech naturalness of children. Normally fluent children (Coughlin-Woods et al., 2005) and spontaneously recovered children who used to stutter (Finn et al., 1997) were the subjects of the only studies focusing on speech naturalness in children. The speech naturalness of children who stutter was also measured as one dependent variable comprising part of larger studies that focused on stuttering treatment (Craig et al., 1996; deKinkelder & Boelens, 1998; Druce, Debney, & Byrt, 1997; Hancock et al., 1998; Ingham & Riley, 1998; Onslow et al., 1990). deKinkelder and Boelens did not use the traditional 9-point scale (Martin et al., 1984) and provided no reliability data for the scale they did use: therefore, their study will not be discussed further.

The largest speech naturalness study conducted with children's speech samples used fluent speakers rather than children who stutter (Coughlin-Woods et al., 2005). The authors wanted to determine if speech naturalness ratings in children were related to sex or age. They hypothesized that the speech of older children would be rated as more natural than the speech of younger children and that speech naturalness scores would vary between genders. Sixty normally fluent children, aged 8, 10, 12, 14, and 16 with 6 boys and 6 girls in each age group, provided the majority of the speech samples. To correspond to the normal versus abnormal

speaker paradigm commonly used in speech naturalness measurement (i.e., comparing the speech of people who stutter to the speech of people who do not stutter), the authors also included 10 speech samples of children with various speech and language disorders. Out of these 10 children, 2 stuttered. The others exhibited articulation, voice, language, or hearing impairments.

Thirty-nine inexperienced listeners rated 70 30-s speech samples for the study and rerated 14 speech samples (20%) during the same session to assess intrarater agreement. Intrarater agreement procedures defined listeners as being reliable with themselves if at least 79% of their ratings were within 1 interval of their original rating on the 9-point speech naturalness rating scale (Martin et al., 1984). The only explanation the authors give regarding the seemingly arbitrary 79% criterion is:

> Responses on replicated samples were analyzed in a procedure similar to that used by Ingham et al., (1985), which defined listeners as reliable with themselves if 75% of their ratings when exposed to the same sample were within one interval of the original rating. (Coughlin-Woods et al., 2005, p. 300)

Ingham, Gow et al. (1985) was the study cited, and in this investigation, they considered intrarater reliability to be "satisfactory if at least 90% of the first and second ratings were within +/- 1 rating unit" (p. 498); therefore, the Coughlin-Woods et al. explanation appears to be inaccurate. Seven people did not have 79% of their ratings within 1 interval of their original rating; therefore, their data were eliminated from further analysis. This left 32 "reliable" listeners (82% of the original listeners) continuing in the study.

Interrater reliability was determined for these 32 listeners by calculating the percentage of the 60 original speech samples for which each listener's rating agreed within +/- 1 rating point

with each of the other 31 listeners. An arbitrary level of 60% agreement was set as the criterion for acceptable interrater reliability. This criterion is much lower than the traditionally used 80%, and using this level only ensures that listeners are reliable on a little more than half of the speech samples. Six people did not meet the interrater agreement criterion and were eliminated from further data analysis. This left 26 listeners (67% of the original listeners) who had "acceptable" (as defined by the study's authors) intrarater and interrater reliability and whose data were subsequently reported in the study. Using only the data from these 26 listeners, the speech naturalness values averaged 1.9 for the entire group of 60 children, with a standard deviation of .8 and range of 1.2-5.3. The mean speech naturalness rating for all the ages and genders was between 1.6 and 2.2 except for 8-year-old girls who had a mean of 3.0. No relationship between speech naturalness and gender was found and only 8-year-olds were found to have different naturalness ratings than the other ages. This is most likely due to the higher mean naturalness ratings of the 8-year-old girls when compared to the 8-year-old boys, as well as to all of the other age groups. No significant difference by gender or gender by age interaction was found (Coughlin-Woods et al., 2005).

Finn et al. (1997) perceptually assessed the speech *normalcy* of children who used to stutter but recovered without formal treatment by comparing their speech to the speech of normally fluent children. This normalcy assessment included speech naturalness as one of the dependent variables. The speakers for this study were 10 preschool and early school aged children who were documented as recovered from stuttering without formal treatment and 10 age and gender matched children who had never stuttered. Experienced speech-language pathologists and inexperienced undergraduate students both rated the speech naturalness of 60 s audiovisual samples, with both groups rerating all of the samples 1 week later. The

inexperienced judges' average speech naturalness rating was 4.24 (SD = 1.22, range = 2.65-6.88) for the recovered stuttering children and 3.82 (SD = 1.05, range = 2.08-5.19) for the normally speaking children. The difference between the means was not significant. The experienced judges' average speech naturalness rating was 3.71 (SD = 1.29, range = 2.57-6.93) for the recovered stuttering children and 3.24 (SD = 0.97, range = 1.64-4.71) for the normally speaking children. The difference between these means was also not significant.

An acceptable level of rater agreement was defined as ratings that differed by no more than +/- 1 rating scale value, and agreement was presented as the percentage of ratings that were within +/- 1 rating scale value of their original ratings (or the other raters' ratings). Mean intrarater agreement levels were 64% and 65%, while interrater agreement levels were 40.6% and 45.1% for inexperienced and experienced judges, respectively. Because agreement levels fell well below the traditional standard of 80%, further research is needed to determine if this low agreement level will persist when the speech naturalness of children who actively stutter and normally speaking children is measured. No difference was found between the two groups of raters, as was the case in the adult speech naturalness literature (Finn et al., 1997).

Several studies measured the speech naturalness of children who stuttered as part of a larger study and did not report any reliability or agreement data for the construct (Craig et al., 1996; Druce et al., 1997; Hancock et al., 1998). Because of this, it is not known if listeners can reliably measure speech naturalness in the speech of children who stutter. All of these studies used one to two trained clinicians to rate naturalness rather than a larger number of raters. Another notable point is that audio only samples were typically used which, as discussed above, have been shown to decrease the sensitivity of speech naturalness ratings (Martin & Haroldson, 1992).

When looking at the reliability of children's speech naturalness ratings, this limited research suggests poor inter- and intrarater agreement in most cases. Intrarater agreement of speech naturalness ratings of children's speech was only reported in two studies (Coughlin-Woods et al., 2005; Finn et al., 1997), while interrater agreement was reported in the two studies discussed above (Coughlin-Woods et al.; Finn et al.) as well as by Ingham and Riley (1998) and Onslow et al. (1990). Finn et al. reported average interrater agreement levels of 40.6% for inexperienced listeners and 45.1% for SLPs. Coughlin-Woods et al. only had 26 of 32 listeners meet their criterion that at least 60% of listeners' ratings be within +/- 1 scale value of all of the other listeners. Ingham and Riley used a single subject design and had two clinicians rate the speech naturalness of 2 preschool children who stuttered. Of these 2 preschool children, 1 had spontaneously recovered without treatment and 1 was in treatment. Child 1 had an interrater agreement rating of 73%. Child 2 had interrater agreement ratings of 90% during baseline, 96% during treatment, and 74% during withdrawal. The total interjudge agreement across conditions for child 2 was 86%. (Interrater agreement was defined as the percentage of the 2 listeners' ratings that were within +/- 1 scale value of each other). Onslow et al. also used a single subject design with two clinicians rating the speech naturalness of 4 preschool children who stuttered. They reported the two clinicians' data graphically and found that one clinician consistently rated naturalness higher than the other in 3 of the 4 speakers with that trend reversed in the fourth speaker. Upon visual inspection, the ratings varied by about 2 scale values on most occasions.

This review suggests that the reliability values found in the adult speech naturalness literature (see Appendices D and E) may be equivalent in the measurement of children's speech naturalness. The body of literature for the speech naturalness of children who stutter is exceedingly small, with most of it conducted with children who did not currently stutter. Only

five children who stuttered participated in all of the studies discussed above (one in Ingham & Riley, 1998 and four in Onslow et al., 1990), and in both of these studies only two raters were used to measure speech naturalness. This low number of raters may have affected the reliability reported in these studies. As the above review outlines, due to the limited research in speech naturalness with children who stutter, there were numerous issues that warranted investigation.

## Statement of the Problem

The literature review above revealed the lack of research that has been done with speech naturalness and children who stutter. Although speech naturalness is recommended as a needed treatment outcome measure with children who stutter (Bloodstein, 1995; Costello, 1983; Ingham & Riley, 1998), little scientific evidence supports that this concept can be reliably measured in this population. Only two published studies focused on the speech naturalness of children as the main dependent variable, and they studied normally fluent children (Coughlin-Woods et al., 2005) and children who used to stutter but recovered without treatment (Finn et al., 1997). Several studies assessed the speech naturalness of children who stutter as one dependent variable comprising part of a larger study with the majority of these studies rating the naturalness of 10 or fewer children (de Kinkelder & Boelens, 1998; Finn et al., 1997; Ingham & Riley, 1998; Onslow et al., 1990). One study used 15 children (Druce et al., 1997) and two reported the short term and follow-up data of the same 97 children who stuttered (Craig et al., 1996; Hancock et al., 1998).

To determine the need for investigating speech naturalness training and to assess the internal definition standard currently used in speech naturalness research, listeners' intra- and interrater agreement when rating the speech naturalness of children who stutter was investigated in a series of two studies. The first study addressed basic speech naturalness ratings, rater

agreement when listeners rated the speech of children who stutter and the effect of *SMS* training on these agreement levels. The specific questions identified for investigation in Study One were (Chapters 2 and 3):

1) What are inexperienced listeners' speech naturalness ratings for speech samples of children who stutter using Martin et al.'s (1984) 9-point scale?

2) Using Martin et al.'s 9-point scale, what are inexperienced listeners' intrarater and interrater agreement levels for speech naturalness of 30-s speech samples of children who stutter?

3) Will there be significant differences in intrarater and interrater agreement of speech naturalness ratings among random assignment of raters to 1 of 3 groups on the first two (of four) rating occasions?

4) Will participants report being able to generalize speech naturalness training using a training program that has speech samples from mostly adult speakers to the measurement of speech naturalness in the speech of children?

5) Will a group of inexperienced listeners' intrarater and/or interrater agreement levels, when rating the speech naturalness of children who stutter, improve after completing a modified version of the *SMS* speech naturalness training when they are compared to an exposure control group and a control group?

The findings related to question 5 provided effect size and power analysis values in order to more specifically determine the required sample size for Study Two. The specific methods and results from Study One are presented in Chapters 2 and 3.

Study One addressed several of the initial questions raised by the previously existing literature of speech naturalness in children. Based on those results, Study Two addressed speech naturalness ratings, rater agreement and reliability with ICCs by stuttering severity, and the

effect of training on listeners' ratings and agreement using the speech samples of children who do and do not stutter. The specific questions identified for investigation in Study Two were:

1) What are inexperienced listeners' speech naturalness ratings of normally speaking children and children who stutter at mild, moderate, and severe levels?

2) Do listeners' speech naturalness scores for children differentiate among normally speaking children and children who stutter at mild, moderate, and severe levels?

3) What are inexperienced listeners' intrarater and interrater agreement levels when rating the speech naturalness of 30-s speech samples of normally speaking children and children who stutter at mild, moderate, and severe levels?

4) Will speech naturalness training using the *SMS* or exposure to the training videos significantly improve listeners' intrarater and interrater agreement levels?

5) Will stuttering severity (normal, mild, moderate, severe stuttering) affect listeners' intrarater and interrater agreement levels?

6) What are the Intraclass Correlation Coefficients (ICCs) for the speech naturalness ratings of listeners rating the speech naturalness of children who do and do not stutter at various severity levels?

Methods and results for Study Two are presented in Chapters 4 and 5.

# CHAPTER 2

# STUDY ONE: METHOD

## Study Design

A pre-test post-test group design, utilizing multiple control groups, was used to investigate inexperienced raters' speech naturalness scores, agreement scores, the effect of random assignment on group equality, and the effect of a modified version of the *SMS* training procedures designed to teach judges to rate speech naturalness. Inexperienced raters were used as judges based on the framework that listeners in a speaker's environment judge what is natural as well as the literature reviewed in Chapter 1 regarding the effect of listener experience on listener reliability. This study was deemed necessary because, as discussed in the above literature review, a study having a large group of listeners rate speech naturalness of the speech of children who stutter has not been conducted. Because of this, several key issues affecting the study presented in Chapters 4 and 5 warranted investigation. Specific goals for Study One were to determine: 1) inexperienced listeners' speech naturalness ratings when rating speech samples from children who stutter with the majority at a moderate or severe level; 2) inexperienced listeners' intra-and interrater agreement levels when rating the speech naturalness of 30-s speech samples of children who stutter; 3) if random assignment of raters to groups resulted in groups with no significant differences in intra- and interrater agreement scores after the first two rating occasions; 4) if participants reported being able to generalize speech naturalness training using a training program with mostly adult speech samples to the measurement of speech naturalness in children; and 5) if the *SMS* speech naturalness training program improved inexperienced

listeners' intra- and interrater agreement to a statistically significant level when compared to an exposure control and control group? Methods to address these goals are described in detail below.

## Method

Forty-three inexperienced student raters were randomly assigned to one of three groups: a Training Group (n=15) that received modified *SMS* speech naturalness training (discussed further below) after the first two (of four) speech naturalness rating occasions; an Exposure Control Group (n=14) that rated the speech naturalness of the video samples seen in the *SMS* training and criterion test, but received no feedback regarding their ratings; and a traditional Control Group (n=14) who had no training or exposure to the training stimuli after the first two (of four) speech naturalness rating occasions. The participants in each group rated the speech naturalness of 24 speech samples of children who stuttered (described in detail below) four times on two separate occasions. Because of the possible problems presented with having listeners only rerate a portion of data as discussed in Chapter 1, this study used the more stringent method of having all of the participants rerate all of the speech samples during each rating occasion.

### *Participants*

The 43 inexperienced female student judges who participated in this study had a mean age of 21. All but four of the students were undergraduates with no formal or informal training in stuttering; all were enrolled in the same audiology class at the University of Georgia and all received 2.5% extra credit in the audiology class for participating. The remaining four judges were Master's level students who reported minimal training in stuttering with their only experience in their undergraduate speech disorders class. No judge stuttered or reported a history of stuttering, and all judges reported English as their primary language. Twenty-one of

the students reported knowing someone who stuttered, but their relationship to the person was not reported. This issue was addressed in Study Two because participants were excluded from participation if they reported knowing someone close to them (family member, significant other, teacher) who stuttered.

*Sample Size*

To ensure adequate statistical power to identify a difference between trained and untrained speech naturalness raters, sample size was determined by a power analysis with power set at .80, alpha set at .05, and assuming an effect size of .40 (Cohen, 1969). Because this study used analysis of variance techniques to assess the effect of training on repeated multiple measurements of speech naturalness, Cohen's recommendations for sample size for "F tests in the Analysis of Variance" were used. Based on Cohen's recommendations, and using the above values, a required sample size of 13 people per group was estimated.

Because the speech naturalness literature reviewed above does not provide group differences to help determine effect size for the current study, the closest relevant standards located were used. These came from Einarsdottir and Ingham (2008), who found a large effect size (d=1.5) when measuring the effect of an Icelandic version of the *SMAAT* training program on the identification of stuttering by preschool teachers. This study was chosen because it assessed the effects of training and used the speech samples of preschool children who stuttered, both of which are addressed in this study. In another related study, an effect size of d=.6 was found when comparing trained and untrained parents of children who stuttered when identifying the presence of stuttering in children (Einarsdottir, 2009). As above, this study was chosen as a reference because it assessed training effects on the identification of stuttering in preschool children. Because these studies do not examine the exact issues as this study, and because it is

not known if the effect size of English speaking participants will be the same as the Icelandic speakers, a lower, more conservative estimate of effect size was chosen (.40).

## Materials

### *Speech Samples*

*Preparation of stimuli.* Audiovisual speech samples previously recorded at The University of Georgia from 9 different children who stuttered (none of these children was part of the *SMS* training videos) were used to develop the stimuli for this study. These recordings, from 5 males and 4 females, averaged 4 minutes in length. Specific age information was not available for the children. All of these speech samples were transferred from a Digital Video Disk (DVD) to a digital video file using Hand Brake software and an iMac 24-inch desktop computer. These samples were chosen because they were all videos of elementary aged children who were sitting at a table talking with no other people present in the video shot. No treatment was being administered in any of the samples. The children's parents previously gave consent for the recorded speech samples to be used for research purposes. Stuttering severity was rated on a three-point scale of mild, moderate, and severe by 2 graduate clinicians with extensive experience collecting stuttering data, and by the researcher using the children's 4-min speech samples in their entirety (see Table 1). Three speakers were classified as exhibiting mild stuttering (contributing 6 speech samples), three moderate (contributing 9 speech samples), and three severe (contributing 9 speech samples). At least two of the three judges agreed on the severity rating for each speaker.

From among these longer samples, 30-s speech samples were chosen based on the following criteria: 30 s of continuous speech with no more than 2 s of someone other than the child speaking; each sample had to contain at least one stutter, as agreed upon by 2 graduate

clinicians unaware of the purposes of the study and the researcher; the speaker exhibited no obvious articulation or language deficits as evidenced by normal Percent Consonant Correct (>90% correct) and Mean Length of Utterance (>8.0) scores (see Table 1), as judged by a speech-language pathologist holding the Certificate of Clinical Competence from the American Speech-Language-Hearing Association and two speech-language pathology graduate students. Speech samples of 30 s were chosen because the small body of research discussed in Chapter 1 does not definitively conclude that the rater agreement for one speech sample length is clearly stronger than the other, and because the only study directly investigating speech sample length used stutterfree utterances as opposed to utterances including stuttering that were used in this study. A major consideration in this decision was the limited availability of continuous 60-s speech samples of children who stutter. Based on the information reviewed in Chapter 1, the researcher felt uninterrupted continuous speech samples should be prioritized over speech sample length and, therefore, chose 30-s samples for this study. Twenty-four samples met these criteria and comprised the final stimuli for this study. Quick Time software was used to edit the original 4-min samples discussed above into these 30-s segments meeting the above criteria. Percent syllables stuttered and syllables per minute were also collected for each 30-s sample to be used in this study by two graduate clinicians unfamiliar with the study. As can be seen in Table 2, these speech samples present a wide range of stuttering frequency and speech rate.

*Experimental stimuli.* The 30-s segments were transferred to iMovie software where breaks and prompts were added. Before each 30-s segment, a 3 s black screen appeared with the stimulus number in large white print in the center of the screen. After each sample, a visual prompt of "Record Naturalness Rating" appeared for 7 s to give the listener time to rate the segment. Four different randomized orders of these 24, 30-s samples were created to control for

any possible order effects in the participants' speech naturalness ratings. These four orders were created by randomly placing the speech samples in order under the restriction that two samples of the same speaker must be separated by at least two samples of other speakers. Each sample was numbered 1 to 24 and then the computerized random number generator Randomness 1.5.2 (Merenbach, 2007) was used to create the order in which the samples were to be placed on the Digital Video Disks (DVDs). Two orders were created where the 24 speech samples appeared twice, and DVDs of the two orders of 48 speech samples were made for the participants to review. Each of the two randomized orders of 30-s speech samples, pauses, and prompts was transferred to two separate DVDs as a QuickTime movie using the iMovie software program.

Each participant was randomly assigned to use one order on her first rating occasion and the other order on her second. Participants watched the DVD on their personal computer and simultaneously recorded their naturalness ratings on a data collection form (Appendix H). The data were hand transferred from these data sheets to an Excel spreadsheet on two separate occasions by the researcher. The data were compared within the spreadsheet to locate entry errors and no data entry errors were found upon comparison.

## Procedure

Each of the 43 participants met with the researcher individually three times: once to complete the necessary IRB paperwork, and intake questionnaire (see Appendix I), to sign a pledge not to discuss the study with anyone (Appendix J), to receive a packet containing detailed instructions on how to complete the first task (see Appendix K), to receive a copy of the 9-point naturalness scale (Appendix B), and to receive a data collection form (Appendix H); once to return the first packet and receive their second one containing detailed instructions on how to complete the second task (Appendix L), the 9-point naturalness scale (Appendix B), and data

collection forms (Appendices H and L); and finally to return their second packet and complete a post-study questionnaire (Appendix M).

Each participant rated each of the 24 speech samples twice during two rating sessions. They were asked to complete these sessions independently. Each participant used her own computer and headphones to complete the tasks. During their initial meeting with the researcher, participants signed a pledge stating they would not discuss the study with anyone other than the researcher (see Appendix J). All participants reported via the post-study questionnaire (Appendix M) that they completed each task independently and that they did not discuss them with their classmates. Members of the Control Group completed two Rating Sessions in which they rated the 24 speech samples twice during each session. Members of the Training Group additionally completed a Training Session and its associated posttest immediately before the second Rating Session. Members of the Exposure Control Group completed one longer session in which they rated the naturalness of the speech samples used in the training program and posttest, but were given no feedback as to the accuracy of their ratings. The participants completed all tasks independently with no direct supervision provided by the researcher. Detailed written instructions were provided to the participants by the researcher for each task (Appendices K and L). The participants reported via a post study questionnaire (Appendix M) that they completed the two sessions an average of 7 days apart. All sessions and procedures are described in greater detail below.

*Training Group*

The 15 member Training Group completed two sessions: a Rating Session and a Training + Rating Session. Each participant completed each session independently with each person reporting she had no difficulty completing the two tasks.

*Pretraining Rating Session.* The pretraining Rating Session took approximately 32 minutes to complete (24 minutes for the 24 speech samples of 30-s duration twice and 10 s of pause time for each sample taking 8 minutes). Listeners rated the 24 speech samples once and then immediately rated them a second time using another randomized order during the same session. They were given specific instructions on how to complete the task (Appendix K). Participants recorded their naturalness ratings on the data collection form containing lines numbered 1 to 48 (Appendix H). Half of the participants were randomly assigned to use one of the randomly ordered DVDs, while the other half used the second order. Each participant was scheduled to return their packet and pick up their final task 5 days after they were given the first task.

*Training + Rating Session.* The final task for members of the Training Group lasted approximately 1 hour and 15 minutes and consisted of the participants independently completing a modified version of the *SMS* naturalness training program and Criterion Test, and rating the 24 speech samples twice using the second randomly ordered DVD not used in the first Rating Session. As before, no time elapsed between the two rating occasions. After completing training, each participant was given specific instructions to complete this session (Appendix L). Participants recorded their naturalness ratings on the *SMS* data collection forms (Appendix L) and on another copy of the same data collection form used in the previous session (Appendix H). The *SMS* modifications referenced are described below. First, rather than complete the *SMS* program as is traditionally done over the Internet, participants watched the training and Criterion Test videos from a DVD. A second modification was that participants recorded their speech naturalness ratings via paper and pencil rather than using the *SMS* online program. Because the computerized *SMS* program was not used, the 2 and 3 min samples were divided into 1 min

samples by the researcher and placed on a DVD rather than having the participants view the samples online. Typically the computerized version of the *SMS* program averages the speech naturalness ratings for the 2- and 3-min samples, but because it was not used the researcher averaged the two naturalness ratings (or three for the three minute samples). This was done after the study was completed to derive the final naturalness rating for each longer sample. Another modification to the original program was that participants completed all of the Criterion Test samples one time and were not given naturalness feedback for the Criterion Test. Whether or not the participant met the Criterion Test standard was determined by the researcher after the study was completed. This was done so participants did not have access to the standards because seeing the standards before completing the task may have influenced their ratings. All 15 participants met the Criterion Test standard of three consecutive speech naturalness ratings within the range specified in the *SMS* program.

*Exposure Control Group*

The 14 members of the Exposure Control Group completed two sessions both with specific instructions: a Rating Session, as described above for the Training Group (Appendix K), and a longer Rating Session in which the participants rated the speech samples used in the *SMS* training program and Criterion Test, followed by the 24 experimental speech samples twice (Appendix L). As with the Training Group, this group viewed the videos from a DVD prepared by the researcher rather than online. As discussed above, the longer samples were broken into 1 min samples by the researcher and were averaged for the final speech naturalness value by the researcher at the end of the study. Participants reported completing each session independently, with each person reporting she had no difficulty completing the two tasks. Seven participants used one randomly ordered DVD, while the other seven used the second order during each task.

Only 4 of the 14 participants in this group met the Criterion Test standard of three consecutive speech naturalness ratings within the range specified in the *SMS* program.

*Control Group*

The 14 members of the control group completed two rating sessions using specific instructions (Appendices K and L). Participants completed each session independently with each person reporting she had no difficulty completing the two tasks. Seven of the participants used one randomly ordered DVD during the first Rating session while the other seven used the second order. Each participant used the order she did not use during session one during session two. The first Rating Session was conducted as described above for the Training Group. The second Rating Session was conducted like the first one with the exception of a statement instructing participants to rate the speech naturalness of the sample without trying to remember their previous ratings (Appendix L).

## Data Analysis

Data from the participants' data collection sheets were entered into the Microsoft Excel computer program to begin data analysis. The researcher transferred these data from the data sheets to an Excel spreadsheet twice, and the data were compared to locate entry errors. No data entry errors were found between the two entries. Once all of the data were successfully entered into Excel, the spreadsheet was exported to SPSS version 18, which was used for continued data analysis.

Mean speech naturalness ratings are presented for each speech sample on each rating occasion (Appendix N) as well as the mean, range, and standard deviation for all 24 speech samples combined for all 43 raters combined on the first two rating occasions (Table 3). Interrater and intrarater agreement levels (or the percentage of ratings within +/- 0 to 8 scale

values of one another) are presented for each group (Tables 4 & 5), along with each participant's agreement ratings for each rating occasion (Appendices O and P). Participant responses to the post study questionnaire were compiled for all participants.

*Independent Variables*

For research questions 1, 2, 3, and 4 (questions not addressed using ANOVA statistical procedures), the independent variables were inexperienced listeners, random assignment of raters to groups, and *SMS* training, respectively. For question 5 (addressed using ANOVA statistical procedures), the independent variable was the presence of *SMS* training.

*Dependent Variables*

For research questions 1, 2, 3, and 4 (questions not addressed using ANOVA statistical procedures), the dependent variables were speech naturalness ratings (for 1), intra- and interrater agreement levels (for 2 and 3), and the participants' self-reports about whether and how they believed they generalized training using the *SMS* to rating the experimental speech samples (for 4). For question 5 (addressed using ANOVA statistical procedures), data analysis was conducted in terms of two dependent variables: intrarater agreement for the participants' speech naturalness scores on the first 2 rating occasions and the last 2 rating occasions, and interrater agreement of the participants' speech naturalness scores on rating occasions 1, 2, 3, and 4. These agreement ratings were calculated for the judges within each group. For interrater agreement, the speech naturalness ratings of each person in the Training Group were compared to each of the other 14 group members, while the speech naturalness ratings for each person in the two control groups were compared to each of the other 13 members in their respective group.

*Speech naturalness ratings.* The mean, range, and standard deviation for all 43 raters for all 24 speech samples combined are presented for occasions one and two prior to any training

(Table 3). The mean speech naturalness scores by group for each speech sample for each rating occasion are presented (Appendix N). Speech naturalness ratings are not presented by stuttering severity as there were more moderate and severe samples than mild.

*Intrarater agreement*. Intrarater agreement was calculated using difference scores which were derived for each listener by subtracting the speech naturalness score for each speech sample on occasion 1 from the speech naturalness rating of the same speech sample on occasion 2 for every rater in each of the three groups. The direction of the difference was disregarded. This was repeated for the speech naturalness scores for each speech sample on occasions 3 and 4 for each rater. The experimental group had a total of 360 difference scores (24 speech samples x 15 judges) and the two control groups had 336 difference scores each (24 speech samples x 14 judges). The percentage of ratings within +/- 0, 1, 2, 3, 4, 5, 6, 7, and 8 scale values of one another were then calculated for each group and reported (Table 4). Individual data for the participants are also presented (Appendix O). As is traditional in stuttering speech naturalness research, the percentage of ratings within +/- 1 scale value or less of each other, for each rater, was used as the standard for intrarater agreement, and these percentages per rater for each occasion were used for further data analysis. A one-way analysis of variance procedure was conducted to determine if the intrarater agreement for occasions 1 versus 2 was different between groups. A two-way analysis of variance procedure (within subjects rating occasion comparison-2 levels x between subjects group-3 levels) was used to determine if the groups' intrarater agreement differed to a statistically significant level for rating occasions 1 versus 2 and occasions 3 versus 4. Individual differences between groups were assessed using Fisher's Least Significant Difference procedure as a statistically significant main effect for group was found. Mean differences between groups using 95% confidence intervals are also presented (Figure 1).

*Interrater agreement.* Difference scores also estimated interrater agreement. These difference scores were calculated by comparing the speech naturalness score for each speech sample at each occasion for the first judge in each group to the speech naturalness rating of the same speech sample on the same occasion for every other judge in the same group. The direction of the difference was disregarded. Judges in the experimental group had 2520 difference scores (24 speech samples x 105 rater comparisons) at each rating occasion. Each control group had 2184 difference scores (24 speech samples x 91 rater comparisons) at each rating occasion. The percentage of ratings within +/- 0, 1, 2, 3, 4, 5, 6, 7, and 8 scale values of one another was then calculated for each group and reported (Table 5). Individual data for the participants is also presented (Appendix P). As is traditional in stuttering speech naturalness research, the percentage of ratings within +/- 1 scale value or less of each other, for each rater, was used as the standard for interrater agreement. These percentages per rater for each occasion were then used for further data analysis. The effect of training and rating occasion on participants' interjudge agreement of speech naturalness ratings was analyzed using a two-way Repeated Measures ANOVA with "Group" the between factor (3 levels-Training, Exposure Control, Control), and "Rating Occasion" the within factor (4 levels). Individual differences between groups were assessed using Fisher's Least Significant Difference as statistically significant main effects were found. Mean differences between groups using 95% confidence intervals are also presented (Figure 2).

# CHAPTER 3

# STUDY ONE: RESULTS AND DISCUSSION

All participants completed both sessions and provided usable data, and all 43 reported no difficulty completing the tasks.

## Speech Naturalness Ratings

The mean speech naturalness rating for all 43 raters for all 24 speech samples was 5.67 on occasion one (with a range of 2.4 to 8.4) and 5.97 on occasion two (with a range of 2.8 to 8.4; Table 3). For the experimental group, speech naturalness values ranged from 2.4 to 8.4, 2.3 to 8.5, 4.3 to 8.2, and 4.7 to 8.2 of all four rating occasions, respectively (Appendix N). The exposure control group's speech naturalness ranges on all four occasions were 2.7 to 8.3, 2.9 to 8.3, 3.0 to 8.0, and 2.8 to 8.3. For the control group, speech naturalness values ranged from 2.4 to 8.5, 3.1 to 8.4, 3.1 to 8.4, and 3.4 to 8.2.

## Intrarater Agreement

Intrarater agreement for the speech naturalness of children who stutter as rated by inexperienced listeners ranged from 74% to 77% (Figure 1; Table 4). Occasion 1 versus occasion 2 intrajudge agreement did not differ across the three groups ($F_{(2, 40)} = .339$, n.s). After training, the training group showed significantly higher intrarater agreement than that achieved by the other two groups ($F_{(1\ 40)} = 21.49$, $p = .000$, partial eta squared = .349). A comparison of the three means using Fisher's LSD procedure found the experimental group's intrarater agreement (M = 92.78) was significantly higher than either the exposure control group (M = 80.36) or the control group (M = 81.55) after training (Figure 1). The two control groups did not

differ from each other to a statistically significant level. A significant interaction effect was found for rating occasion by group ($F_{(2, 40)} = 4.662$, p = .015, partial eta squared = .189).

## Interrater Agreement

Inexperienced listeners' interrater agreement when rating the speech naturalness of children who stutter ranged from 50.14% to 60.28% on the first two rating occasions, well below the standard 80% criterion (Table 5; Figure 2). The training group had significantly higher interrater agreement levels than the control groups on the first two occasions (Figure 2). This potentially negatively affected the gain of the treatment group, but the treatment still increased their interrater agreement for trials 3 and 4 when compared to the pretraining trials to a statistically significant level. A two-way RM-ANOVA found a nonsignificant main effect for rating occasion ($F_{(3, 38)} = .747$, n.s.), but a significant interaction effect for rating occasion x group ($F_{(6, 78)} = 4.517$, p = .001). A comparison of the group means using Fisher's LSD procedure found the experimental group had significantly higher interrater agreement levels when rating occasions were combined than the exposure control group (mean difference = 11.176, p = .000) and control group (mean difference = 12.493, p= .000). The two control groups did not differ from each other to a statistically significant level for combined rating occasions (mean difference = 1.317, n.s.).

## Post-Study Questionnaire

When asked, via the post study questionnaire (Appendix M) if they compared the experimental speech samples to children or adults, all of the participants responded that they compared the speech samples to other children's speech. Participants' responses included: "I compared each individual to their respective social age group norms" (Participant 30); "I compared it to the kids that I have been around that have normal speech of around the same age,

and also to the kids on the video that I rated as 'natural'" (Participant 33); "I did not compare the children to the adults because children aren't that fluent all the time and adults typically are" (Participant 2); and "I tried to compare the children's speech to the way I hear children talk around me, I tried to be conscious that they were children and not rate them too harsh" (Participant 9). Out of the 43 participants, only one suggested she might have had difficulty comparing the experimental speech samples to children's speech. She reported "I would like to think that I always compared kids to regular child speech, but sometimes I'm sure I might have compared the children to what I am used to hearing on a day to day basis, which would be adult speech" (Participant 1).

Out of the 29 participants that saw the *SMS* training and criterion test samples, only 3 reported difficulty transitioning from adults' speech samples back to children. Two participants reported that they were able to transition back to rating children, but had to cue themselves to compare the children's speech samples to other children and not adults stating "I had to remind myself of what the child's normal speech for their age was before I measured their naturalness" (Participant 37), and "I had to remember when I went back to the kids that they were kids and not adults" (Participant 28). The third participant reported "it was difficult to go from hearing how to judge naturalness of an adult voice to trying to judge a child's speech sample because I felt there are different expectations for children than adults" (Participant 18).

## Study One: Discussion and Interpretations

The first goal of Study One was to gather information about speech naturalness in a group of children who stuttered as measured by a group of judges. The range of averaged speech naturalness rating for all 43 raters of 5.67 to 5.97 suggests inexperienced listeners rate the speech naturalness of children who stutter at predominantly moderate and severe levels in the middle of

the 9-point rating scale. The range of speech naturalness ratings from 2.4 to 8.4 indicates several of the samples were rated in the extreme ranges, including in the "normal" speaker range of less than 3 as discussed in Chapter 1 (Appendix N). Because only six speech samples of children who stutter at a mild level were used and nine of children who stutter at moderate and severe levels, speech naturalness ratings by severity levels were not calculated in this study. Three speech samples of children who stutter at a mild level as well as nine speech samples of normally speaking children were added in Study Two to make this comparison.

The second research question for this study addressed intra- and interrater agreement of inexperienced listeners' ratings of speech naturalness for children who stutter. Obtained intrarater agreement fell between 74.1% and 77.4%. These values approach the typically accepted standard of 80%, but they should be interpreted with caution because rating occasions 1 and 2 were conducted during the same session. As is typical in speech naturalness literature and shown in Figure 2, interrater agreement for each group was lower, ranging from 50% to 60% on occasions 1 and 2. These starting values provide data that inexperienced listeners' agreements, especially interrater agreement values, are low enough to warrant another investigation of a possible way to improve them. These results support research designed to further investigate the effect of training on rater agreement as was done in Study Two.

Another question addressed in this study was if random assignment of inexperienced listeners to groups would result in groups with equal intra- and interrater agreement levels prior to any training. This technique did result in groups that were not significantly different in their intrarater agreement for occasions 1 and 2, but resulted in a treatment group with a significantly higher interrater agreement than both of the control groups on occasions 1 and 2. Because of this, results of this study should be interpreted with caution. To protect against such problems in

Study Two, an ANOVA was used after the first two rating occasions to ensure that the randomization process had resulted in groups with equal speech naturalness ratings, intrarater agreement, and interrater agreement (but see Chapter 5).

Participants' ability to transition from training with predominantly adult speakers to children's speech was addressed in Study One by asking the participants who viewed adult speakers (via a modified version of the *SMS*) questions about their ability to transition between adults and children (see Appendix M for questionnaire). Because the majority of the participants said they had no difficulty transitioning between the speech samples in the *SMS* and the experimental samples used in Study One, the *SMS* was used in Study Two as planned in an attempt to determine if listeners generalize speech naturalness training using longer speech samples of adults and children to the rating of shorter, children-only speech samples.

To address question 5, the *SMS* training program did appear to have a positive effect on the training group's agreement levels when compared to the 2 control groups, as their intra- and interrater agreement levels were significantly higher post training than the control groups'. This finding provided support for further investigation of the *SMS* in Study Two, but should be interpreted with caution due to Study One's limitations. First, 21 subjects reported knowing someone who stuttered, but the relationship to the people who stuttered is not known. If some of the participants had close relationships with someone who stuttered, it may have affected their ratings. This was controlled for in Study Two as participants who had immediately family members, spouses, close friends or teachers who stuttered were excluded from participation. Secondly, as discussed above, the groups' interrater agreement levels were not equivalent prior to the training group's *SMS* exposure. Another major limitation of Study One was that participants independently completed all steps of the study. Although group participants

reported they were able to complete each step of the study independently and without difficulty, because the researcher was not present with the participants as they completed the study, it cannot be guaranteed that each participant followed the step-by-step instructions exactly. Evidence that they completed the tasks as instructed was based solely on participants' self report. Greater control was asserted in Study Two as the researcher was present with all participants throughout all sessions to ensure study compliance. Another limitation to Study One that may have affected results was the modifications made to the *SMS* training program. As discussed above, modifications to the *SMS* in this study were: not viewing the videos online but from a DVD with all of the clips being 60 s in length; not entering the speech naturalness data using the *SMS* program, rather having participants use a paper and pencil to record speech naturalness ratings; and not giving the training group feedback regarding their Criterion Test ratings as is done in the *SMS* program. Criterion Test results were not given to the participants because this was a take home independent task and the researcher did not want to risk participants looking at the criterion prior to rating the speech samples thereby influencing their ratings. Any of these issues may have affected the results of Study One and were controlled for in Study Two by using the computerized version of the speech naturalness portion of the *SMS* without any significant changes (see Chapter 4).

Although no one in the training group reported they had difficulty completing the second (Training + Rating) task, 7 people did note that it was long "which made it difficult to stay focused" (Participant 11). Because the task was completed independently, the researcher can not know for sure if the participants did the tasks in their entirety or if they took unscheduled breaks while completing a task, an issue that was controlled for in Study Two. Also, the length of the Training + Rating task was shortened in Study Two as the participants only rated the

experimental samples one time after training. This was done in hopes of controlling for the internal validity threat that may have occurred during Study One due to fatigue. The results of Study One, combined with the literature reviewed in Chapter 1 lead to the hypotheses investigated in Study Two, which are presented below.

## Study Two: Research Questions

This study investigated the following research questions:

1) What are inexperienced listeners' speech naturalness ratings of normally speaking children and children who stutter at mild, moderate, and severe levels?

2) Do listeners' speech naturalness scores for children differentiate among normally speaking children and children who stutter at mild, moderate, and severe levels?

3) What are inexperienced listeners' intrarater and interrater agreement levels when rating the speech naturalness of 30-s speech samples of normally speaking children and children who stutter at mild, moderate, and severe levels?

4) Will speech naturalness training using the *SMS* or exposure to the training videos significantly improve listeners' intrarater and interrater agreement levels?

5) Will stuttering severity (normal, mild, moderate, severe stuttering) affect listeners' intrarater and interrater agreement levels?

6) What are the Intraclass Correlation Coefficients (ICCs) for the speech naturalness ratings of listeners rating the speech naturalness of children who do and do not stutter at various severity levels?

The specific research methods used to address these six questions are described in the next chapter.

# CHAPTER 4

# STUDY TWO: METHOD

This chapter describes the participants, materials, and procedures used in Study Two. Participants were randomly assigned to one of three groups: a Training Group, an Exposure Control Group, or a Control Group. Each group rated the speech naturalness of 36 speech samples, 9 from normal speaking children and 9 each from children who stutter at mild, moderate, and severe levels. Because of the possible problems presented with having listeners only rerate a portion of data, as discussed in Chapter 1, this study used the more stringent procedure of having all of the participants rerate all of the speech samples during each rating occasion for purposes of calculating rater agreement. All 36 speech samples were rated four times, with the Training Group and Exposure Control Group rating the samples two times before they were trained or viewed the training samples without receiving feedback, and two times after they were trained or viewed the training samples. Rating occasions occurred 7-14 days apart. The steps used for the preparation and analyses of collected data, including reliability procedures, are also described.

## Sample Size

To ensure adequate statistical power to identify a difference between trained and untrained speech naturalness raters, sample size was calculated using the power analysis program G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2009). The various effect sizes for each effect tested in Study One, along with the sample size calculations using these effect sizes, are reported in Table 6. For these calculations power was set at .80 and alpha was set at .05. The

recommendations for sample size of the F test were used to calculate sample size using these values. For intrarater agreement, the smallest effect size was f = .338, which was for the main group effect when comparing the intrarater agreement of the experimental and two control groups. Using this effect size, a total sample size of 27 (n=9) was estimated. For interrater agreement, the smallest effect size was found for the main effect of rating occasion and was f = .105. Using this effect size, a total sample size of 15 (n = 5) was estimated. The sample size of 54 (n= 18) is well above any of the above projections. Effect size values for testing each research question resulting from using 54 participants will be presented in the results section.

## Participants

Fifty-four women aged 19 to 32 (mean age = 20.70, SD = 2.16) who were enrolled in a basic audiology class at the University of Georgia and who had no formal or informal training in stuttering served as participants. Inexperienced raters were used as judges based on the framework that listeners in a speaker's environment judge what speech sounds natural as well as the literature reviewed in Chapter 1 regarding the effect of listener experience on rater reliability. Inclusion criteria for all participants were: English as their primary language; no neurological disorders or learning difficulties; no formal clinical or academic training in stuttering; did not participate in Study One reported in Chapters 2 and 3; had no immediate family members, spouses, significant others, teachers, or close friends who stuttered; and had no history of ever stuttering themselves. This information was assessed by self-report via a short pre-study questionnaire (see Appendix Q). Fifty-seven people volunteered for the study, but three participants were disqualified from further participation because they had immediate family members, significant others, or close friends who stuttered. Participants also had to pass a vision and hearing screening to participate.

Vision was screened using a standard Snellen vision chart. Participants had to be able to read the following line or smaller when standing 20 feet away.


A pure tone air conduction hearing screening was conducted at 20 dB bilaterally at 500, 1000, 2000, & 4000 Hz using a portable audiometer. All participants passed both of these screenings with or without aided help. Participants meeting these criteria were randomly assigned to one of three groups. Experimental group participants had to successfully complete the naturalness training portion (Part Four) of the *SMS* program to have their data included in the study (see below). Because neither the mean speech naturalness ratings nor intra- or interrater agreement ratings for the three groups were found to differ to a significant level after the first two rating sessions, participants were not matched into groups.

## Materials

### *Speech Samples*

Speech sample preparation was the same as those described above in Study One (see Chapter 2) except 9 speech samples from normally speaking children were added, along with three additional samples of children who stuttered at a mild level. This resulted in nine speech samples at each severity level. In addition to the criteria outlined in Chapter 2, for the normal speakers, the speech samples had to be deemed stutter free as agreed upon by the researcher and 2 graduate clinicians unaware of the purposes of the study. The final 36 stimuli contained speech samples from 9 different children who stuttered, 5 boys and 4 girls, and 3 different normally speaking children, 2 boys and 1 girl. Of the 36 30-s samples, 27 were from children who stuttered (9 each at mild, moderate, and severe levels) and 9 were from normally speaking

children. For the children who stuttered, 15 speech samples were from boys and 12 were from girls, and for the normally speaking children, 4 were from boys and 5 were from girls. For the children who stuttered, one child had only one speech sample included, four children had two, and a different child had three, four, five, and six speech samples included. For the normally speaking children, one child had five speech samples included, while the other two children had two speech samples each in the final stimuli used for this study. As can be seen in Table 2 these speech samples present a wide range of stuttering frequency and speech rate.

*Experimental stimuli.* Four randomized orders of these 36 speech samples were created and transferred to DVD as described in Chapter 2.

## Procedure

Each of the 54 participants met with the researcher on four separate occasions, and all participants completed all four sessions in the researcher's presence. Participants completed each session either individually or with up to three other participants completing the same task at the same time. When participants contacted the researcher with interest in participating in the study, they completed the intake questionnaire (Appendix Q). If they met all of the inclusion criteria discussed above, they were scheduled for their first session. During the first session, all participants completed the necessary IRB paperwork (Appendix R), completed a vision and hearing screening, signed a pledge not to discuss the study with anyone (Appendix J), and completed their first session task (session task details are discussed in greater detail below). During the remaining three sessions, participants completed their assigned tasks (discussed in detail below). In addition, after their fourth session, all participants completed a post study questionnaire (Appendix S) and asked any questions they had.

All sessions were conducted in the same room in Aderhold Hall at The University of Georgia with each participant using a Dell Optiplex GX620 computer with a 19 inch screen and EckoUnltd sound reducing headphones . The room was arranged so that participants could only see their own computer with visual access to no others during each session. Sessions were scheduled for participants 7-14 days apart with no details given to them about their upcoming sessions. The next paragraph will present an overview of the study's tasks for each group, and each group's tasks are then described in detail.

As discussed in the "Participants" section, after the first two rating sessions each person was randomly assigned to one of three groups: an Experimental Group, an Exposure Control Group, and a Control Group (Table 7). All participants rated each of the 36 experimental speech samples four times during four separate rating sessions spaced 7 to 14 days apart. In addition to these four rating sessions, members of the Training Group also completed one training session consisting of the presentation of a naturalness training program and its associated posttest immediately followed by rating the naturalness of the 36 experimental speech samples during this third session. Members of the Exposure Control Group also completed one longer third session in which they rated the naturalness of the speech samples used in the training program and posttest followed by the experimental speech samples, but these participants did not receive any feedback as to the accuracy of their ratings on the training samples. The Control Group only rated the 36 speech samples one time during their third session. All sessions and procedures for each group are described in greater detail in the following sections.

*Training Group*

The Training Group completed four sessions: two rating sessions, one training + rating session, and a final rating session. The researcher was present throughout each session**.**

*First pretraining rating session.* The first pretraining rating session lasted approximately 45 min (10 min to complete the necessary IRB paperwork and answer questions, 10 min to complete a vision and hearing screening, and 24 min to watch 36 samples of 30-s duration and 10 s of pause time to record naturalness rating for each sample). Once the participant signed the consent form (Appendix R) agreeing to participate in the study and asked any questions she had, a pure-tone air-conduction hearing screening at 20 dB bilaterally (at 500, 1000, 2000, and 4000 Hz) was conducted, as well as a vision screening using the Snellen Vision chart. All participants passed both of these screenings, with or without aided help, and continued to participate in the study. After passing the vision and hearing screenings, each participant read and signed a pledge agreeing not to discuss the study or speech samples with other people until the completion of the study (Appendix J). Participants were seated in an experimental room at a table 1.5 feet from a computer screen. Next, the researcher read specific instructions to rate speech naturalness (based on Martin et al.'s, 1984 instructions; Appendix T) as each participant followed along on a written copy. Each participant repeated the instructions back to the researcher to ensure understanding and the participants were given the data collection form containing lines numbered 1 to 36 on which to write their naturalness ratings (Appendix U). A separate 8.5 x 11 sheet of paper was taped on the table in front of the participant with the 9-point naturalness scale drawn across the middle of the page and labeled "1 = highly natural 9 = highly unnatural" as a reminder of the scale values (Appendix B; Martin et al., 1984).

After the researcher read the instructions and the participant had the opportunity to ask any questions she had, the experimenter activated the DVD on one of the four University of Georgia stuttering lab computers. Participants watched the DVD on one of these computers in a window filling the entire screen using headphones to circumvent ambient noise, and

simultaneously recorded their naturalness ratings on a data collection form (Appendix U). All 36 speech samples were played without stopping. One fourth of the participants used each of the four randomly ordered DVDs containing the 36 experimental samples. Although participants completed the tasks individually at a University of Georgia stuttering lab computer, up to four participants worked in the stuttering lab at the same time. All participants completing a session together were completing the same task. Although it was possible for up to four raters to complete the task simultaneously, the computers were placed at opposite ends of a table with the two tables across the room from each other so there was no way to see any of the other computers except the one each participant was facing. All four computers had identical monitors. Additionally, all raters completing the task together worked from different DVD randomizations. At the end of this session, participants were reminded of the pledge they signed not to discuss the study or speech samples with other people until the completion of the study (Appendix J), and were scheduled for their next rating session to be completed 7-14 days later. They were given no information about the upcoming session.

*Second pretraining rating session.* The second pretraining rating session for members of the Training Group lasted approximately 32 min (8 min for instructions, questions, and to complete the "Discussion of Study Scale"; 36 samples of 30-s duration taking 18 min; and 10 s of pause time for each sample taking 6 min). Upon arrival, the participants were asked to complete a "Discussion of Study Scale" (Appendix V) asking them to rate how much they had discussed the study with someone other than the researcher since the last research session. After being seated in front of one of the stuttering lab's computer screens, the researcher read instructions similar to the ones used in the first session, except these instructions included a statement asking participants not to attempt to remember what their ratings were in a prior

session but to rate how natural they thought the speech sample sounded as they listened to it during this session (Appendix W). Participants followed along on a written copy of the instructions as the researcher read them aloud. The participants were asked to repeat the instructions back to the researcher to ensure they understood the task. The rating task was conducted exactly like the first rating session with the exception that the participant saw one of the other four randomly ordered DVDs containing the 36 samples.

After completing the second pretraining rating task, each participant was reminded of the promise they initialed during the first session to not discuss the study or speech samples with other people and was scheduled for her next session, naturalness training and first posttraining rating session, to be completed 7-14 days later. Participants were given no information about their upcoming session.

*Training + rating session.* Participants in the training group completed this session individually. Upon arrival for the third session, participants were asked to complete another "Discussion of Study Scale" (Appendix V). Next, members of the Training Group completed the *SMS* naturalness training. This third session lasted between 85 min and 145 min depending on how many times the participant had to complete the *SMS* training program. Seventeen of the participants successfully completed the *SMS* program, as evidenced by achieving the Criterion Test's standard of correctly rating three consecutive samples within the *SMS* defined target range (discussed in greater detail below) after one attempt, and their session lasted approximately 85 minutes (1 min to complete the "Discussion of Study Scale," 60 min for *SMS* training and Criterion Test and 24 min to watch and rate the 36 experimental samples one time). One participant did not successfully complete the *SMS* training program after one attempt; therefore,

she repeated the program and Criterion Test and her session lasted 145 minutes. She passed the Criterion Test on the second attempt.

The study's naturalness training was the naturalness section of the *SMS* online training program (http://sms.id.ucsb.edu/index.html; Ingham et al., 2010). The *SMS* program is designed to train clinicians/users to watch and listen to a speech sample and concurrently identify the number of syllables spoken, the number of stutters uttered, and to rate speech naturalness using Martin et al.'s 1984 9-point scale, using a computerized mouse-button identification system (see Chapter 1 for a description of the *SMS* program). This study used both the speech naturalness training portion (Part 4) of the *SMS* system and the final Criterion Test samples (Part 5). Upon initiation of the training program, the researcher read the speech naturalness training *SMS* instructions to the participant while the participant followed along on a written copy of the instructions (Appendix X). After the instructions were read, the participant repeated the instructions back to the researcher to ensure she understood the task. First, basic information about speech naturalness was presented followed by a description of the Martin et al. (1984) speech naturalness rating scale that was used to measure speech naturalness in the *SMS* program. The speech samples to be presented in the training program were discussed next, followed by detailed instructions on rating speech naturalness using the *SMS* training program (see Appendix X). In order to score a participant's speech naturalness rating for a given sample as "correct" the rating had to fall within the *SMS's* predetermined target range for that sample. (See Chapter 1 for a discussion of how the target ranges were assigned.) If a rating was correct, she moved on to the next sample. If it was incorrect, the participant rerated the naturalness of the speech sample "keeping in mind the naturalness rating assigned to that sample" (Ingham & Ingham, 2010, p. 76). This process was repeated until the participant's rating did fall within the target

range. No participant had to rerate a sample more than three times. When the participant's speech naturalness ratings of all 12, 1-min samples and 4, 2-min samples were within the *SMS* target range, she had successfully completed training and continued to the *SMS* Criterion Test.

After completing the training portion of the program, the researcher read aloud the instructions on how to complete the *SMS* Criterion Test (Appendix Y) while the participant followed along on a written set of instructions. The participant repeated the instructions back to the researcher to ensure she understood the task, and then rated the speech naturalness of all eight 3-min Criterion Test speech samples every 60 s without stopping using the *SMS* computer program and the computer keyboard. After each sample, the *SMS* computer program presented the speech naturalness average for the sample and the participant recorded that average on the Criterion Test data sheet (Appendix Y). Once all 8 samples were completed, the researcher compared the participant's speech naturalness ratings recorded on the Criterion Test "Data Recording Sheet" (Appendix Y) to the target ranges presented in the *SMS* program manual. Seventeen of the participants had three consecutive samples within the *SMS* target range after the first attempt, thereby successfully completing the Criterion Test. One participant did not meet this criterion on the first attempt, so she retrained by repeating part 4 of the *SMS* system and retook the Criterion Test during the same session. She passed the Criterion Test on the second attempt.

Immediately after participants successfully completed the Criterion Test, they rated all of the experimental samples as described above in the "first pretraining rating session" section, but this time instructions emphasized that the participant should use the information she learned during training to help her make her naturalness ratings (Appendix Z). After the instructions were read to the participant as she followed along on a written copy, she repeated the instructions

back to the researcher to ensure understanding. Each participant used a different randomly ordered DVD containing the 36 samples and recorded her ratings on a new copy of the data collection form (Appendix U). After rating these experimental samples, the participant was reminded of the promise she had earlier initialed not to discuss the study. At the completion of this session, each participant was scheduled for one final 30-min posttraining rating session. This session was scheduled 7-14 days after the completion of the rating + training session and as before, participants were given no information about their upcoming session.

*Posttraining rating session.* The last session for members of the Training Group lasted approximately 30 min (8 min for instructions, questions, and to complete the "Discussion of Study" scale and a post study questionnaire; 36 samples of 30-s duration taking 18 min; and 10 s of pause time for each sample taking 6 min). This session was similar to the first posttraining rating session (completed immediately after naturalness training and the Criterion Test) described above except that each participant used the fourth randomly ordered DVD containing the 36 samples. Like the first two rating sessions one to four participants could complete the task at the same time as long as they were all in the same group completing the same task. Again, the instructions emphasized using the training they received in the previous session were read to participants (Appendix Z) and each participant was asked to repeat the instructions to the experimenter to ensure that she understand the task. Upon completion of this session, the participant completed a post study questionnaire (Appendix S) and the experimenter answered any questions participants had about the study.

*Exposure Control Group*

Members of the Exposure Control Group completed two pretraining rating sessions as described above for the Training Group. In their third session, members of the Exposure Control

group were shown speech samples included in the "Naturalness Training" section and the "Criterion Test" of the *SMS* training program and were instructed to rate their speech naturalness, but were given no feedback regarding their ratings. Participants were read the instructions by the experimenter while they followed along on a written copy (these detailed instructions are presented in Appendix AA) and repeated the instructions back to the researcher to ensure understanding. Participants were told they were going to rate, every 60 s, the speech naturalness of 1, 2, and 3 min speech samples using a computer program called "SMS," and rated the speech naturalness of the 24 speech samples in the training and Criterion Test portions of the *SMS* without stopping. Immediately following these ratings, they again rated the 36 experimental samples as described above in the "first pretraining rating session" section for the Training Group, except the instructions from the second pretraining rating session were used (Appendix W). At the completion of this session, the exposure control group participants were scheduled for one final session 7-14 days later. Participants were given no information regarding their upcoming session. The Exposure Control Group's final session was conducted as described above in the "second pretraining rating session" section for the Training Group. Upon completion of this session, the experimenter answered any questions participants had about the study.

*Control Group*

Members of the Control Group completed two pretraining rating sessions as described above for the Training Group. In their third session, members of the Control Group rated the 36 experimental samples as described above in the "second pretraining rating session" section for the Training Group. At the completion of each of these first three sessions, the Control Group participants were scheduled for another session 7-14 days later. Participants were given no

information regarding their upcoming sessions. The Control Group's final session was conducted as described above in the "second pretraining rating session" section for the Training Group. Upon completion of this section, the experimenter answered any questions participants had about the study.

## Data Analysis

Data from the participants' data collection sheets was entered into the Microsoft Excel computer program to begin data analysis. Two research assistants transferred these data from the data sheets to an Excel spreadsheet independently, and the data were compared to locate entry errors. Where the data from the two assistants were not identical, the researcher went back to the data sheets to find and correct any entry errors. Once all of the data were successfully entered into Excel, the spreadsheet was exported to SPSS version 18, which was used for data analysis.

*Independent variables.* Independent variables for this study were raters' group assignment (a between subjects variable-Training Group, Exposure Control or Control) and stuttering severity (a within subjects variable- normal, mild, moderate, severe stuttering).

*Dependent variables.* Data analyses were conducted for this study in terms of four dependent variables: speech naturalness 1-9 ratings, intrarater agreement of raters' speech naturalness scores, interrater agreement of raters' speech naturalness scores, and ICCs from speech naturalness scores. The first data analysis to be conducted on these variables was based on descriptive statistics with the mean, standard deviation, and range reported for each group for each rating session (or comparison of 2 rating sessions for intrajudge reliability). Intrarater and interrater agreement analysis was conducted for 100% of the rating tasks, because participants rated all 36 speech samples on four separate occasions. Listeners' speech naturalness ratings and agreement scores are presented by group, by speakers, as well as for each individual participant.

These data are discussed descriptively and presented in graph or table form for visual inspection in the following pages. After descriptive statistics were used to address research questions one and three, inferential statistical procedures were used to address the other questions and are described below.

*Speech naturalness ratings.* The mean speech naturalness scores by group for each speech sample for each occasion as well as each participant's speech naturalness ratings for each speech sample on each occasion by group are presented in Appendices AB and AC. Group data are also presented for speech naturalness ratings by speaker group (Tables 8 and 9). A two-way RM-ANOVA (within subjects factor rating occasion- 2 levels x between subjects factor group- 3 levels) after rating occasions one and two to determine group equality ($F_{(2,51)} = 2.008$, n.s., partial eta squared = .073) indicated groups were not significantly different after the first two rating occasions and therefore procedures to equate the groups were not necessary. To address research question two, "do listeners' speech naturalness scores in children differentiate between normal speaking children and children who stutter at mild, moderate, and severe levels?", a three way (two within subjects variables speaker group-4 levels x rating occasion- 4 levels x one between subjects variable treatment group-3 levels) RM-ANOVA procedure was conducted. Because main and interaction effects were statistically significant, the naturalness scores were analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate specific differences between the variables that had significant effects. Group means and 95% confidence intervals are also presented collapsed across rating occasions (Figure 3) and across speaker group (Figure 4).

*Intrarater agreement.* Intrarater agreement was calculated using difference scores which were derived from each listener by subtracting the speech naturalness rating for each speech

sample on occasion one from the speech naturalness rating of the same speech sample for occasion two for every rater in each of the three groups. The direction of the difference was disregarded. This was repeated for the speech naturalness ratings for each speech sample on occasions three and four for each rater. Each group had a total of 648 difference scores (36 speech samples x 18 raters) for each rating occasion. The number of reratings that achieved difference scores of +/- 0 to 8, along with cumulative percentages for each group is presented in a table format (Table 10). Individual data for the 18 participants in each group is also presented (Appendix AD). As is traditional in stuttering speech naturalness research, the percentage of ratings with differences of +/- 1 scale value or less was used as the standard for intrarater agreement data and these percentages were used for further data analyses. The effect of training on the consistency of participants' speech naturalness ratings was analyzed using a two-way (pre/post *SMS* exposure-2 levels x group-3 levels) RM-ANOVA. Group and pre and post *SMS* exposure main effects as well as the interaction between these three variables were tested. Because the main effect for pre and post *SMS* exposure was statistically significant, the percentage of reratings with differences +/- 1 scale value or less was analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate the specific differences between pre and post *SMS* exposure. Mean speech naturalness ratings and 95% confidence intervals by training group and for both rating occasion comparisons (1 versus 2 and 3 versus 4) are also presented (Figure 5).

The percentage of ratings with differences of +/- 1 scale value or less for intrarater agreement was also used to investigate research question five, "Will stuttering severity (normal, mild, moderate, severe stuttering) affect listeners' intrarater agreement levels?" using a three-way (stuttering severity-4 levels x group-3 levels x pre/post *SMS* exposure- 2 levels) RM-

ANOVA procedure. Mean intrarater agreement ratings for all three groups for all stuttering severity levels (normal, mild, moderate, severe) along with the range and standard deviations are presented in Table 11 for both rating occasion comparisons (1 versus 2 and 3 versus 4). Individual data are also presented (Appendix AE). Stuttering severity level, group and rating occasion comparison main effects as well as the interaction among these three variables were tested. Because main and interaction effects were statistically significant, these data were analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate specific differences between the variables that had significant effects. Mean differences between speaker groups and 95% confidence intervals are also presented for all raters combined (Figure 6).

*Interrater agreement.* Interrater agreement was also determined via difference scores. These difference scores were calculated by comparing the speech naturalness score for each speech sample on each occasion for the first rater with the speech naturalness score of that same speech sample on the same occasion for every other rater in their assigned group. The direction of the difference was disregarded. Each group had a total of 5508 difference scores (36 speech samples by 153 rater comparisons) at each rating occasion. The number of ratings that achieved difference scores of +/- 0 to 8, along with cumulative percentages for each group, are presented in Table 12. Individual data for the 18 participants in each group are also presented (Appendix AF). As is traditional in stuttering speech naturalness research, the percentage of ratings with differences of +/- 1 scale value or less were used as the standard for interrater agreement data, and these data were used for further data analyses. The effect of training and rating occasion on the interrater agreement of participants' speech naturalness ratings was analyzed using a two-way (rating occasion- 4 levels x group-3 levels) RM-ANOVA. Group and rating occasion main

effects, as well as the interaction between these three variables, was tested. Because statistically significant effects were found, the percentage of ratings within +/- 1 scale value or less was analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate specific differences between the variables that had significant effects. Mean differences between groups using 95% confidence intervals are also presented (Figure 7).

As stated earlier in this chapter, mean speech naturalness ratings, inter- and intrarater agreement ratings for all three groups were assessed after rating occasions one and two and were found not to differ to a statistically significant level. When the above analysis was conducted for interrater agreement on all four rating occasions, however, a statistically significant difference was found both before and after training. The researcher and a statistical consultant compared the two data sets and located an error in the SPSS syntax used in the original ANOVA procedure for occasions one and two. This error showed the groups to be equivalent on occasions one and two when in actuality they were not. To mitigate the influence of this error on the conclusions drawn from these data, the data were re-run as explained above without the outlier (the group member with the lowest interrater agreement in each group) to determine if the lowest group member's data affected the groups' equivalence. The outliers did not affect the groups' equivalence; they were still statistically different, so the original data set with the outliers included was statistically corrected. This was done by running a RM-ANCOVA using the group difference on occasions one and two as the covariant. This procedure establishes the association between the covariant and the effect being studied and then performs the original analysis with that portion of the variance removed. Because statistically significant main and interaction effects were found, the percentage of ratings within +/- 1 scale value or less were analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate specific

differences between the variables that had significant effects. Mean differences between groups using 95% confidence intervals are also presented (Figure 8).

Because the corrected data did not change the results, the original, uncorrected percentage of ratings with differences of +/- 1 scale value or less for interrater agreement data (Table 12) were also used to investigate research question five, "Will stuttering severity (normal, mild, moderate, severe stuttering) affect listeners' interrater agreement levels?" using a three-way RM-ANOVA (stuttering severity-4 levels x treatment group-3 levels x rating occasion-4 levels). Stuttering severity, treatment group and rating occasion main effects as well as the interaction between these three variables were tested. Because main and interaction effects were statistically significant, these data were analyzed using Fischer's Least Significant Difference planned comparisons procedure to investigate specific differences between the variables that had significant effects. Mean differences between treatment groups and 95% confidence intervals for each speaker group are presented in Figures 9-12. Table 13 presents the mean, range, and standard deviation for the interrater agreement for each stuttering severity level and treatment group. Appendix AG presents individual raters' interrater agreement for each stuttering severity level. Figure 13 presents the mean interrater agreement and 95% confidence intervals for all speaker groups on all four rating occasions for all raters combined.

*Interrater reliability as ICCs.* Because the goal of the present study is to assess the relationship between the absolute values of speech naturalness ratings, as discussed in Chapter 1, ICCs are not the preferred method. In order to meet this goal, agreement scores were calculated in this study, but ICCs were also calculated to allow comparisons to speech naturalness research that uses ICCs as the main reliability metric. To address the last research question "what are the ICCs (both for the group of raters and the average individual raters) for untrained and trained

listeners when rating the speech naturalness of children who stutter at mild, moderate, and severe levels as well as normal speaking children?", rater reliability was calculated via ICCs. As discussed in Chapter 1, the ICC (R) procedure used in this study uses an ANOVA technique to determine the proportion of the total variance in a set of ratings due to the variance across the samples being rated. An R of 1.0 indicates perfect interrater reliability and an R of 0.00 indicates the total absence of agreement (Martin et al., 1984). The ICCs for the mean reliability of all 54 raters ($R_{54}$) and for the average individual rater ($R_1$) are presented as well as the single ($R_1$) and average ($R_{18}$) ICCs for each group. The entire group's ICCs were calculated using the speech naturalness ratings from the first two rating occasions for all speech samples combined for all 54 participants. The second round of ICCs was calculated for each individual group for each of the speaker groups separately (Table 14). The results of these data analyses are presented in the next chapter.

# CHAPTER 5

# STUDY TWO: RESULTS AND DISCUSSION

All participants completed all four sessions with the researcher and provided usable data. All 54 participants reported no difficulty completing the four tasks via a post study questionnaire (Appendix S).

## Speech Naturalness Ratings

Mean speech naturalness ratings by group for each speech sample on each occasion ranged from 2.01 to 8.09 (Table 8) and individual speech naturalness ratings for each rater for each speech sample on each occasion ranged from 1 to 9 (Appendix AC). A RM-ANOVA was conducted at the completion of the study (within subjects factor rating occasion- 4 levels x between subjects factor group- 3 levels) and only found a main effect for rating occasion ($F_{(3,49)}$ = 6.849, p = .001, partial eta squared = .295). When the rating occasion main effect was investigated via Fisher's Least Significant Difference Test, differences were found between rating occasions one (M=4.97) and three (M=5.27; mean difference = -.299, p = .001), two (M=5.11) and three (M=5.27; mean difference = -.159, p = .036), and three (M=5.27) and four (M=5.06; mean difference = .213, p = .001). These differences were found with all three groups combined; there were no significant between group differences.

### *Speech Naturalness Ratings and Speaker Group*

All 54 raters combined rated normal speakers the lowest (most natural; M = 2.04) and severe stutterers the highest (least natural; M = 7.85). Mild and moderate speakers fell in between, with means of 3.57 and 6.43, respectively (Table 9). The standard deviations were

smaller for the extreme speaker groups (normal and severe) than for mild and moderate speakers and this was confirmed with a one group t-test ($t_{(3)} = 10.57$, $p = .002$).

Speech naturalness ratings ranged from 1 to 4 for normal speakers, 1.7 to 5.4 for mild speakers, 4.3 to 8.4 for moderate speakers, and 6.2 to 9 for severe speakers (Table 9). Statistically significant main effects were found for severity ($F_{(3,49)} = 1022.63$, $p = .000$, partial eta squared = .984) and rating occasion ($F_{(3,49)} = 6.848$, $p = .001$, partial eta squared = .295). Statistically significant interaction effects were found for severity by group ($F(6,98) = 2.62$, $p = .021$, partial eta squared = .138), rating occasion by group ($F_{(6,98)} = 3.21$, $p = .006$, partial eta squared = .164), and severity by rating occasion by group ($F_{(18,86)} = 4.063$, $p = .000$, partial eta squared = .460). The between subject effect of group was not statistically significant ($F_{(2,51)} = 3.074$, $p = .055$, partial eta squared = .108). Fishers Least Significant Difference Test showed statistically significant differences between speaker groups. Mean differences between the speakers groups are normal and mild = -1.80, normal and moderate = -4.65, normal and severe = -5.90, mild and moderate = -2.85, mild and severe = -4.10, and moderate and severe = -1.25.

## Intrarater Agreement

The intrarater agreement (percentage of listeners with speech naturalness ratings within +/- 1 scale value of each other) for the speech naturalness of children who do and do not stutter, as rated by inexperienced listeners, ranged from 76.39% to 78.09% on rating occasions 1 and 2 and 81.64% to 85.03% on rating occasions 3 and 4 (Table 10; Figure 5).

### *Intrarater Agreement and Severity*

A three way (rating occasion comparison-2 levels x stuttering severity-4 levels x group-3 levels) RM-ANOVA procedure found significant main effects for rating occasion comparison ($F_{(1,51)} = 11.346$, $p = .001$, partial eta squared = .182) and stuttering severity ($F_{(3,49)} = 39.385$, $p =$

.000, partial eta squared = .707) along with a significant rating occasion comparison by stuttering severity interaction ($F_{(3,49)} = 3.158$, p = .033, partial eta squared = .162; Tables 10, 11 and Figure 6). Rating occasion comparison by group ($F_{(2,51)} = .436$, ns) and stuttering severity by group ($F_{(6,98)} = .950$, ns) interactions were not significant. A comparison of the two means (rating occasion comparison) using Fisher's Least Significant Difference procedure found intrarater agreement on occasions one versus two (77.42%) to be significantly lower than intrarater agreement on rating occasions three versus four (82.92%) for all samples and all raters combined. This difference was significant at the .001 level. Significant differences were found between raters' intrarater agreement levels of normal speakers and children who stutter at a mild level (88.7% versus 73.4%), normal speakers and children who stutter at a moderate level (88.7% versus 72.3%), children who stutter at mild and severe levels (73.4% versus 86.8%) and children who stutter at moderate and severe levels (72.3% versus 86.8%) across rating occasion comparisons for all 54 raters.

## Interrater Agreement

Interrater agreement ranged from 56.88% (Table 12; control group on rating occasion three) to 70.66% (training group rating occasion three) with average agreement of 62.9% for all raters across all rating occasions. A statistically significant main effect was found for rating occasion ($F_{(3,49)} = 7.579$, p = .000, partial eta squared = .317) and a statistically significant interaction effect of occasion by group ($F_{(6,98)} = 4.146$, p = .001, partial eta squared = .202). A comparison of the rating occasion means using Fisher's Least Significant Difference procedure found interrater agreement on occasions one versus two (61.1% versus 62.7%) to be statistically significantly (p = .038), and also one versus three (61.1% versus 63.9%, p = .008), one versus four (61.1% versus 65.2%, p = .000), and two versus four (62.7% versus 65.2%, p = .007) to be

statistically significant. All other comparisons were not significant. Fisher's Least Significant Difference Test was also used to investigate the group by occasion interaction and found significant differences at rating occasion one between the experimental and exposure control groups (65.2% versus 59.6%, p = .02), and the experimental and control groups (65.2% versus 58.5%, p = .010). At rating occasion two significant differences were found between the experimental and control groups (65.2% versus 58.0%, p = .007) and the exposure control and control groups (65.0% versus 58.0%, p = .008). For rating occasion three, significant differences were found between the experimental and exposure control groups (70.6% versus 63.6%, p = .008), the experimental and control groups (70.6% versus 57.6%, p = .000), and the exposure control and control groups (63.6% versus 57.6%, p = .016), while for rating occasion four the only significant difference was between the experimental and control groups (68.4% versus 62.6%, p = .019).

Another two-way RM-ANOVA was conducted (rating occasion- 4 levels x group 3 levels) with the person with the lowest interrater agreement in each group excluded from the analysis which replicated the above results with a significant main effect for rating occasion ($F_{(3,46)} = 10.334$, p = .000, partial eta squared = .403), and a statistically significant interaction effect of rating occasion by group ($F_{(6,92)} = 3.642$, p = .003, partial eta squared = .192). Fisher's Least Significant Difference Test failed to find any statistically significant differences when comparing groups by occasions using this data set. The original data set with the outliers included was statistically corrected using a two-way RM-ANCOVA (rating occasion- 4 levels x group- 3 levels) with the group difference on occasions one and two a covariant. There was not a significant main effect for rating occasion ($F_{(3,48)} = 2.759$, n.s), but there was a significant

interaction effect between rating occasion and groups ($F_{(6,96)} = 4.715$, p = .000, partial eta squared = .228; Figure 8).

*Interrater Agreement and Stuttering Severity*

Children who stutter at a severe level and normal speaking children had the highest interrater agreement (severe-64.8% to 80.7%, normal 62.2% to 81.1%). Children who stutter at mild and moderate levels were lower (mild-45.6% to 66.8%, moderate-43.4% to 74.2%; Table 13). Significant main effects for stuttering severity ($F_{(3,49)} = 162.869$, p = .000, partial eta squared = .909) and rating occasion ($F_{(3,49)} = 7.683$, p = .000, partial eta squared = .320) were found along with a significant stuttering severity by group interaction ($F_{(6,98)} = 3.547$, p = .003, partial eta squared, .178), rating occasion by group interaction ($F_{(6,98)} = 4.239$, p = .001, partial eta squared, .206), rating occasion by stuttering severity interaction ($F_{(9,43)} = 5.899$, p = .000, partial eta squared = .552), and stuttering severity by rating occasion by group interaction ($F_{(18,86)} = 6.16$, p = .000, partial eta squared = .563).

Post hoc analysis using Fisher's Least Significant Difference Test for the main effect of stuttering severity found interrater agreement for normally speaking children (71.87%) to be statistically higher than children who stutter at a mild level (51.82%, p = .000) and a moderate level (55.21%, p = .000), the interrater agreement when rating the speech naturalness of children who stutter at a mild level (51.82%) to be significantly lower than children who stutter at a moderate level (55.21%, p = .021) and children who stutter at a severe level (72.75%, p = .000; Figure 12), and interrater agreement when rating the speech naturalness of the speech of children who stutter at a moderate level (55.21%) significantly lower than the interrater agreement of children who stutter at a severe level (72.75%, p = .000).

## Intraclass Correlation Coefficients

The ICCs for the mean reliability of all 54 raters ($R_{54}$) was .946 and was .197 for the average individual rater ($R_1$), both of which are significant at the .000 level. For both the experimental and control groups, group correlations ($R_{18}$) were the highest for the normal speakers (.950 and .964, respectively) and for both of these groups children who stutter at a severe level had the lowest ICCs at .853 and .785 respectively (Table 14). The exposure control group had the highest ICCs for the severe speakers (.950) and the lowest for the mild speakers (.867).

## Study Two: Discussion and Interpretations

### *Speech Naturalness Ratings of Children Who Do and Do Not Stutter*

The first basic finding of this study was inexperienced listeners' speech naturalness ratings when rating the speech samples of normal speaking children and children who stutter at mild, moderate, and severe levels. Normal speaking children had the lowest (most natural) speech naturalness scores (M=2.04) and children who stuttered at a severe level had the highest speech naturalness scores (M=7.85). Children who stuttered at mild and moderate levels fell in between the two extremes (M=3.57 and M=6.43 respectively). This replicates findings in the adult speech naturalness literature as normal speaking adults were also rated as more natural than adults who stutter in numerous studies (see Appendix C). Martin et al., (1984) reported speech naturalness ratings of 6.52 and 5.84 for adults who stutter and adults who stutter using DAF and 2.12 for normal speaking adults. Martin and Haroldson's (1992) findings replicated this study with ratings of 6.81 and 6.04 for adults who stutter and 2.3 and 2.27 for normal speaking adults. All studies comparing the two groups replicated these results with adults who stuttered rated less natural than adults who did not stutter (Ingham et al., 1985; Ingham, Warner et al., 2006;

Mackey et al. 1997; Metz et al., 1990; O'Brian, Onslow et al., 2003; O'Brian, Packman, Onslow, Cream et al., 2003; Onslow, Hayes et al., 1992; Runyan et al., 1990; Stuart & Kalinowski, 2004; Stuart et al., 2006; Van Borsel & Eeckhout, 2008).

Although all of the research comparing normal speakers to people who actively stutter shared these results, two studies comparing recovered stutterers with normal speakers did not. Finn (1997) found speech naturalness values of 2.36 for recovered stutterers and 1.77 for normal speaking adults. Although technically normal speakers are rated more naturally than the people who used to stutter, both values are within the "normal" speech naturalness range of 3 or less (Ingham & Ingham, 2010). Overlapping speech naturalness values were also found when the speech naturalness of spontaneously recovered children and normal speaking children was rated (Finn et al., 1997). Speech-language pathologists rated recovered children's speech naturalness 3.71 whereas inexperienced listeners rated it 1.77. The speech naturalness of normal speaking children was rated 3.24 and 3.82 by speech pathologists and inexperienced listeners respectively.

The speech naturalness ratings for normal speaking children reported in the current study do coincide with the ratings reported by Coughlin-Woods et al. (2005), although the majority of the children in their study were older than the speakers used in the current study. They found a speech naturalness mean of 1.9 for 60 normal speaking children aged 8-16 with a range of 1.6-3 compared to a mean of 2.04 and a range of 1-4 in the current study with elementary aged speakers' speech samples. These values coincide with the literature reporting speech naturalness of normal speaking adults ranging from 1 (Hewat et al., 2006) to 3.6 (O'Brian, Onslow Cream, & Packman, 2003) were reported with the majority of the ratings under 3 (Finn, 1997; Ingham et al., 1985; Ingham, Sato, et al., 2001; Mackey et al., 1997, Martin & Haroldson, 1992; Martin et al., 1984; O'Brian, Packman, Onslow, Cream et al., 2003; Runyan et al., 1990; Stuart &

Kalinowski, 2004; Stuart et al., 2006; Van Borsel & Eeckhout, 2008). The study reported here provides the first speech naturalness data collected from a large group of listeners using speech samples from children who do and do not stutter. The speech naturalness ratings reported here coincide with the adult literature and complement the small body of work already completed with children (Coughlin-Woods; Finn et al.). The next research question addresses whether these speech naturalness ratings for the various speaker groups were statistically significant.

*Speech Naturalness Ratings Differentiating Speaker Groups*

The second research question investigated in this study was if inexperienced listeners' speech naturalness ratings would differentiate children who stuttered at mild, moderate, and severe levels and normal speaking children at a statistically significant level. Statistically significant differences were found between stuttering severity because mean differences between the following speakers were significant: normal speakers and children who stutter at a mild level (-1.80); normal speakers and children who stutter at a moderate level (-4.65); normal speakers and children who stutter at a severe level (-5.90); children who stutter at mild and moderate levels (-2.85); mild and severe levels (-4.10); and moderate and severe levels (-1.25). All comparisons were significant at the .000 level, providing support that this scale can be successfully used to differentiate between speaker groups in children as has been previously reported in adults (Martin et al., 1984).

In the adult speech naturalness literature, Finn (1997) found statistically significant differences between normal speaking adults and recovered adults who used to stutter. Metz et al. (1990) found a statistically significant difference between normal speaking adults and treated adults who stutter, and several studies found statistically significant differences between the speech naturalness ratings of adults who currently stutter and normal speaking adults (Ingham et

al., 1985; Mackey et al., 1997; Martin & Haroldson, 1992; Martin et al., 1984; O'Brian, Onslow, Cream, & Packman, 2003; O'Brian, Packman, Onslow, Cream et al., 2001; Onslow, Hayes et al., 1992; Runyan et al., 1990; Stuart & Kalinowski, 2004; Van Borsel & Eeckhout, 2008). One of Martin et al.,'s purposes when initially using this 1-9 scale to rate speech naturalness in adults who stutter was to determine if it would differentiate between speaker groups. They found speech naturalness when measured with this 9 point scale could discriminate between the speech of adults who stutter with and without DAF and normal speaking adults. Based on this study's results, it can also discriminate between normal speaking children and children who stutter at mild, moderate, and severe levels.

Although the current study's findings support the adult literature, the only other study measuring speech naturalness in children did not find the scale could discriminate between the two speakers groups in their study. Finn et al. (1997) did not find a significant difference between the speech of children who naturally recovered from stuttering and normal speaking children. This finding is rational, however, because none of the children in Finn et al.'s study currently stuttered and the recovered children did not receive any formal treatment that may have influenced their speech patterns in an unnatural fashion. Although this study does not provide support for this speech naturalness scales' ability to differentiate between speaker groups, it does not provide contradictory evidence for the concept either as none of the children in the study stuttered or had received any stuttering therapy. The overall conclusion based on past and current research is that this speech naturalness scale can effectively discriminate between normal speakers and speakers who stutter at various severity levels in both adults and children.

*Intrarater Agreement Variations by Stuttering Severity*

The addition of the normally speaking children to Study Two's speech samples allowed for an investigation of listeners' intrarater agreement levels when rating the speech samples of normally speaking children and children who stutter at mild, moderate, and severe levels. Research question five asks if stuttering severity (normal speakers and mild, moderate, and severe stuttering) will affect that inexperienced listeners intrarater agreement. Data supported this question as listeners' intrarater agreement when rating the "extreme" speech samples (normal speaking children and children who stutter at a severe level; 87.4% and 86.3%, respectively on rating occasion comparison 1 versus 2 and 90% and 87.4% on rating occasion comparison 3 versus 4) was significantly higher than when rating the speech samples from children who stutter and mild and moderate levels (70.3% and 66.2%, on rating occasion comparison 1 versus 2 and 76.5% and 78.3% on rating occasion comparison 3 versus 4, respectively). This finding is logical as well as supported by data; it is easier to agree with yourself when rating the more extreme cases than the ones that do not have as many polarizing speech patterns or no abnormal speech issues. The intrarater agreement levels did not approach 80% for mild and moderate speakers on the first two rating occasions but exceeded 80% for normal and severe speakers. These results warrant further investigations into ways to train listeners to improve their intrarater agreement when rating speech naturalness at least with some speaker groups. Future research also needs to be conducted with matched speech samples from these speaker groups to determine if the current study's results would be replicated when the speech samples from the various speaker groups were matched on age, gender, speech rate, mean length of utterance and percent consonants correct.

As shown in Figure 6, intrarater agreement was higher for the second rating occasion comparison versus the first for all stuttering severity levels. Listeners' intrarater agreement increased the least for the extreme speakers; normal speakers (+2.6% between the two rating comparisons) and speakers who stutter at a severe level (+1.1% between the two rating occasions). Listeners' intrarater agreement when rating the speech samples of children who stutter at a moderate level increased the most (+12.1%) and listeners' agreement levels when rating the speech samples of children who stutter at a mild level were between the two extremes (+6.2%). It should be noted that even at their highest, the intrarater agreement for children who stutter at mild and moderate levels was almost 10 percentage points lower than the starting agreement for normal and severe speakers. As discussed above, this finding is logical, but has not been replicated in the limited adult research investigating stuttering severity and intrarater agreement of speech naturalness.

The only study in the adult speech naturalness literature that directly investigated listeners' intrarater agreement when rating speech naturalness in adults who stuttered with differing severity levels was Kalinowski et al. (1994). In this study, the researchers had 32 undergraduate listeners rate the speech naturalness of adults who stuttered at mild and severe levels both before and after treatment with the Precision Fluency Shaping Program. The listeners rerated the samples one week after their initial ratings. Listeners' intrarater agreement when rating the speech naturalness of adults who stuttered at a mild level pre and post therapy was 89% and 86%, respectively. For the severe speakers, listeners' intrarater agreement was 83% and 89%. When compared to the intrarater agreement levels in the present study, Kalinowski et al.'s data coincide with the values found for normal speakers and children who stutter at a severe level, but their adult intrarater agreement for people who stutter at a mild level

are 10-19 percentage points higher than the values reported here. Several major differences between the studies may account for this. First, the Kalinowski et al. study used audiotaped samples when having their listeners rate speech naturalness whereas the current study used audiovisual samples. As discussed in Chapter 1, this has been shown to make a difference in raters' agreement levels in previous stuttering speech naturalness research (Martin & Haroldson, 1992). Secondly, where the current study used 30-s conversational speech samples to have listeners rate speech naturalness, the adult study used 1-min reading passages. No study has directly investigated the difference between the speech naturalness ratings collected from spontaneous speech and reading samples, but it is commonly recommended by stuttering researchers that other stuttering treatment outcome variables be measured using various types of speech samples as these variables have been known to be effected by speech sample type (Bothe, Davidow, Bramlett, & Ingham, 2006). Based on this information, it is safe to assume speech naturalness and agreement levels associated with it may change based on type of speech sample as well. Although type of speech sample has not been directly investigated, speech sample length has and as reviewed in Chapter 1, has been shown to affect listeners' speech naturalness agreement (Onslow, Adams, & Ingham, 1992). The difference in speech sample length may account for differences between the two studies. Next, the current study had listeners rate the speech naturalness of normal speaking children and children who stutter at all three severity levels. The introduction of not only normal speakers, but also a larger variety of speakers who stuttered, may have impacted intrarater agreement data. Finally, whereas the listeners in the current study rated speech naturalness independently (each at her own computer with noise isolating headphones), in the adult study raters listened to the speech samples through amplified speakers in small groups rather than with headphones. Ambient room noise may have affected

the results found by Kalinowski et al. Neither study reported matching speech samples between groups, a limitation to the current study discussed in more detail below. This may have affected both studies' results. Any or none of these study design differences may account for the intrarater agreement differences between the two studies and could be investigated in future research to start determining what variables do effect the intrarater agreement of listeners' speech naturalness ratings.

*Interrater Agreement Variations by Stuttering Severity*

As with intrarater agreement, the addition of normally speaking children and three additional speech samples from children who stutter at a mild level, allowed for the investigation of interrater agreement by stuttering severity (normally speaking children and children who stutter at mild, moderate, and severe levels). Also similar to intrarater agreement, the interrater agreement values were significantly higher for the extreme speech samples (normal speakers 71.9% and children who stutter and a severe level 72.8%) than for the speech samples of children who stutter at mild (51.8%) and moderate levels (55.2%). As discussed above, this finding is logical as typically it is easier for raters to agree on samples with marked stuttering features (as in the severe samples) or samples with no stuttering features (as in the normal speakers) than samples with less well defined speech characteristics.

Although the values reported above are significantly different from one another, it is important to note that none of them, when reported as all 54 raters combined, for any speaker group on any of the four rating occasions approaches the common standard of accepted interrater agreement of 80% (see Figure 13). The highest interrater agreement was for severe speakers at rating occasion two (76.04%). The next highest agreement rating was for normal speakers on rating occasion four (74.87%) and the lowest interrater agreement was found for mild speakers at

rating occasion two (48.78%). When addressing the interrater agreement for each stuttering severity level by each group on all four rating occasions (see Figure 12), the training group did have interrater agreement approaching 80% when rating severe speakers on three of the four rating occasions, with the only exception being rating occasion four (72.26%). The exposure control group also had interrater agreement approaching 80% at rating occasion two for severe speakers (77.19%), but it dropped on the other three occasions (68.92%-71.02%). This group's highest agreement was for normal speakers in which three of the four occasions were 74%, 76%, and 81% for occasions two through four, respectively (Figure 9). The control group's highest interrater agreement was for normal speakers as well, but it only approached 80% for occasion four (76.1%). On the other three occasions, it was 71.59% or less. When rating the mild and moderate speakers, the experimental group had the highest ratings at 48% to 74% whereas the two control groups mean interrater agreement ratings were much lower at 43% to 55% (Figures 10 and 11).

In the adult speech naturalness literature, the studies comparing people who do and do not stutter found higher interrater agreement values than those reported in the current study and some reported higher values for normal speaking adults. The interrater agreement of listeners who rated the speech samples of normal speaking children ranged from 62.2% to 76.1% (see Table 13) whereas in the adult literature they were 80% to 81% (Mackey et al., 1997; Martin & Haroldson, 1992) and 75% (Martin et al., 1984). The only study in the speech naturalness literature with adults that directly compared adults who stutter at mild and severe levels (Kalinowski et al., 1994) found higher interrater agreement for the mild speakers (67% pretreatment) than the values reported here (48% to 56%), but lower values for the severe speakers (60% pretreatment) than the current study (71% to 76%). Limitations of the current

study's design related to this and all hypotheses that may have affected the results will be discussed in the next chapter after a discussion of ICCs.

*ICCs of the Average Individual Rater and Groups of Raters When Rating Speech Naturalness in Children Who Do and Do Not Stutter*

ICCs for all 54 raters for all speech samples combined were .946 for the group and .197 for the individual. Group ICCs for normal speakers were the highest for the training group and control group at .950 or above and for both of these listener groups ICCs decreased for mild, moderate, and severe speakers respectively (Table 14). ICCS for the exposure control group were highest for the severe speakers (.950) with moderate (.949) and normal (.927) in between and mild the lowest (.867). The lowest ICC reported in the current study was for the control group's ratings of speech naturalness in severe speakers (.785), clearly still in the excellent range according to Fleiss (1986). For speech naturalness studies that only used adults who stuttered as speakers, group ICCs were reported in the same range as the ones in the current study at .90 or greater (see Appendix G; Armson & Kiefte, 2008; Martin et al., 1984; Martin & Haroldson, 1992; Packman et al., 1994; Tasko et al., 2007). In studies that combined adults who do and do not stutter when calculating ICCs, the values were .95 or greater (Metz et al., 1990; Schiavetti et al., 1994; Tasko et al., 2007). The ICCs reported in the adult literature are equivalent to .946 in this study. For normal speaking adults, ICCs were reported as .74-.77 (Martin & Haroldson, 1992; Martin et al., 1984), clearly lower than the values of .950, .927, and .964 found in the current study. In the only study that directly compared the speech naturalness of adults who stutter at mild and severe levels, ICCs of .93 to .99 for mild speakers and .97 to .89 for severe speakers were reported (Kalinowski et al., 1994). These values also coincide with the ICCs for

the children who stutter at a severe level reported here (.853, .950, .785, for the three listener groups, respectively).

As has also been reported throughout the adult speech naturalness literature, group ICCs were higher than ICCs for the average individual rater. The current study had individual ICCs of .197 for all raters and all speaker groups combined. Equivalent adult literature found ratings of .695 (Metz et al., 1990) to .935 (Schiavetti et al., 1994), obviously higher than the current study's values. Studies reporting individual ICCs for only people who stutter ranged from .34 (for 30-s speech samples, Onslow, Adams, & Ingham, 1992) to .91 (Packman et al., 1994). Speakers under the influence of DAF had ICCs of .57 (Martin et al., 1984). All of these values are higher than the individual ICCs for children who stutter at all three severity levels for the training group and for mild speakers in the exposure control group and severe speakers in the control group. For normal speaking adults, individual rater ICC values were reported of .10 (Martin et al.) to .16 (Martin & Haroldson, 1992). These are clearly much lower than the values of .346, .260, and .424 found in the current study. Corresponding with the adult literature, ICC values, for the group of raters was highly reliable, but the reliability of the individual raters did not meet the "fair" minimum value (Fleiss, 1986), with the exception of the control group when rating normal speaking children. As has often been recommended in the adult speech naturalness literature, based on ICCs, speech naturalness should be measured by groups of raters when being used in research studies, but could be considered applicable with fewer raters for single subject design studies and clinical use (Martin et al.).

ICCs and interrater agreement are two types of interrater reliability (see Chapter 1), but as can be seen in this study, their values are not equivalent. When comparing group ICCs to group interrater agreement, ICC values are much higher at .946 than interrater agreement, which

ranges from 56.88% to 70.66% (Table 12). If ICCs were the only reliability metric calculated in the present study, it would appear that Martin et al.'s (1984) 9-point scale is very reliable when measuring speech naturalness in children who do and do not stutter, with speech samples accounting for 94.6% of the variance in speech naturalness ratings. When using interrater agreement as the reliability metric, the reliability of this scale is not as strong because the highest interrater agreement (percent of comparisons between raters with scale value differences of +/- 1 scale value or less) is only 70.66% (Table 12). Using the more stringent criterion of the percentage of raters with identical speech naturalness ratings, interrater agreement is only 24.98% to 30.27%. The comparison of these two reliability methods support the literature reviewed in Chapter 1 that ICCs, because they assess the relationship of ratings to one another rather than absolute ratings, are better suited for classical test theory research versus observational research. A discussion of the other research questions and their relation to both studies and the adult speech naturalness literature as well both studies limitations will follow in the final chapter.

# CHAPTER 6

# GENERAL DISCUSSION

The two studies presented in this dissertation were designed to investigate inexperienced listeners' speech naturalness ratings and the intrarater and interrater agreement of their speech naturalness ratings when rating the speech of children who do and do not stutter. The effect of the *SMS* speech naturalness training program on raters' agreement levels was also investigated by comparing the speech naturalness ratings of a training group that learned to rate speech naturalness using the *SMS* program, an exposure control group that was exposed to the *SMS* training samples but were not given feedback regarding them, and a true control group that did nothing with the *SMS*. The remainder of this chapter provides a discussion of the results presented in Chapters 3 and 5 and the implications of these findings. The chapter concludes with a discussion of the limitations of the present studies and future research directions given the current studies' findings and limitations.

## Intrarater Agreement Values After Rating Occasions One and Two

The two studies presented in Chapters 2 through 5, when combined, provide the opportunity to assess multiple estimates of the above variable. The percentage of untrained judges who showed agreement levels of +/- 1 scale value for intrarater agreement, for example, fell between 74.11% and 78.09% for all six groups (training, exposure control, and control in both studies). These values are lower than some of the intrarater agreement values reported in the adult literature for groups of listeners rating the speech naturalness of adults who stutter (see Appendix E; Kalinowski et al., 1994; Martin et al., 1984). The lowest of the values obtained in

these studies are also lower than those reported by Finn (1997) with recovered stutterers and normal speakers providing the speech samples, but were also higher than other studies reporting intrarater agreement of speech naturalness in adults who do stutter (Martin & Haroldson, 1992; Onslow, Adams, & Ingham, 1992; Stuart & Kalinowski, 2004). The values reported in the current studies are also higher than the only group study reporting the intrarater agreement of children who recovered from stuttering without formal treatment and normally speaking children (Finn et al., 1997).

The intrarater agreement reported in these studies parallels values reported in the adult literature when listeners are not given any specialized training to use the speech naturalness measure. It is interesting to note that in two completely unrelated samples of inexperienced college students in the current study (two years separated the groups therefore none of them took classes together), the intrarater agreement levels on rating occasion comparison 1 and 2 were very close. The initial intrarater agreement for all three groups combined was 75.3% for Study One and 77.4% in Study Two. Although there is little more than two percentage points between the combined groups, the actual starting difference was 3.4% when comparing the training groups for the two studies (74.4% One and 77.4% Two), 4% for the exposure control groups (74.1% One and 78.1% Two), and 1% between the two control groups (77.4% One and 76.4% Two). Although all of the values are very close (less than 4% difference), it is of interest to note that the training and exposure control groups started with higher intrarater agreement in Study Two, whereas the control group subjects started higher (minimally, but higher) in Study One. The intrarater agreement values on rating occasion comparison three versus four were also similar between the two studies for the control groups. The exposure control groups for both studies had intrarater agreement values on rating occasions three versus four of 80.4% in Study

One and 81.6% in Study Two. Less than a percentage point difference between the two was found when comparing rating occasions three and four for the control groups as well (81.5% One versus 82.1% Two). This is worth noting because Study Two included normally speaking children, which, according to classical test theory, could have possibly inflated the agreement ratings (Crocker & Algina, 1986). According to classical test theory, the more heterogeneous the experimental samples used (in this case speech samples), typically the higher the reliability because it is easier to identify differences in more varied samples than in samples that are similar. Using this model, one would assume the initial intrarater agreement in Study Two would be higher than that found in Study One because the samples in Study Two are more varied (normal and stuttering) when compared to stuttering only samples that were used in Study One. The intrarater agreement data from these two studies does not appear to support this theory because the agreement values between the two studies are so similar.

### The Effect of *SMS* Training on Intrarater Agreement

Another goal of these studies was to assess the effects of a specific training program on listeners' intrarater agreement when rating speech naturalness. In Study One the training group (92.78%) was found to have significantly higher intrarater agreement on rating occasion comparison three versus four than the two control groups (80.36% and 81.55%; Figure 1). The data presented in Chapter 5 for Study Two, however, do not provide support for this question as no difference between the three listener groups was found on rating occasion comparison three versus four (see Figure 5). Although the intrarater agreement values for both rating occasion comparisons for the control groups were similar, the major difference between the two studies was the post training intrarater agreement levels for the training groups. In Study One, post training intrarater agreement values were 92.8%, which was a significant improvement over the

groups 74.4% starting value, as well as the two control groups values on rating occasion comparison 3 versus 4 (80.4% and 81.5%). For Study Two, however, the training groups post training intrarater agreement was lower, 85%, which was not a statistical improvement over the groups' pretraining agreement of 77.8% or over the two training groups ratings on rating comparison three versus four of 81.6% and 82.1% ($F_{(2,51)} = .428$, n.s.). In Study Two, there was a significant main effect between rating occasion comparisons (one versus two and three versus four) as the first comparison was statistically lower than the second, but this difference was not seen by treatment group as all three groups improved on the second rating comparison.

The obvious question becomes then, what could account for the contrasting findings between the two studies? The first step in investigating this question is to assess the effect of outliers on intrarater agreement. This was investigated and was found to make no difference in regard to the above findings. Because outliers did not affect results, other possible explanations are discussed below. First, although both studies used the *SMS* speech naturalness training program and Criterion Test to train listeners to rate speech naturalness, Study One participants completed the program independently as a self study (as is typical when used by students and clinicians), whereas Study Two participants completed the training program with the researcher present in the stuttering lab at The University of Georgia. This difference in task administration may have contributed to the difference in the two studies' results. All participants in Study One reported completing the tasks independently and as explicitly directed, but other than the self report completed by participants upon completion of the study, the researcher has no way of verifying this information.

Another possible difference that could explain the variation in intrarater agreement between the two studies is that 21 people in Study One (48.8%) reported knowing someone who

stuttered, but the relationship to the people they knew was not reported. If participants went to elementary school with a student who stuttered, but has not known any one since that time this knowledge probably does not impact the study's results. If, on the other hand, the person's spouse, parent, teacher, or sibling stutters, then her experience with stuttering would be extensive and this could quite possibly affect the study's results. This was controlled for in Study Two as anyone with immediate family members, spouses, close friends, or teachers who stuttered were excluded from participation. Whereas excluding people with direct experience with stuttering strengthened the study's methods, it may have affected the results as maybe people who have experience with stuttering respond more positively to the training than do people who have no direct experience with stuttering. This is only one potential hypothesis to explain the differences between the studies' results, but future research could be designed to investigate this issue by assessing the effects of training on speech naturalness and agreement ratings of listeners with no experience with stuttering when compared to people with a direct connection with stuttering (i.e. the family members of people who stutter) and speech-language pathologists who not only have academic and clinical training in stuttering, but also work with people who stutter.

One issue differentiating the two studies that is expected to affect listeners' intrarater agreement when rating the speech naturalness of children who do and do not stutter is the timing of listener reratings. All participants in both studies rated 100% of the speech samples on each rating occasion, but listeners in Study One rerated the samples during the same session whereas listeners in Study Two rerated the samples 7-14 days later. Although one would expect the rerating timing to make a difference in listeners' intrarater agreement levels, it does not appear to as the initial intrarater agreement for the two studies differed by only two percentage points. In the adult speech naturalness literature, studies that had listeners complete reratings during the

same session had the lowest intrarater agreement (47% audio only, Martin & Haroldson, 1992), but ratings of 62% were also reported (Martin & Haroldson audio visual) as well as a high of 100% (O'Brian, Onslow et al., 2003). It's important to note in the O'Brian et al. study that only three raters were used to rate naturalness, and as discussed in Chapter 1 it is easier for a lower number of raters to agree with each other than when a larger number of raters are used. Also, they only had raters rerate 16.6% of the samples for people who stuttered and 5.5% for people who did not stutter. As also discussed in Chapter 1, the low number of samples rerated may have affected the agreement scores because rerating less than 17% of the speech samples is not enough to be representative of the entire corpus of the original speech samples. If a different 17% had been chosen for listeners to rerate, intrarater agreement may have been completely different. Stuart & Kalinowski (2004) was the only adult speech naturalness study that had listeners rerate 100% of the speech samples during the same session and they reported intrarater agreement at 70%. This is slightly lower than the intrarater agreement reported in Study One, but is still comparable.

In the adult speech naturalness literature, the studies that had listeners rerate speech samples 1-3 weeks after the initial rating had the highest intrarater agreement values ranging from 76% (Mackey et al., 1997; Onslow, Hayes et al., 1992) to 94.3% (Onslow et al., 1996). The Onslow et al. (1996) study only used 2 raters possibly inflating the intrarater agreement, as discussed above. In other studies with rerates in the same time frame using 10 or more raters, intrarater agreement was at least 83% (Finn, 1997; Kalinowski et al., 1994). Overall, the majority of these studies reported higher intrarater agreement than the values reported in the current studies (see Appendix E). The only study that had listeners rerate samples 3 or more weeks after the initial rating reported intrarater agreement levels of 65.6% and 75.2% (Onslow,

Adams, & Ingham, 1992). Based on the literature reviewed above (and in more detail in Chapter 1 and Appendix E), one would expect the starting values to be lower for Study One participants than the Study Two participants and technically they were, but only by 2%, not a remarkable amount. From the data presented here, timing of reratings does not appear to have a significant effect on intrarater agreement, but to have a better understanding of this issue, it would have to be scientifically investigated in future studies as no one study has directly addressed this issue in speech naturalness research to date.

Another possible explanation for the differences found between the intrarater agreement of the two studies is the inclusion of normal speakers in Study Two, which may have had an effect on the results according to classical test theory. This could affect the study's results because it is theoretically easier to agree with yourself when rating samples that vary markedly from one another (i.e. discriminating the naturalness of normal speakers versus severe stutterers) than when rating similar samples (discriminating the difference between people who stutter at mild and moderate levels; Crocker & Algina, 1986). If this were the case however, one would expect the starting intrarater agreement values to be higher for the Study Two participants than the Study One participants and that was not the case; they were equivalent. Because the intrarater agreement between the two studies on rating occasion comparisons one and two are so similar, the possibility that the inclusion of normal speakers affected intrarater agreement seems unlikely, but the only way to be sure is for the idea to be scientifically investigated. Future research designed to control for other variables outside of the inclusion of normal speech samples could provide data addressing whether normal speakers make a difference in listeners' intrarater agreement. A final point to address here is that although a discussion of possible reasons why *SMS* training worked in Study One, but did not in the follow up study is warranted,

some may argue that the entire discussion is fruitless as intrarater agreement approaches 80% prior to any attempt at training. Although this argument is valid, as the next research question addresses, the intrarater agreement for all speaker groups did not approach 80% and therefore some sort of training to increase intrarater agreement in at least some speaker groups is of value.

Interrater Agreement Values After Rating Occasions One and Two

The initial interrater agreement reported in Study Two is roughly 10 percentage points higher than the interrater agreement reported in Study One. The interrater agreement for all 54 raters on rating occasion one was 64.4% and was 62.7% for rating occasion two. When presented by group, the training group's interrater agreement was 65.2%, the exposure control group's was 59.6%, and the control group's was 58.5%. For occasion two, the three groups' interrater agreement was 65.2%, 65%, and 58% for the three groups respectively. As reported in Chapter 3, interrater agreement for all 43 raters combined was 53.7% on occasion one and 58.3% for the training group, 60.5% for the exposure control group, and 52.5% for the control group. On occasion two, the interrater agreement for all 43 raters was 54.7% and the agreement for the three separate groups' was 60.3%, 53.8%, and 50.1%, respectively. As discussed above, one possible explanation for the difference between the two studies interrater agreement is the addition of normal speaking children to the Study Two speech samples as this addition may have inflated the interrater agreement values in Study Two as is theorized in classical test theory (Crocker & Algina, 1986).

When comparing the values obtained in Study Two to the adult literature, the dissertation interrater agreement values are lower than other studies that used a group of raters to rate speech naturalness in adult speakers (see Appendix D; Hewat et al., 2006; Martin et al., 1984; Martin & Haroldson, 1992). Study Two's agreement is higher than the agreement from adult speech

samples at 30-s intervals for sophisticated and inexperienced listeners, sophisticated listeners' 15-s intervals and inexperienced listeners' 60-s intervals (Onslow, Adams, & Ingham, 1992), but is similar to the agreement reported in several adult studies (Finn et al., 1997; Kalinowski et al., 1994; Mackey et al., 1997; Onslow, Adams, and Ingham, 1992, for sophisticated listeners at 60 s and inexperienced listeners at 15 s only). In the only other study rating speech naturalness in children that reported interrater agreement in this manner, Finn et al. (1997), reported values of 45.1% when speech-language pathologists rated speech naturalness and 40.6% when inexperienced listeners rated the concept. These values are very similar to Study Two's interrater agreement and when taken together, provide a rather clear picture that interrater agreement when rating the speech naturalness of children who do and do not stutter is indeed less than 80%. The current studies' initial interrater agreement values, along with the literature review were the catalyst for the investigation of the *SMS* speech naturalness training discussed further below.

### The Effect of *SMS* Training on Interrater Agreement

When comparing Study Two to Study One, the first major difference between the two is that *SMS* training was shown to be effective at improving rater agreement in Study One, whereas it was not shown to be effective in Study Two. As discussed in the intrarater agreement section above, possible explanations for this are; administration methods for the *SMS*, participants' experience with a person who stutters, and the inclusion of normal speakers in Study Two. The point must be made that although training was shown to be "effective" in Study One, interrater agreement only improved from 58.29% to 65.24% before and after training with the *SMS* program. Even though this change is statistically significant, the clinical significance of it is questionable because the value is still well below the typically agreed upon standard in speech

naturalness research of 80%. Additionally, the training group in Study Two had higher interrater agreement post training (70.6%) than Study One's training group, but it was not a statistically significant improvement as the group's interrater agreement started at 65.20%. Also, the interrater agreement of both control groups on occasions three and four in Study Two was higher than Study One's training group on occasions three and four (see Figures 2 and 7). These results indicate that although the *SMS* was deemed effective in the Study One because the treatment group's ratings did increase to a statistically significant level after training, the interrater agreement for all three groups in Study Two was higher on rating occasions three and four than the training group in Study One. Possible explanations for the higher interrater agreement in Study Two are discussed further below.

When comparing the data presented above in Figure 7 to the Study One values presented in Figure 2 it can be seen that interrater agreement for occasions one and two in Study Two is eight to ten percentage points higher than the Study One data (64.43% versus 53.74% for all raters combined on occasion one and 62.73% versus 54.66% on occasion two). Unlike the two studies' intrarater agreement in which the initial values were almost identical, these differed by a significant amount. Several explanations for this difference are possible. First, as discussed above in regard to intrarater agreement, 21 people in Study One reported knowing someone who stuttered whereas this was controlled for in Study Two as people with immediate family members, spouses, close friends, or teachers who stuttered were excluded from participation. Based on prior research in the adult speech naturalness literature presented in Chapter 1, experience with stuttering does not typically effect agreement ratings when rating speech naturalness in adults who stutter (Onslow, Adams, & Ingham, 1992) or children who recovered from stuttering without treatment when compared to normal speaking children (Finn et al.,

1997), but this difference could have had an effect in the current studies. As discussed above, the difference in the way the *SMS* was administered, as a self-study in Study One and researcher administered in Study Two, could have affected the results and attributed to the differences found in the two studies. No previous research has been located investigating this issue, therefore outside of a follow-up study designed to address the effect of *SMS* administration on agreement values, it can not be determined for sure if *SMS* administration method affected interrater agreement. Another possible explanation for the difference in interrater agreement found in the two studies is the inclusion of normal speakers in Study Two. The normal speakers may have introduced "anchoring" into raters' speech naturalness ratings, which may have affected their interrater agreement.

Anchoring is a judgmental bias that occurs on an unconscious level (Verdantam, 2006) whereby ratings are assimilated to the starting point of the judge's deliberations (Qu, Zhou, & Luo, 2008). Anchoring bias is, therefore, the effect of these uninformative anchoring numbers on raters' judgments (Brewer & Chapman, 2002). In other words, when answering a question, whatever starting point the rater has plays a powerful role in determining the answer they determine to be correct (Tversky & Kahneman, 1974). In speech naturalness studies comparing the speech samples of people who stutter to people who do not stutter, the latter group's speech samples are considered the "norm" or anchor value that the former group's naturalness ratings are assimilated into. When comparing the two groups, this adjustment would cause the 9-point speech naturalness scale to become smaller for the speech samples of people who stutter because the naturalness ratings from the speech samples of people who do not stutter would comprise the lower, more natural end of the scale and would leave fewer scale values to rate the disordered speech samples. When these two groups are compared, any stuttered speech may sound more

unnatural when compared to a fluent, natural sounding speaker than it would when compared to other stuttered speech. According to classical test theory, the higher interrater agreement scores may be due to the inclusion of normal speakers because this inclusion could have inflated the Study Two data because it included a broader range of speech samples (it included speech samples from children who stuttered and children who did not stutter together), whereas Study One used a more homogeneous sample by not including the normal speakers (Crocker & Algina, 1986). Although this could have been the explanation for the interrater agreement differences found between the two studies, this can not be assumed without a single study examining the issue of the differences between interrater agreements based on the inclusion/exclusion of normal speakers. Limitations of the two studies are presented next with the chapter ending with a discussion of future research based on the findings of the two studies presented here.

## Limitations of the Present Studies

Like all research conducted these two studies have limitations. First, the speech samples used in the studies were not matched therefore other variables between them such as age, gender, or speech rate may have affected the results. Because a convenience sample of prerecorded speech samples was used for the children who stuttered, the specific age information was not available to the researcher. This limitation should be kept in mind when addressing the findings presented in these pages. Another limitation related to the speech samples used in Study Two was the difference in video quality between the videos of the children who stuttered and the normal speaking children. Although the researcher made every effort to degrade the quality of the normal speech samples collected 10+ years after the samples from the children who stutter to equate video quality, the difference between them was still apparent to at least a couple of participants as reported via informal conversations after the conclusion of the study. This ability

to discriminate between the normal speaking children and the children who stuttered based on video quality alone provides a confounding variable that may have affected the study's results. It can be controlled for in future research by collecting new samples from children who do and do not stutter using the same recording equipment.

One limitation of both studies is that the *SMS* program uses mainly adult speakers to train listeners to rate speech naturalness, whereas both studies assessed raters' ability to rate speech naturalness in children. Study Two was conducted based on Study One's results in which the *SMS* not only appeared to improve both intrarater and interrater agreement levels for listeners rating speech naturalness in children, but also provided participants' comments from a post study questionnaire in which the majority reported having no difficulty transitioning from rating mostly adults in the training program to children in the experimental samples (see Chapter 3). The ineffectiveness of the *SMS* in changing raters' intrarater and interrater agreement levels may be at least partially due to the fact that the training population was predominantly adults and the experimental samples were children. Future research to address this and other issues is presented below.

## Future Research

The studies presented in this document were the first to have a large group of listeners rate the speech naturalness of children who currently stutter and in Study Two normal speaking children as well. Study Two provided mean speech naturalness ratings for children who stutter at all three severity levels and normal speaking children, as well as rater agreement and ICC reliability data. Intrarater agreement was found to exceed the accepted 80% for normal and severe speakers and approach acceptable for mild and moderate speakers for both rating occasion comparisons. Interrater agreement was found to be low for all speaker groups. A line of

research stemming from these findings is to replicate this study using speech samples matched by gender, age, speech rate, percent consonants correct, and mean length of utterance collected using the same recording equipment to control for variations in video quality. This study would not only provide speech naturalness, agreement, and reliability data, as the current studies do, but would also provide the ability to correlate speech naturalness with gender, age, speech rate, stuttering frequency and severity variables to begin the process of determining what other speech outcome variables have significant relationships with speech naturalness. This study would be the first step in a line of research attempting to identify variables that affect speech naturalness in an effort to determine an external definition of speech naturalness. If this definition could be scientifically developed, it may provide the piece needed for raters to reliability rate speech naturalness without the need for formal training.

Future research can use the results of the current studies to design follow up studies in an effort to determine the effectiveness of the *SMS* for training listeners to rate speech naturalness. One relevant question is, can the speech naturalness portion of the *SMS* be isolated to train listeners to rate speech naturalness? One possible explanation for its lack of effectiveness in Study Two is that maybe speech naturalness training needs to be conducted in the context of the entire *SMS* program rather than isolated as was done in these studies. This could be investigated by having two groups of matched participants complete speech naturalness training using the *SMS*, one being trained with only the speech naturalness portion as was done here and one by completing the entire *SMS* program, to assess the effect of the *SMS* on listeners' speech naturalness scores and agreement ratings when rating the speech naturalness of people who stutter.

Another future study related to the *SMS* is to replicate Study Two using adult speech samples in an effort to determine if the current *SMS* training program works when listeners rate the speech samples of adults versus children. This study would also provide data regarding the effect of speaker severity on speech naturalness and rater agreement in adults because only one study has investigated severity with only mild and severe people who stutter providing speech samples. Depending on the results of this future study, either the current version of the *SMS* will be shown to work, this providing support for the development of a *SMS* version for children, or the *SMS* will be shown not to work in children or adults and the development of another training program would be worth investigating (or research investigating an exact definition of speech naturalness as discussed above). One possibility for an alternative training program would be related to the work done in stuttering interval identification. Researchers in this area had reliable stuttering experts create a corpus of "normed" samples to identify stuttering. These normed samples were used to create a training program to train raters to identify stuttering when speech samples were presented as intervals of differing lengths. The same model could be used to develop speech naturalness training. Videos of children and adults who stutter could be judged by stuttering experts with at least 80% intrarater and interrater agreement when rating speech naturalness. These videos and expert naturalness ratings collected in this follow up study could then be used to train listeners to rate speech naturalness.

A final related study would be to investigate the effect of listener experience when rating speech naturalness using clinicians, parents of children who stutter, and inexperienced listeners. Although this has been done with adults and children who recovered from stuttering without treatment using trained speech-language pathologists and inexperienced listeners, it has not been done with children who currently stutter, nor has the parental group been introduced. This study

would address the effect of experience with stuttering both from a clinical and personal perspective. It could also be designed with adults who stutter, their significant others, speech-language pathologists and inexperienced listeners using speech samples taken from the adults who stuttered. These studies would begin the task of determining in which listener and speaker population this measure is best suited and to determine if listeners with various levels of experience have differing speech naturalness agreement when rating speech samples using Martin et al.'s (1984) 9-point scale.

## Summary

The studies presented in these pages provide initial data from inexperienced listeners using Martin et al.'s (1984) 9-point speech naturalness scale to rate the speech naturalness of children who do and do not stutter and their agreement levels when rating speech naturalness. They also investigated the effect of the *SMS* training program on these ratings. The effectiveness of the *SMS* naturalness rating training program is inconclusive at this time because data supporting its effectiveness were reported in Study One, but it was not shown to be effective in Study Two. Possible explanations for this as well as future research addressing identified issues are discussed above. With the addition of normal speaking children in Study Two, the relationship of stuttering severity and speech naturalness ratings and agreement data became apparent. Raters had higher agreement for normal speakers and children who stuttered at a severe level, when compared to children who stuttered at mild and moderate severity levels. Replication studies as discussed above need to be conducted to isolate variables that may have affected the results presented in this document. Future research with adults and children based on the current studies' results to address these issues is planned. This follow up research is

needed before this speech naturalness measure can be recommended for use with children who do and do not stutter.

Table 1
Characteristics of the speakers whose speech samples comprised the experimental stimuli. Data taken from 4-minute audiovisual, recorded speech samples.

| Speaker | Gender | Severity[a] | MLU[b] | PCC[c] | %SS[d] | SPM[e] | # SS[f] |
|---|---|---|---|---|---|---|---|
| NK1* | Female | Normal | 10.3 | 100 | 0 | 161.8 | 5 |
| NK2* | Male | Normal | 8.8 | 100 | 0 | 131.1 | 2 |
| NK4* | Male | Normal | 8.8 | 99.15 | 0 | 99.0 | 2 |
| 4 | Female | Mild | 9.2 | 91.0 | 1.7 | 121.7 | 2** |
| 9 | Male | Mild | 10.1 | 100 | 4.2 | 105.2 | 2 |
| 11 | Female | Mild | 9.9 | 100 | 2.1 | 109.8 | 5*** |
| 6 | Female | Moderate | 8.4 | 97.2 | 8.7 | 118.7 | 3 |
| 7 | Female | Moderate | 8.0 | 97.5 | 11.3 | 105.2 | 2 |
| 14 | Male | Moderate | 8.5 | 100 | 11.2 | 147.3 | 4 |
| 5 | Male | Severe | 9.0 | 97.1 | 18.4 | 81.4 | 1 |
| 8 | Male | Severe | 9.4 | 100 | 17.5 | 111.1 | 6 |
| 13 | Male | Severe | 8.0 | 98.8 | 14.7 | 93.5 | 2 |

[a] Rated on a three point scale of mild, moderate, and severe and agreed upon by 2 out of 3 experienced judges.
[b] Mean Length of Utterance
[c] Percent Consonants Correct.
[d] Percent Syllables Stuttered
[e] Syllables Per Minute
[f] # of speaker's 30-s speech samples used
[b-e] average count from two trained undergraduate students with interrater agreement > 80%

*only included in Study Two

**only 1 included in Study One

***only 3 included in Study One

Table 2
Stuttering frequency and speech rate data for speech samples included in studies.

| Sample | %SS Judge 1 | %SS Judge 2 | SPM Judge 1 | SPM Judge 2 |
|---|---|---|---|---|
| 4.1 | 1.45 | 1.35 | 138 | 148 |
| 4.3* | 1.36 | 1.40 | 142 | 151 |
| 5.2 | 11.11 | 10.9 | 90 | 92 |
| 6.1 | 12.5 | 11.76 | 112 | 102 |
| 6.4 | 8.45 | 10.91 | 142 | 110 |
| 6.5 | 21.95 | 22.58 | 82 | 62 |
| 7.2 | 12.28 | 12.0 | 114 | 100 |
| 7.3 | 18.87 | 17.31 | 106 | 104 |
| 8.1 | 22.4 | 28.0 | 116 | 100 |
| 8.2 | 22.41 | 24.53 | 96 | 106 |
| 8.3 | 16.39 | 14.93 | 122 | 134 |
| 8.4 | 27.69 | 27.27 | 130 | 132 |
| 8.5 | 22.45 | 24.4 | 98 | 90 |
| 8.6 | 30.91 | 28 | 110 | 100 |
| 9.2 | 1.59 | 1.89 | 126 | 106 |
| 9.3 | 2.56 | 2.7 | 78 | 74 |
| 11.1* | 2.48 | 3.01 | 104 | 87 |
| 11.2* | 3.15 | 3.35 | 92 | 89 |
| 11.3 | 3.09 | 3.36 | 110 | 88 |
| 11.4 | 2.38 | 2.27 | 84 | 88 |

| | | | | |
|---|---|---|---|---|
| 11.5 | 3.77 | 4.0 | 106 | 100 |
| 13.1 | 8.62 | 12.28 | 116 | 114 |
| 13.2 | 12.24 | 3.7 | 98 | 108 |
| 14.1 | 10.81 | 8.1 | 148 | 196 |
| 14.4 | 14.29 | 12.96 | 112 | 108 |
| 14.5 | 3.57 | 3.8 | 224 | 208 |
| 14.6 | 18.42 | 18.57 | 152 | 140 |
| NK1.2* | 0 | 0 | 172 | 178 |
| NK1.3* | 0 | 0 | 156 | 161 |
| NK1.4* | 0 | 0 | 176 | 170 |
| NK1.5* | 0 | 0 | 165 | 162 |
| NK1.6* | 0 | 0 | 158 | 163 |
| NK2.1* | 0 | 0 | 136 | 132 |
| NK2.2* | 0 | 0 | 128 | 131 |
| NK4.1* | 0 | 0 | 106 | 102 |
| NK4.2* | 0 | 0 | 99.6 | 103 |

Percent syllables stuttered and syllables per minute data collected by 2 experienced judges not familiar with the study.

*only included in Study Two

Table 3
Study One: Speech naturalness mean, range, and standard deviation (SD) for all 43 raters for all speech samples combined on rating occasions one and two.

| Rating Occasion | | | | | |
|---|---|---|---|---|---|
| 1 | | | 2 | | |
| $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| 5.67 | 2.4-8.4 | 1.80 | 5.97 | 2.8-8.4 | 1.86 |

For all 43 raters combined

Table 4
Study One: Average group intrarater agreement of speech naturalness scores cumulative number and percentage (in parentheses).

| Occ. | +/- 0 | +/- 1 | +/- 2 | +/- 3 | +/- 4 | +/- 5 | +/- 6 | +/- 7 | +/- 8 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Training Group | | | | | |
| 1 v 2 | 127 | 268 | 327 | 349 | 356 | 360 | | | |
| | (48.85) | (74.44) | (90.83) | (96.94) | (98.89) | (100) | | | |
| 3 v 4 | 195 | 334 | 354 | 360 | | | | | |
| | (54.17) | (92.78) | (98.33) | (100) | | | | | |
| | | | | Exposure Control Group | | | | | |
| 1 v 2 | 126 | 249 | 303 | 322 | 334 | 335 | 335 | 336 | |
| | (37.5) | (74.11) | (90.18) | (95.83) | (99.4) | (99.7) | (99.7) | (100) | |
| 3 v 4 | 151 | 270 | 316 | 332 | 335 | 336 | | | |
| | (44.94) | (80.36) | (94.05) | (98.81) | (97.7) | (100) | | | |
| | | | | Control Group | | | | | |
| 1 v 2 | 127 | 260 | 295 | 312 | 332 | 334 | 335 | 336 | |
| | (37.8) | (77.38) | (87.8) | (92.86) | (98.81) | (99.4) | (99.7) | (100) | |
| 3 v 4 | 143 | 274 | 317 | 327 | 330 | 335 | 336 | | |
| | (42.56) | (81.55) | (94.35) | (97.32) | (98.21) | (99.7) | (100) | | |

Table 5
Study One: Average group interrater agreement of speech naturalness scores for each rating occasion cumulative number and percentage (in parentheses).

| Occ. | +/- 0 | +/- 1 | +/- 2 | +/- 3 | +/- 4 | +/- 5 | +/- 6 | +/- 7 | +/- 8 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Training Group | | | | |
| 1 | 580 | 1469 | 1975 | 2276 | 2412 | 2494 | 2516 | 2520 | |
| | (23.02) | (58.29) | (78.37) | (90.32) | (95.71) | (98.97) | (99.84) | (100) | |
| 2 | 588 | 1519 | 1954 | 2241 | 2402 | 2484 | 2514 | 2519 | 2520 |
| | (23.33) | (60.28) | (77.54) | (88.93) | (95.32) | (98.57) | (99.76) | (99.96) | (100) |
| 3 | 666 | 1644 | 2178 | 2396 | 2497 | 2516 | 2519 | 2520 | |
| | (26.43) | (65.24) | (86.43) | (95.08) | (99.09) | (99.84) | (99.96) | (100) | |
| 4 | 625 | 1636 | 2148 | 2384 | 2485 | 2514 | 2520 | | |
| | (24.8) | (64.9) | (85.2) | (94.6) | (98.6) | (99.8) | (100) | | |
| | | | | | Exposure Control Group | | | | |
| 1 | 390 | 1102 | 1577 | 1883 | 2033 | 2129 | 2169 | 2183 | 2184 |
| | (17.86) | (50.46) | (72.21) | (86.22) | (93.09) | (97.48) | (99.31) | (99.95) | (100) |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 476 | 1170 | 1643 | 1935 | 2087 | 2151 | 2177 | 2184 | |
| | (21.79) | (53.57) | (75.23) | (88.6) | (95.56) | (98.49) | (99.68) | (100) | |
| 3 | 383 | 1082 | 1563 | 1877 | 2028 | 2122 | 2171 | 2184 | |
| | (17.54) | (49.54) | (71.57) | (85.94) | (92.86) | (97.16) | (99.4) | (100) | |
| 4 | 400 | 1114 | 1581 | 1875 | 2055 | 2136 | 2176 | 2184 | |
| | (18.32) | (51.01) | (72.39) | (85.85) | (94.09) | (97.8) | (99.63) | (100) | |
| | | | | Control Group | | | | | |
| 1 | 421 | 1146 | 1478 | 1837 | 1946 | 2109 | 2167 | 2184 | |
| | (19.28) | (52.47) | (67.67) | (84.11) | (89.1) | (96.57) | (99.22) | (100) | |
| 2 | 418 | 1095 | 1490 | 1738 | 1931 | 2069 | 2162 | 2182 | 2184 |
| | (19.14) | (50.14) | (68.22) | (79.58) | (88.42) | (94.73) | (98.99) | (99.91) | (100) |
| 3 | 418 | 1086 | 1521 | 1795 | 1960 | 2092 | 2166 | 2183 | 2184 |
| | (19.14) | (49.73) | (69.64) | (82.19) | (89.74) | (95.79) | (99.18) | (99.95) | (100) |
| 4 | 421 | 1036 | 1464 | 1732 | 1914 | 2060 | 2158 | 2183 | 2184 |
| | (19.28) | (47.44) | (67.03) | (79.3) | (87.64) | (94.32) | (98.81) | (99.95) | (100) |

Table 6
Sample size calculations using Study One data reported in Chapter 3.

| | Intrarater agreement analyses | |
|---|---|---|
| F test | Effect size f | Total sample size needed* |
| Main effect Rating Occasion | .732 | 9 |
| Main effect Group | .338 | 27 |
| Rating Occasion x Group Interaction | .482 | 15 |
| | **Interrater agreement analyses** | |
| F test | Effect size f | Total sample size needed* |
| Main effect Rating Occasion | .105 | 18 |
| Main effect Group | .941 | 6 |
| Rating Occasion x Group Interaction | .510 | 6 |

*as estimated by G*Power 3

Table 7
Tasks completed by each group during each session in Study Two.

| Group[a] | Session 1 | Session 2 | Session 3 | Session 4 |
|---|---|---|---|---|
| Training | Rate Only | Rate Only | Training/ Assessment + Rating | Rate Only |
| Exposure Control | Rate Only | Rate Only | Exposure to Training Stimuli + Rating | Rate Only |
| Control Group | Rate Only | Rate Only | Rate Only | Rate Only |

Note. Each session was conducted 7-14 days after the previous session.
[a] n= 18 for each group.
36 experimental speech samples rated one time during each session.

Table 8
Study Two: Average speech naturalness scores, range, and standard deviation (SD) for each group on each rating occasion by stuttering severity.

| Speaker Group | Rating Occasion | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Training Group** | | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 2.01 | 1-3.7 | 0.69 | 1.78 | 1.1-3.8 | 0.74 | 2.75 | 1.7-4.2 | 0.78 | 2.50 | 1.2-4.3 | 0.87 |
| Mild | 3.69 | 1.7-5.3 | 1.01 | 3.73 | 1.7-6 | 1.12 | 4.90 | 3-6.9 | 0.94 | 4.41 | 3.2-5.7 | 0.63 |
| Moderate | 6.60 | 4.9-8 | 0.92 | 7.04 | 5.1-8.2 | 0.86 | 7.26 | 5.9-8.6 | 0.63 | 6.89 | 4.7-8.4 | 0.89 |
| Severe | 8.09 | 7.1-9 | 0.55 | 8.04 | 6.9-8.7 | 0.50 | 7.98 | 7.1-8.7 | 0.50 | 7.72 | 6.1-8.6 | 0.61 |
| **Exposure Control Group** | | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 2.01 | 1-4 | 0.79 | 1.95 | 1-3.3 | 0.56 | 1.84 | 1-3.3 | 0.61 | 1.59 | 1-2.7 | 0.50 |
| Mild | 3.65 | 2.2-5.3 | 0.89 | 3.86 | 2.6-5.1 | 0.69 | 4.37 | 2.7-5.6 | 0.92 | 3.83 | 2.8-5.4 | 0.87 |
| Moderate | 6.52 | 4.3-8.4 | 1.11 | 6.96 | 5.2-8.6 | 0.92 | 6.93 | 4.1-8.6 | 1.17 | 6.69 | 4.4-8 | 1.04 |
| Severe | 7.70 | 6.2-9 | 0.76 | 7.99 | 6.7-8.8 | 0.56 | 8.0 | 6.2-9 | 0.76 | 7.91 | 5.3-8.9 | 0.88 |
| **Control Group** | | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 2.09 | 1-3.8 | 0.92 | 2.14 | 1-3.98 | 0.99 | 1.86 | 1-5.2 | 1.08 | 1.68 | 1-3.4 | 0.74 |
| Mild | 3.37 | 1.8-5.4 | 1.16 | 3.46 | 1.6-5.8 | 1.36 | 3.37 | 1.8-6 | 1.11 | 3.14 | 1.8-6 | 1.10 |
| Moderate | 6.17 | 4.3-8.2 | 1.05 | 6.42 | 4.4-9 | 1.19 | 6.21 | 4.2-8.1 | 1.17 | 6.26 | 3.9-7.7 | 1.04 |
| Severe | 7.75 | 6.4-8.9 | 0.71 | 7.94 | 7.2-8.9 | 0.49 | 7.78 | 6.2-8.8 | 0.60 | 8.06 | 7.3-9 | 0.52 |

n=18 per group

Table 9
Study Two: Mean, range, and standard deviation of speech naturalness assigned by 54 naturalness raters to normal speaking children and children who stutter at mild, moderate, and severe levels on rating occasion 1.

| Speaker Group | Mean | Range | Standard Deviation |
|---|---|---|---|
| Normal | 2.04 | 1.00-4.00 | .798 |
| Mild | 3.57 | 1.67-5.44 | 1.02 |
| Moderate | 6.43 | 4.33-8.44 | 1.03 |
| Severe | 7.85 | 6.22-9.00 | .692 |

*Statistically significant differences between each speaker group were found (see p. 98)

Table 10
Study Two: Average group intrarater agreement of speech naturalness scores for each of the 3 groups cumulative number and percentage (in parentheses).

| Occ. | +/- 0 | +/- 1 | +/- 2 | +/- 3 | +/- 4 | +/- 5 | +/- 6 | +/- 7 | +/- 8 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Training Group | | | | | |
| 1 v 2 | 264 | 504 | 597 | 633 | 644 | 648 | | | |
| | (40.74) | (77.78) | (92.13) | (97.69) | (99.38) | (100) | | | |
| 3 v 4 | 278 | 551 | 629 | 644 | 647 | 648 | | | |
| | (42.90) | (85.03) | (97.07) | (99.38) | (99.85) | (100) | | | |
| | | | | Exposure Control Group | | | | | |
| 1 v 2 | 275 | 506 | 591 | 626 | 641 | 646 | 648 | | |
| | (42.44) | (78.09) | (91.20) | (96.60) | (98.92) | (99.69) | (100) | | |
| 3 v 4 | 280 | 529 | 610 | 638 | 644 | 646 | 647 | 648 | |
| | (43.21) | (81.64) | (94.14) | (98.46) | (99.38) | (99.69) | (99.85) | (100) | |
| | | | | Control Group | | | | | |
| 1 v 2 | 282 | 495 | 590 | 622 | 640 | 645 | 647 | 648 | |
| | (43.52) | (76.39) | (91.05) | (95.99) | (98.77) | (99.54) | (99.85) | (100) | |
| 3 v 4 | 316 | 532 | 604 | 627 | 643 | 646 | 646 | 647 | 648 |
| | (48.77) | (82.10) | (93.21) | (96.76) | (99.23) | (99.69) | (99.69) | (99.85) | (100) |

Table 11
Study Two: Average intrarater agreement scores, range, and standard deviation (SD) for each group for each rating occasion comparison (occasion 1 v 2 and 3 v 4) by stuttering severity.

| Speaker Group | Rating Occasion Comparison | | | | | |
|---|---|---|---|---|---|---|
| **Training Group** | | | | | | |
| | Occasions 1v 2 | | | Occasions 3 v 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 87.04 | 55.6-100 | 15.36 | 88.89 | 11.1-100 | 22.22 |
| Mild | 72.22 | 44.4-100 | 15.83 | 78.40 | 11.1-100 | 23.06 |
| Moderate | 62.96 | 33.3-100 | 23.49 | 85.80 | 55.6-100 | 14.66 |
| Severe | 88.89 | 77.8-100 | 8.52 | 87.04 | 44.4-100 | 14.39 |
| **Exposure Control Group** | | | | | | |
| | Occasions 1v 2 | | | Occasions 3 v 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 87.04 | 66.7-100 | 11.59 | 90.12 | 44.4-100 | 15.19 |
| Mild | 70.99 | 44.4-100 | 14.83 | 70.99 | 44.4-100 | 17.93 |
| Moderate | 67.90 | 22.2-100 | 22.51 | 75.31 | 44.4-100 | 14.03 |
| Severe | 86.42 | 22.2-100 | 20.01 | 90.12 | 77.8-100 | 9.25 |
| **Control Group** | | | | | | |
| | Occasions 1v 2 | | | Occasions 3 v 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 87.65 | 55.6-100 | 15.19 | 90.74 | 44.4-100 | 16.28 |
| Mild | 67.28 | 33.3-100 | 21.38 | 79.63 | 33.3-100 | 16.28 |
| Moderate | 67.28 | 33.3-88.9 | 14.50 | 73.46 | 44.4-100 | 16.67 |
| Severe | 83.33 | 55.6-100 | 14.88 | 84.57 | 55.6-100 | 12.14 |

n=18 per group

*No significant difference between groups

Table 12
Study Two: Average group interrater agreement of speech naturalness scores for each group for each rating occasion cumulative number and percentage (in parentheses).

| Occ. | +/- 0 | +/- 1 | +/- 2 | +/- 3 | +/- 4 | +/- 5 | +/- 6 | +/- 7 | +/- 8 |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | **Training Group** | | | | |
| 1 | 3130 | 7126 | 9318 | 10336 | 10754 | 10938 | 11002 | 11014 | 11016 |
| | (28.41) | (64.69) | (84.59) | (93.83) | (97.62) | (99.29) | (99.87) | (99.98) | (100) |
| 2 | 3334 | 7206 | 9176 | 10234 | 10676 | 10904 | 11004 | 11012 | 11016 |
| | (30.27) | (65.41) | (83.30) | (92.90) | (96.91) | (98.98) | (99.89) | (99.96) | (100) |
| 3 | 3098 | 7784 | 10070 | 10764 | 10974 | 11006 | 11016 | | |
| | (28.12) | (70.66) | (91.41) | (97.71) | (99.62) | (99.91) | (100) | | |
| 4 | 2926 | 7504 | 9954 | 10804 | 10990 | 11016 | | | |
| | (26.56) | (68.12) | (90.36) | (98.08) | (99.76) | (100) | | | |
| | | | | | **Exposure Control Group** | | | | |
| 1 | 2800 | 6528 | 8920 | 10088 | 10710 | 10910 | 11002 | 11016 | |
| | (25.42) | (59.26) | (80.97) | (91.58) | (97.22) | (99.04) | (99.87) | (100) | |
| 2 | 3216 | 7130 | 9468 | 10440 | 10820 | 10964 | 11014 | 11016 | |
| | (29.19) | (64.72) | (85.95) | (94.77) | (98.22) | (99.53) | (99.98) | (100) | |
| 3 | 3056 | 6972 | 9312 | 10286 | 10758 | 10970 | 11012 | 11016 | |
| | (27.74) | (63.29) | (84.53) | (93.37) | (97.66) | (99.58) | (99.96) | (100) | |
| 4 | 3212 | 7086 | 9336 | 10406 | 10866 | 10998 | 11014 | 11016 | |
| | (29.16) | (64.32) | (84.75) | (94.46) | (98.64) | (99.84) | (99.98) | (100) | |

**Control Group**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2752 | 6402 | 8502 | 9818 | 10522 | 10852 | 10946 | 11002 | 11016 |
| | (24.98) | (58.12) | (77.18) | (89.12) | (95.52) | (98.51) | (99.37) | (99.87) | (100) |
| 2 | 2950 | 6326 | 8400 | 9632 | 10352 | 10718 | 10912 | 10988 | 11016 |
| | (26.78) | (57.43) | (76.25) | (87.44) | (93.97) | (97.29) | (99.06) | (99.75) | (100) |
| 3 | 2974 | 6266 | 8272 | 9582 | 10258 | 10676 | 10842 | 10938 | 11016 |
| | (27) | (56.88) | (75.09) | (86.98) | (93.12) | (96.91) | (98.42) | (99.29) | (100) |
| 4 | 3312 | 6832 | 8866 | 10030 | 10634 | 10870 | 10962 | 11000 | 11016 |
| | (30.07) | (62.02) | (80.48) | (91.05) | (96.53) | (98.67) | (99.51) | (99.85) | (100) |

Table 13
Study Two: Interrater agreement, range, and standard deviation (SD) for each of the 3 groups for each of the 4 rating occasions by stuttering severity.

| Speaker Group | Rating Occasion | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Training Group | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 72.06 | 38.6-85.6 | 11.5 | 76.88 | 25.5-88.9 | 16.7 | 67.12 | 43.1-77.8 | 8.36 | 67.41 | 33.3-78.4 | 11.68 |
| Mild | 51.05 | 37.9-63.4 | 6.8 | 48.73 | 28.1-61.4 | 7.69 | 60.64 | 35.9-75.8 | 10.73 | 66.81 | 51.6-81.7 | 8.46 |
| Moderate | 57.52 | 29.4-71.9 | 10.24 | 59.04 | 27.5-74.5 | 11.19 | 74.22 | 49.0-85.6 | 10.35 | 66.01 | 20.3-80.4 | 15.67 |
| Severe | 78.14 | 58.8-90.2 | 8.58 | 76.98 | 58.2-89.5 | 7.90 | 80.68 | 68.0-90.8 | 6.42 | 72.26 | 42.5-88.9 | 9.99 |
| | Exposure Control Group | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 69.24 | 41.8-82.4 | 11.28 | 74.18 | 47.1-84.3 | 9.29 | 76.06 | 44.4-88.9 | 10.94 | 81.12 | 54.2-90.8 | 10.32 |
| Mild | 48.29 | 32.0-59.5 | 8.22 | 52.00 | 27.5-64.7 | 11.25 | 50.84 | 37.9-61.4 | 5.97 | 51.42 | 38.6-62.1 | 7.70 |
| Moderate | 50.62 | 29.4-58.8 | 7.62 | 55.55 | 27.5-68.6 | 9.54 | 55.26 | 17.6-69.9 | 13.75 | 54.68 | 27.5-66.0 | 11.48 |
| Severe | 68.92 | 41.2-82.4 | 12.08 | 77.19 | 46.4-87.6 | 9.04 | 71.02 | 37.9-85.0 | 11.71 | 70.08 | 13.1-86.3 | 15.94 |
| | Control Group | | | | | | | | | | | |
| | 1 | | | 2 | | | 3 | | | 4 | | |
| | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD | $\bar{x}$ | Range | SD |
| Normal | 68.53 | 44.4-80.4 | 11.00 | 62.24 | 40.5-75.2 | 11.20 | 71.59 | 30.7-83.0 | 16.17 | 76.12 | 50.3-85.6 | 10.34 |
| Mild | 48.66 | 31.4-59.5 | 9.46 | 45.61 | 26.1-57.5 | 9.61 | 47.78 | 20.9-62.7 | 9.84 | 49.96 | 20.9-60.8 | 10.13 |
| Moderate | 48.95 | 34.6-63.4 | 8.65 | 47.93 | 24.8-59.5 | 10.62 | 43.35 | 27.5-54.2 | 7.73 | 49.31 | 30.1-67.3 | 10.7 |
| Severe | 66.31 | 52.3-78.4 | 7.31 | 73.93 | 59.5-86.3 | 8.16 | 64.78 | 30.7-76.5 | 11.61 | 72.69 | 48.4-81.7 | 8.55 |

n=18 per group

*Post hoc analysis (Fisher's LSD) found statistically significant differences between the following speaker groups:

**normally** speaking children were significantly **higher** than children who stutter at a **mild** level

**normally** speaking children were significantly **higher** than children who stutter at a **moderate** level

children who stutter at a **mild** level were significantly **lower** than children who stutter at a **moderate** level

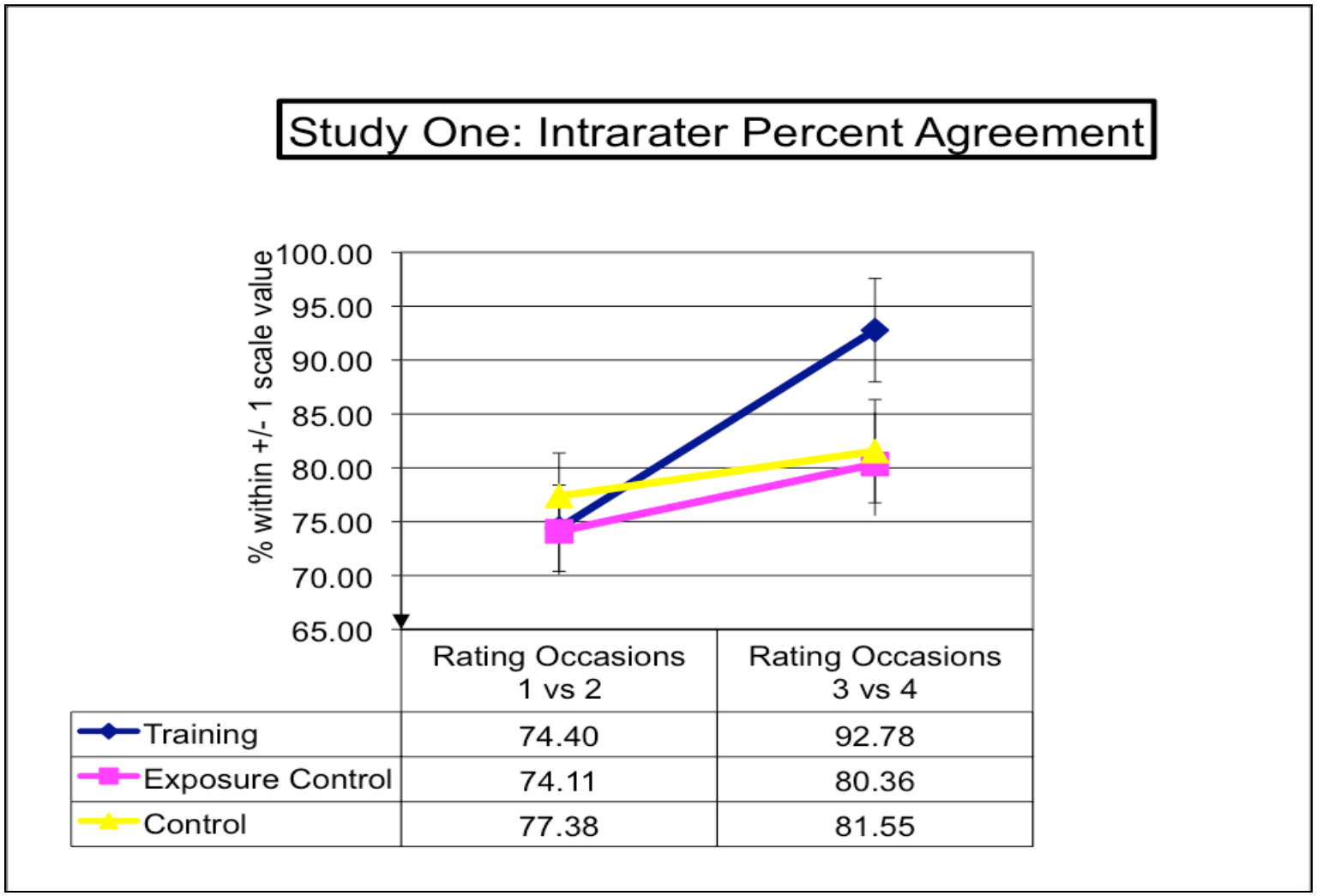
children who stutter at a **mild** level were significantly **lower** than children who stutter at a **severe** level

children who stutter at a **moderate** level were significantly **lower** than children who stutter at a **severe** level

Table 14.
Study Two: Intraclass correlations for the mean speech naturalness rating of 18 raters per group ($R_{18}$) and the average individual rater ($R_1$) for each stuttering severity level.

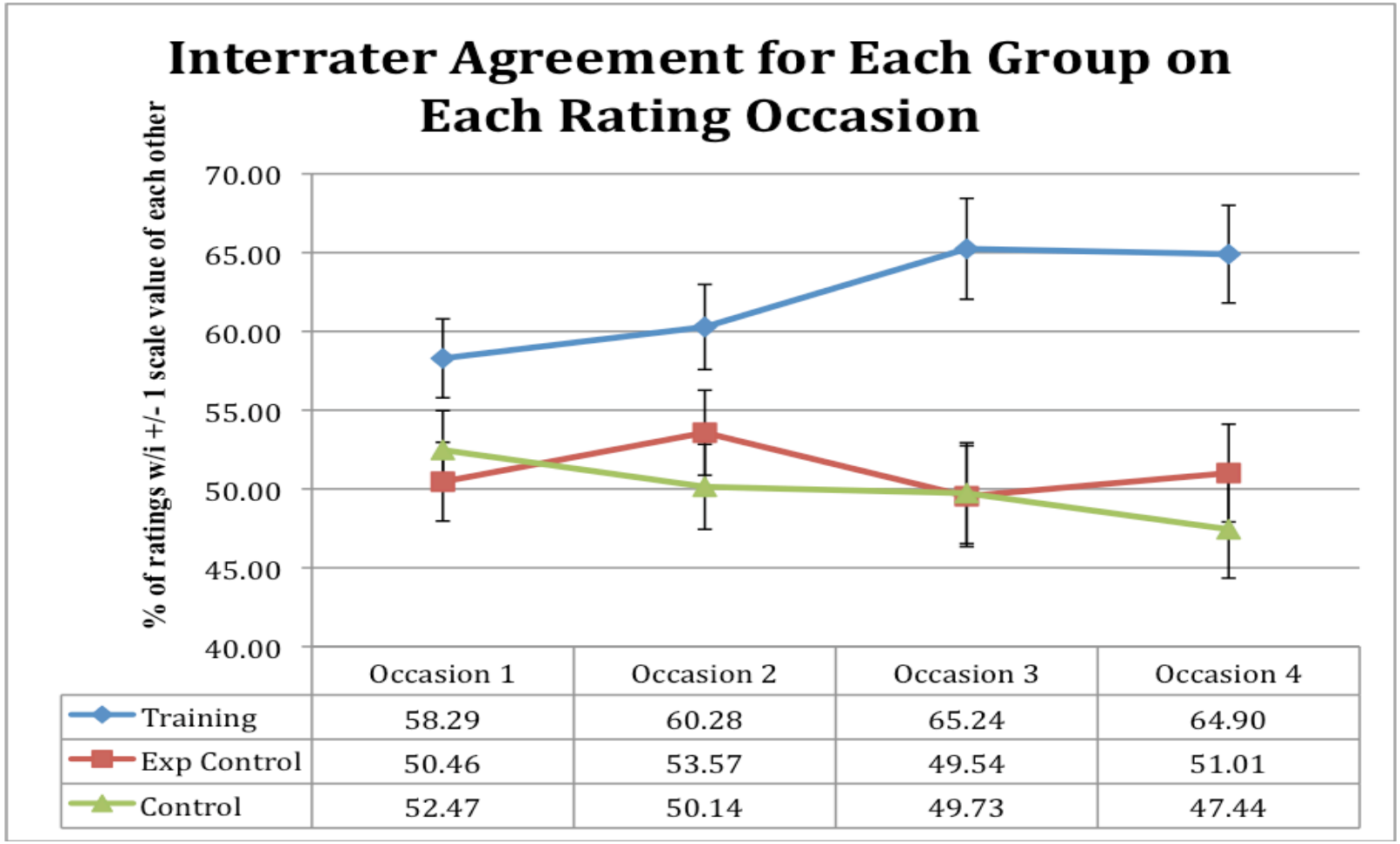
| | Experimental Group | |
|---|---|---|
| Speaker Group | $R_{18}$ | $R_1$ |
| Normal | .950* | .346* |
| Mild | .933* | .279 |
| Moderate | .921* | .244* |
| Severe | .853* | .138* |
| | Exposure Control Group | |
| Normal | .927* | .260* |
| Mild | .867* | .154* |
| Moderate | .949* | .342* |
| Severe | .950* | .344* |
| | Control Group | |
| Normal | .964* | .424* |
| Mild | .955* | .373* |
| Moderate | .949* | .341* |
| Severe | .785* | .092* |

* significant at the .000 level.



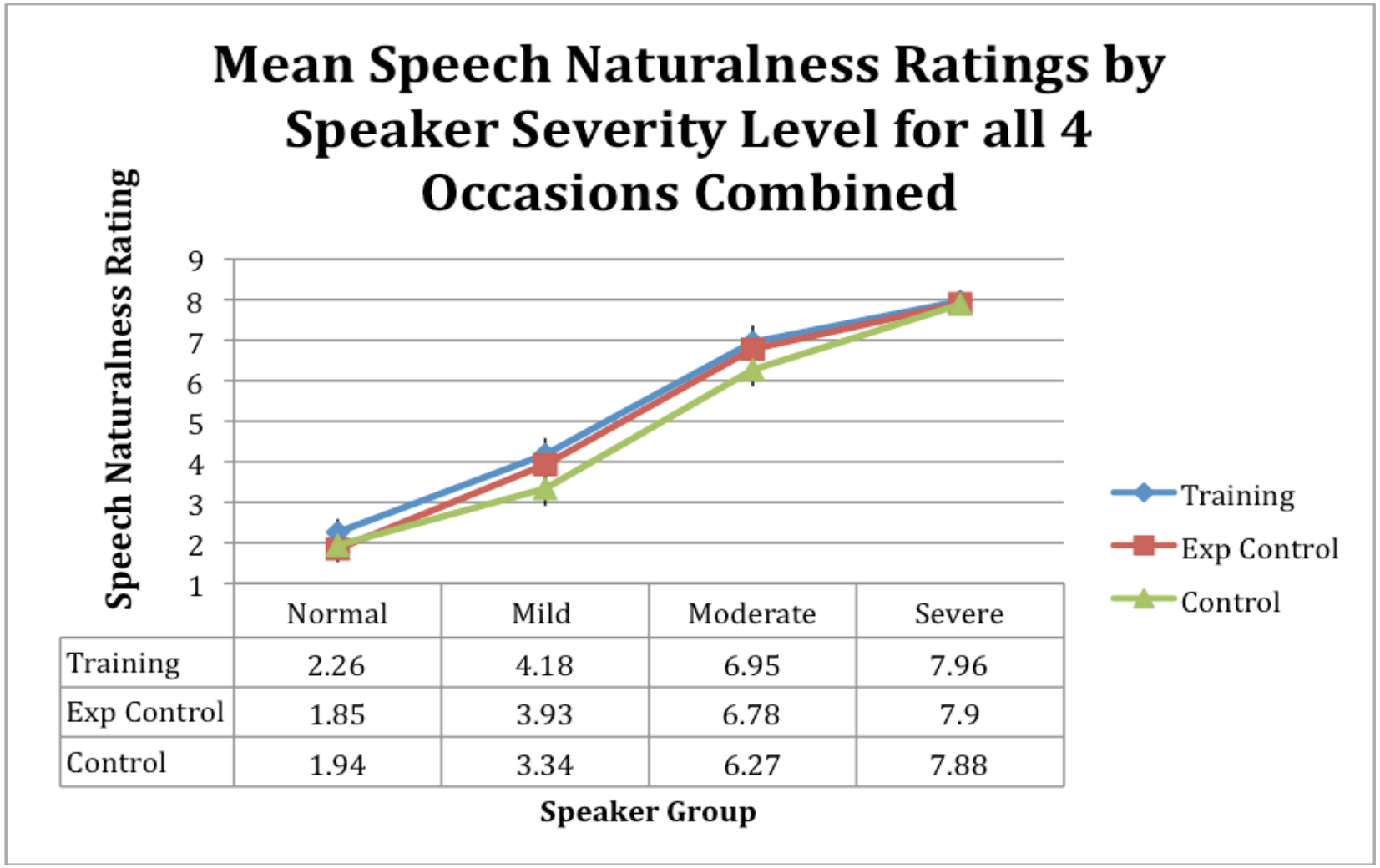

| | Rating Occasions 1 vs 2 | Rating Occasions 3 vs 4 |
|---|---|---|
| Training | 74.40 | 92.78 |
| Exposure Control | 74.11 | 80.36 |
| Control | 77.38 | 81.55 |

*Figure 1*. Study One: Average intrarater agreement for the 43 participants separated by group. Error bars for 95% confidence intervals are also presented.
*Statistically significant differences between **treatment group** and **each** of the **control groups**.




| | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| Training | 58.29 | 60.28 | 65.24 | 64.90 |
| Exp Control | 50.46 | 53.57 | 49.54 | 51.01 |
| Control | 52.47 | 50.14 | 49.73 | 47.44 |

*Figure 2.* Study One: Average interrater agreement for participants separated by group for each rating occasions. Error bars for 95% confidence intervals are also presented.
***Training Group** significantly **higher** than the **two control groups** on all four occasions




| | Normal | Mild | Moderate | Severe |
|---|---|---|---|---|
| Training | 2.26 | 4.18 | 6.95 | 7.96 |
| Exp Control | 1.85 | 3.93 | 6.78 | 7.9 |
| Control | 1.94 | 3.34 | 6.27 | 7.88 |

*Figure 3.* Study Two: Speech naturalness ratings for participants separated by group collapsed across rating occasions. Error bars for 95% confidence intervals are also presented.
*Between subject effect of group was not statistically significant

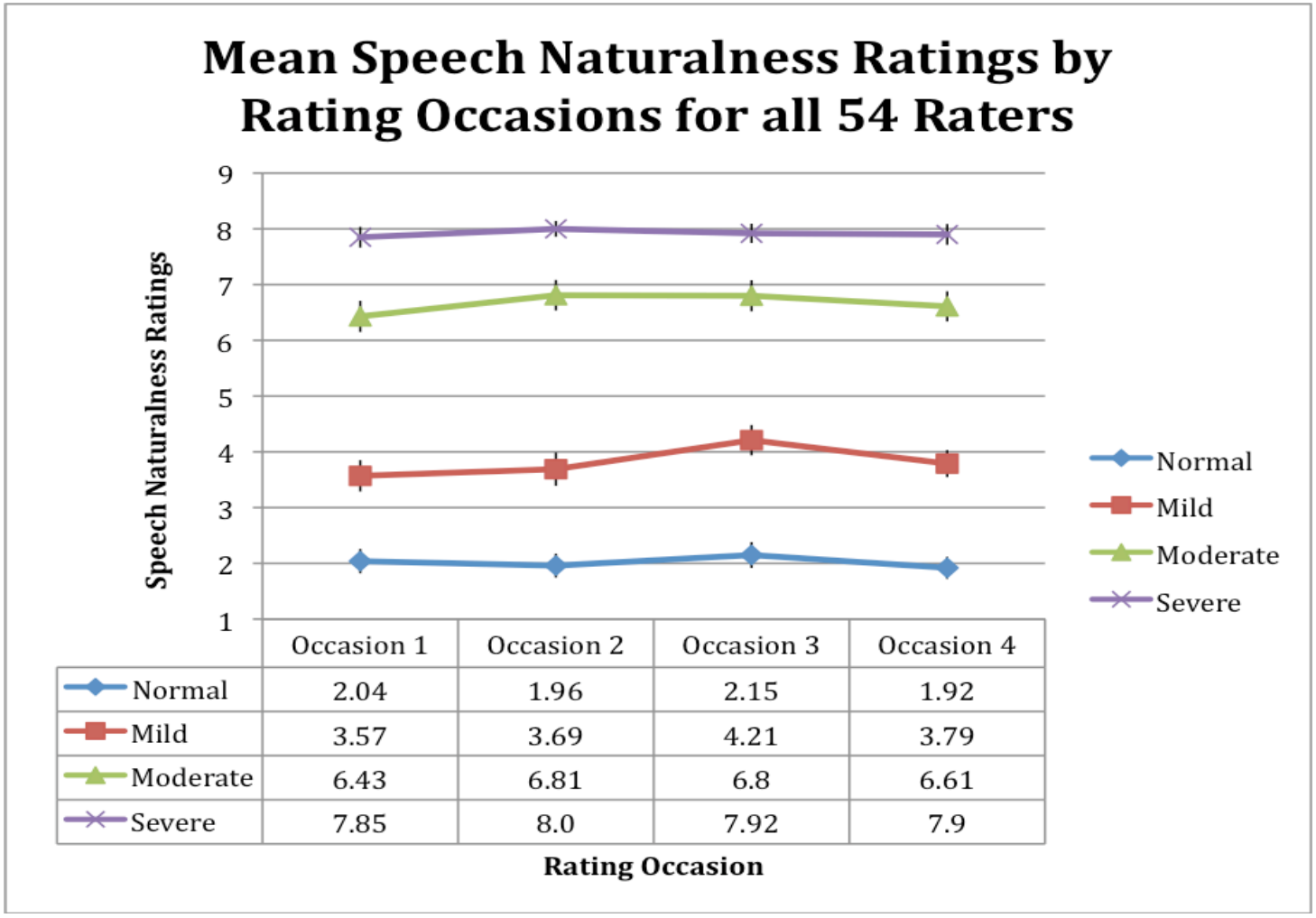



*Figure 4.* Study Two: Mean speech naturalness ratings for each stuttering severity level on each rating occasion for all groups combined (N=54). Error bars for 95% confidence intervals are also presented.
*Statistically significant differences between each **speaker group combination** (see p. 98).

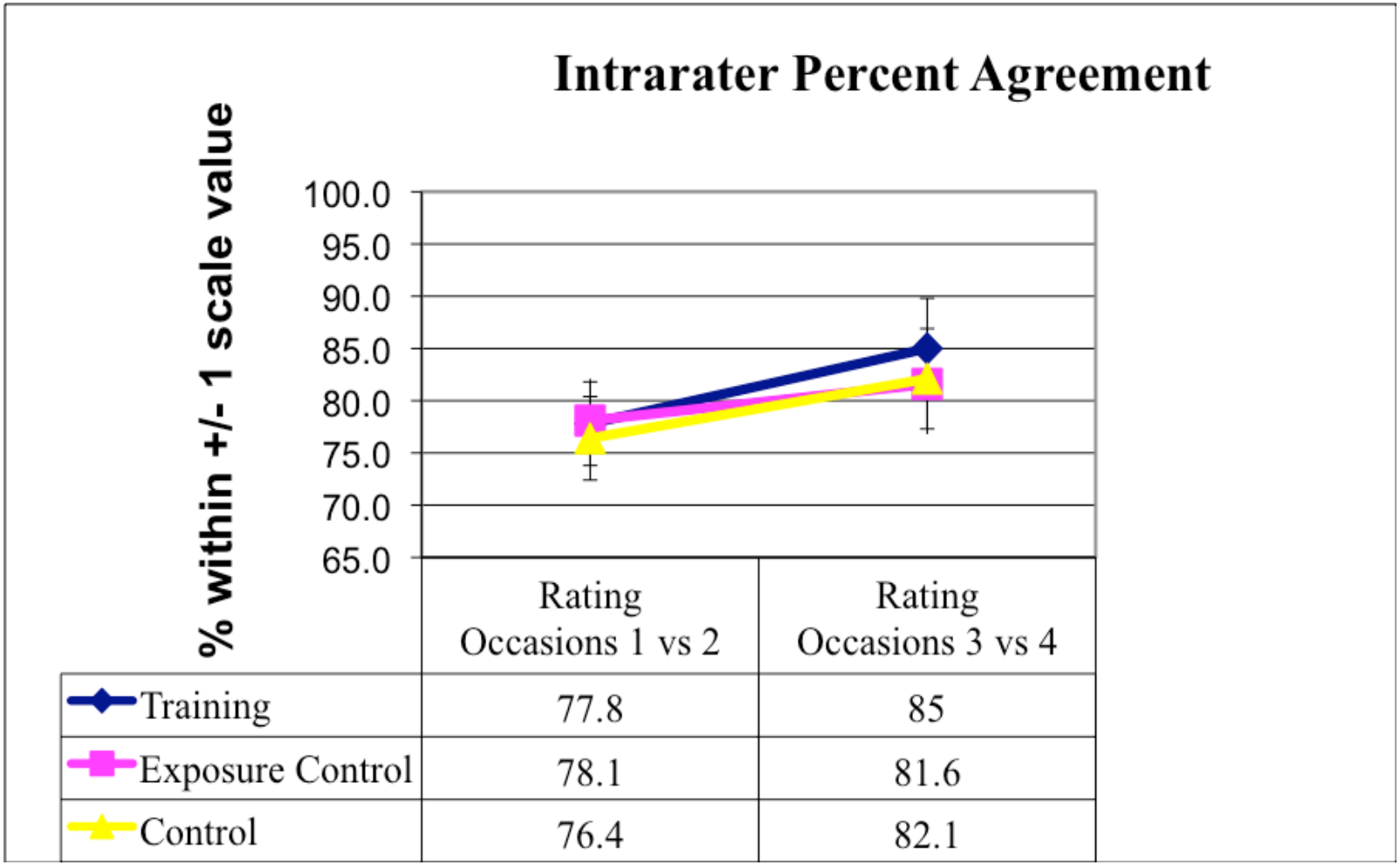



*Figure 5*. Study Two: Intrarater agreement for participants separated by group. Error bars for 95% confidence intervals are also presented.
*Rating occasion comparisons by training group not significant
**Post hoc (Fisher's LSD) found intrarater agreement on **rating occasions 1 vs 2** to be **significantly lower** than on **rating occasions 3 vs 4** for **all** three **groups**.

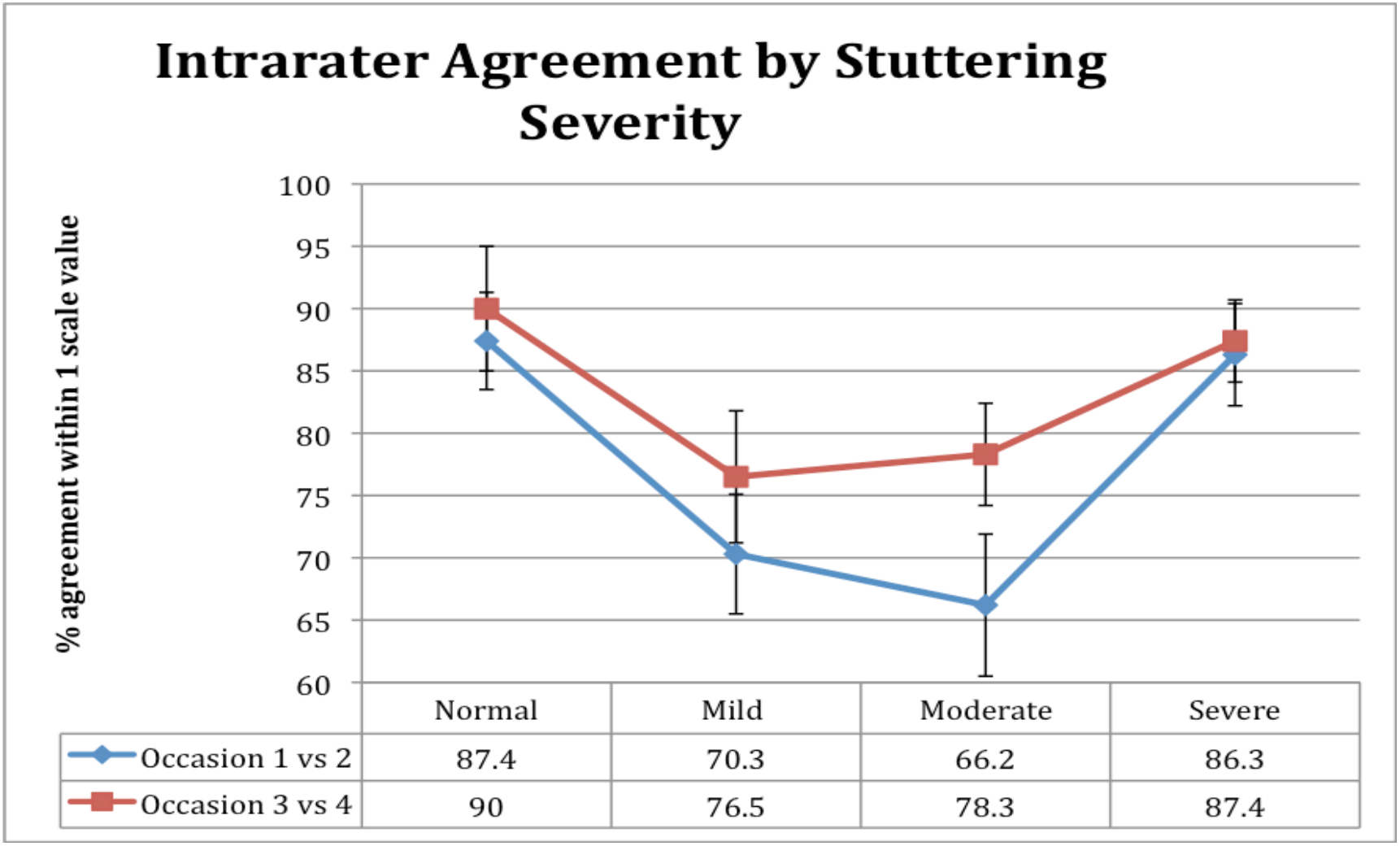
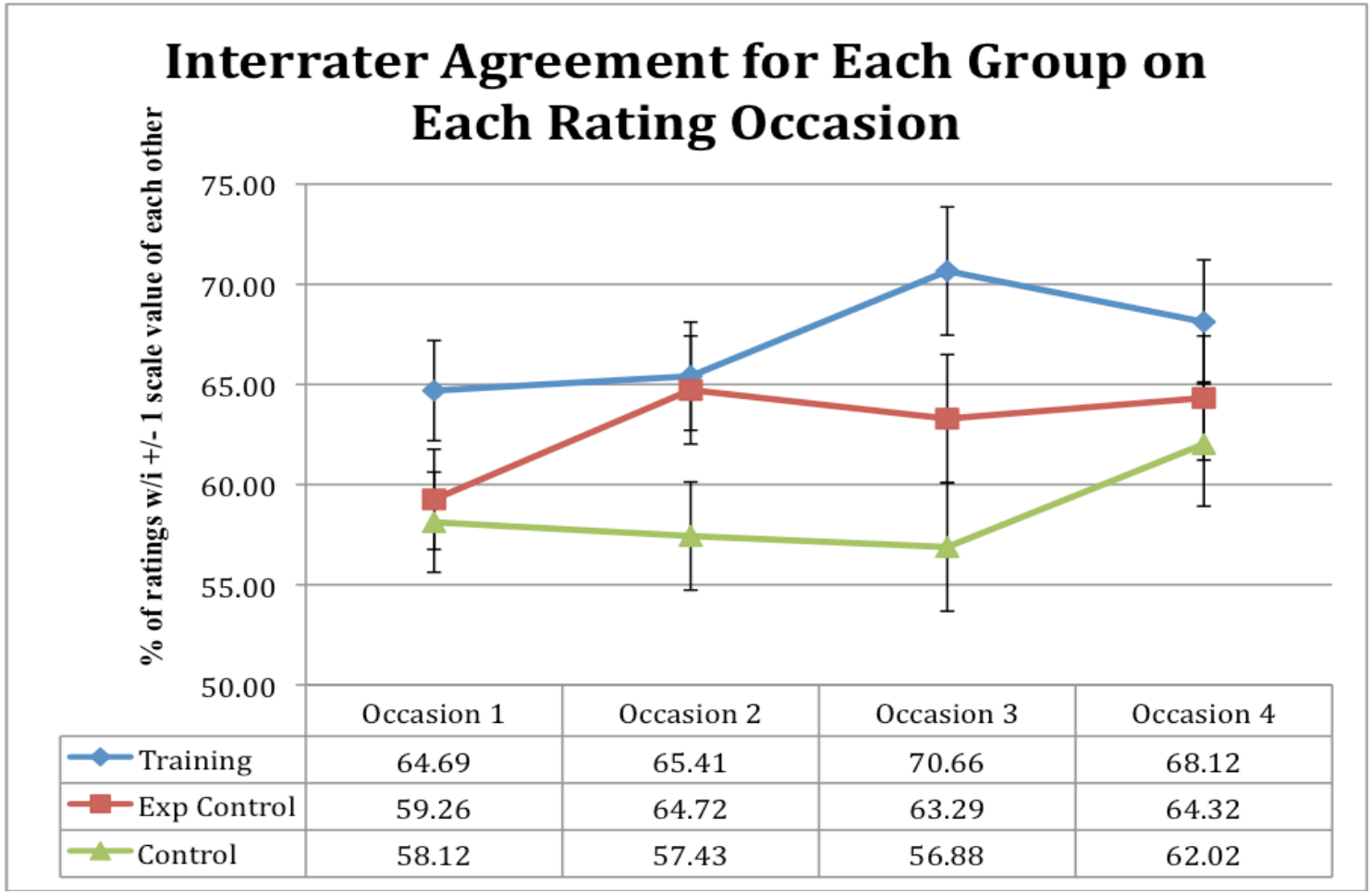



*Figure 6*. Study Two: Intrarater agreement for participants for each stuttering severity level for both rating occasion comparisons for all groups combined. Error bars for 95% confidence intervals are also presented.
*Statistically significant difference between **normal** and **mild** speakers, **normal** and **moderate** speakers, **mild** and **severe** speakers, and **moderate** and **severe** speakers across rating occasions.




| | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| Training | 64.69 | 65.41 | 70.66 | 68.12 |
| Exp Control | 59.26 | 64.72 | 63.29 | 64.32 |
| Control | 58.12 | 57.43 | 56.88 | 62.02 |

*Figure 7.* Study Two: Interrater agreement for participants separated by group for each rating occasion. Error bars for 95% confidence intervals are also presented.
*Significant difference at Occasion 1 between **Training** and **each control group**
*Significant difference at Occasion 2 between **Training** and **Control** groups
*Significant difference at Occasion 3 between **Training** and **each control group** and between the **two control groups**
*Significant difference at Occasion 4 between **Training** and **Control** groups

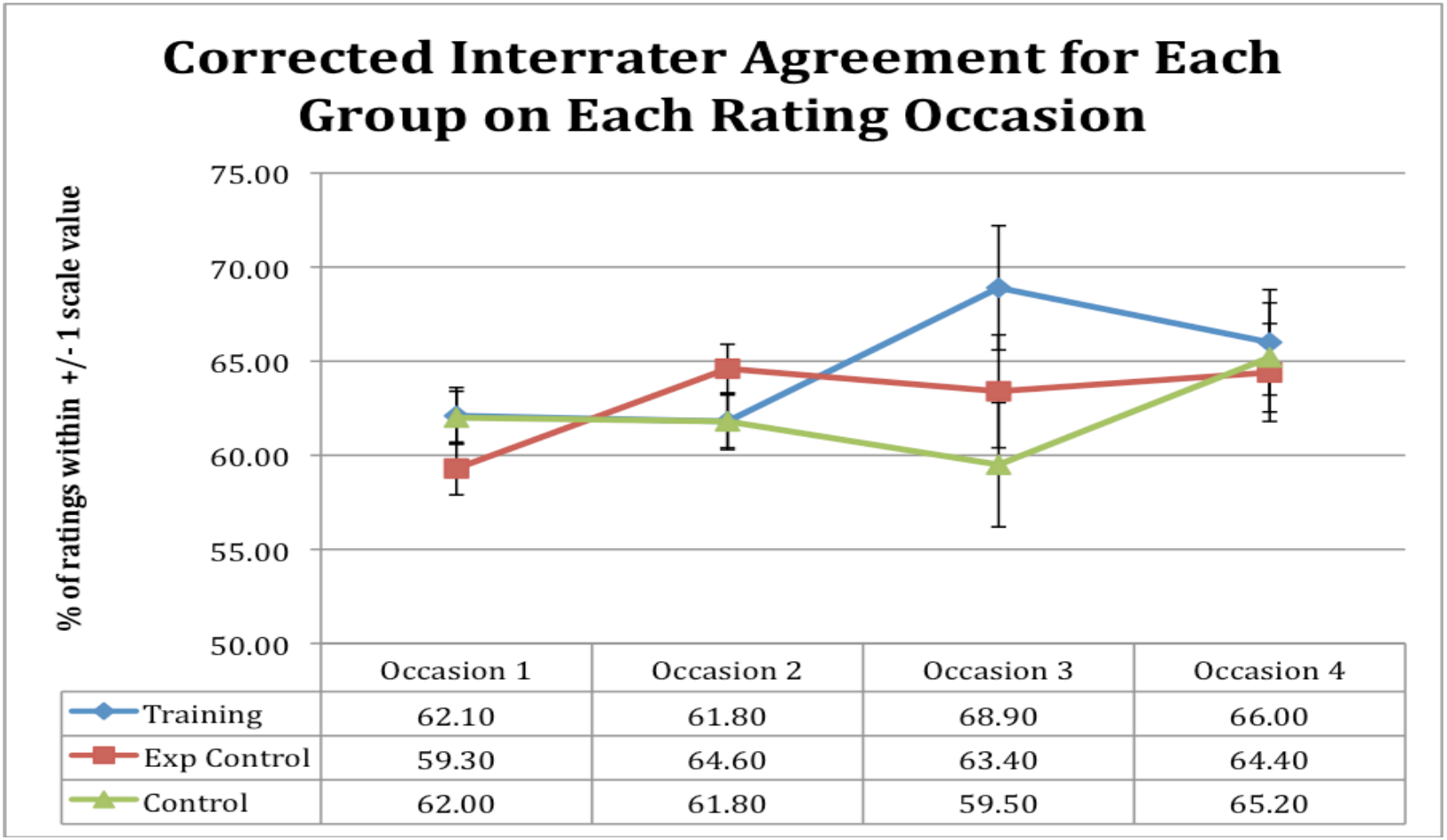



*Figure 8.* Study Two: Corrected interrater agreement for participants separated by group for each of the four rating occasions using group differences on occasions 1 and 2 as a covariate. Error bars for 95% confidence intervals are also presented.

*Only statistically significant difference was between **Training** and **Control** groups on **Occasion 3**

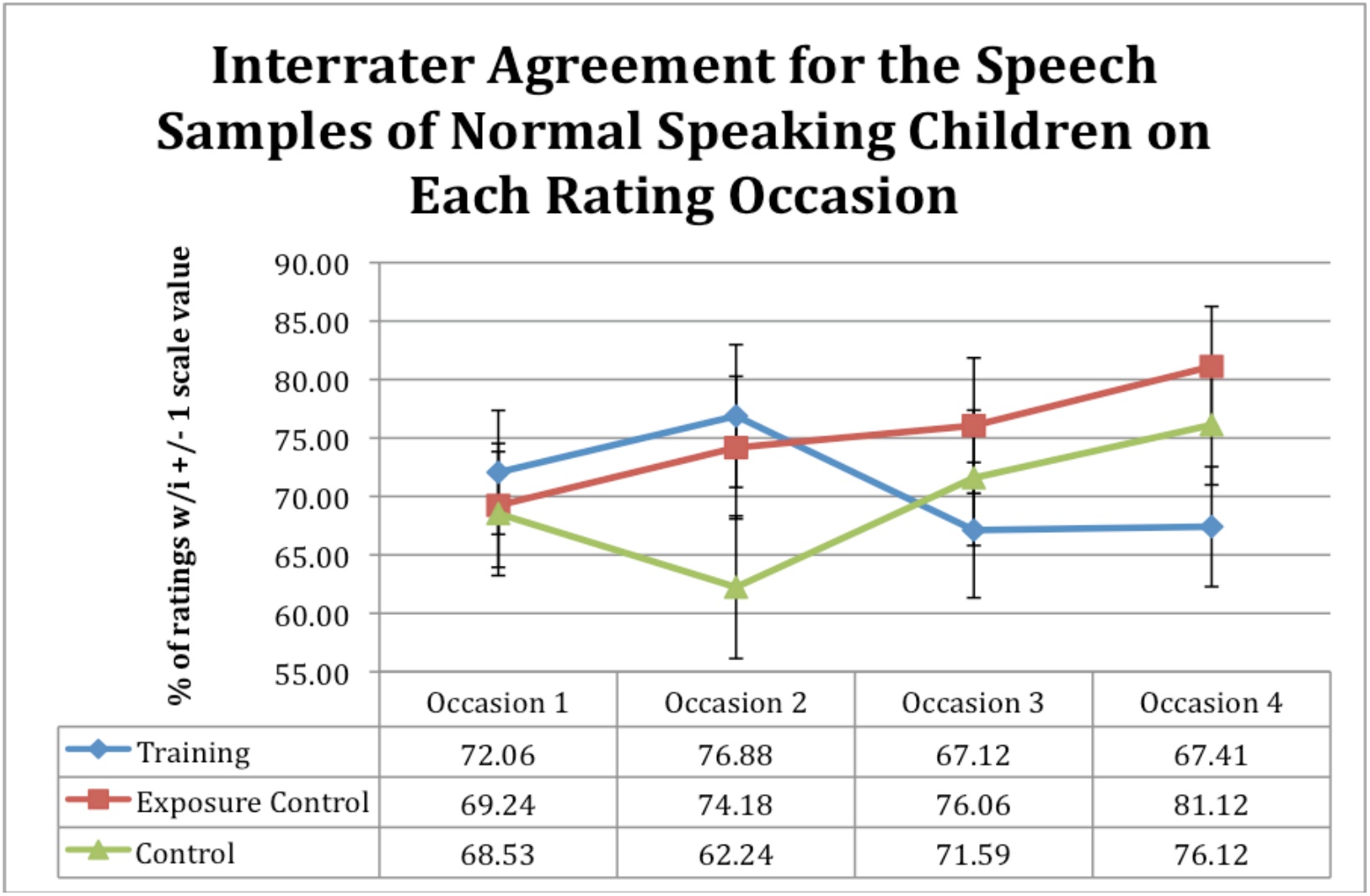



*Figure 9.* Study Two: Interrater agreement for participants separated by group for each rating occasion for normal speaking children only. Error bars for 95% confidence intervals are also presented.

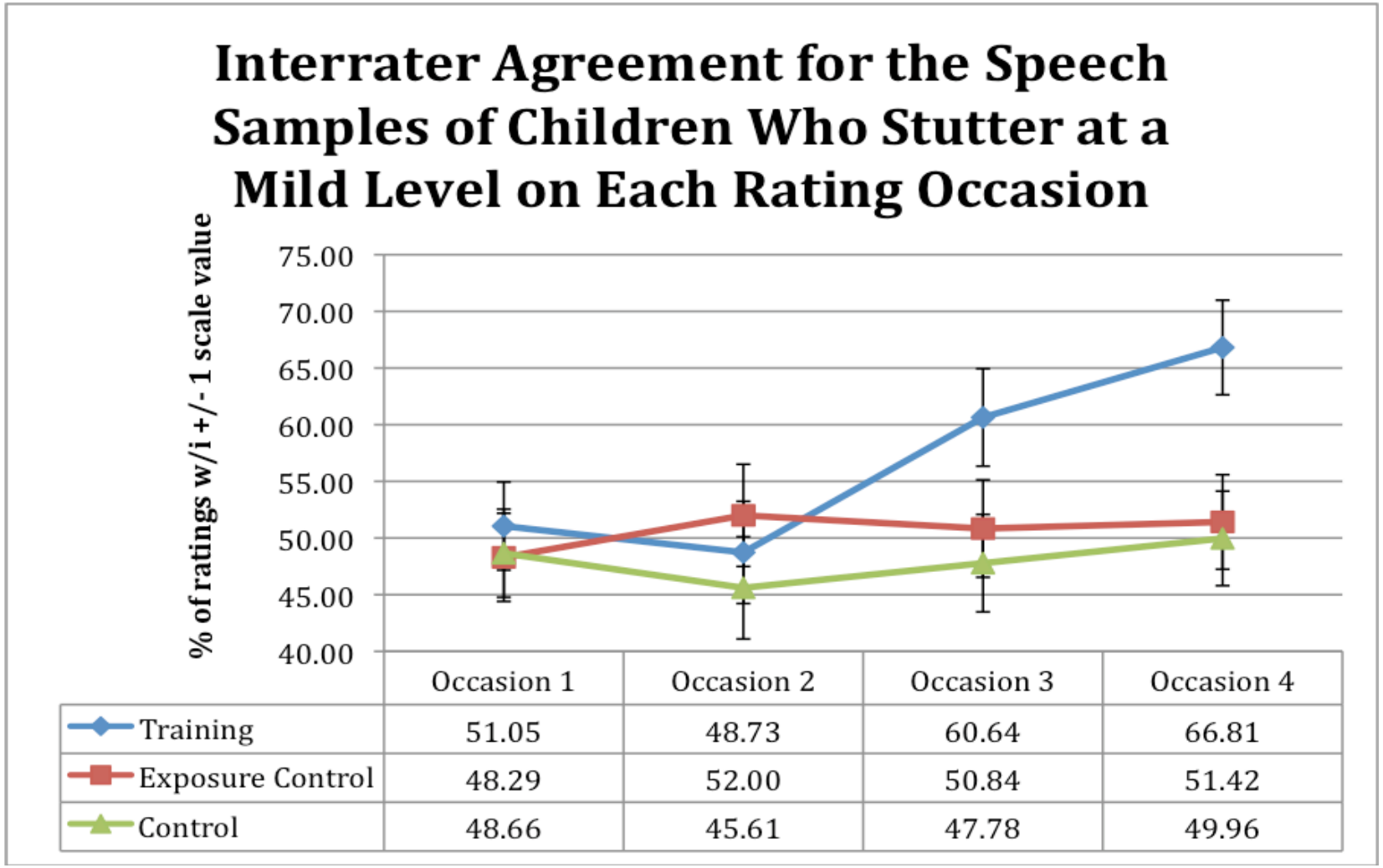



*Figure 10.* Study Two: Interrater agreement for participants separated by group for each rating occasion for children who stutter at a mild level only. Error bars for 95% confidence intervals are also presented.

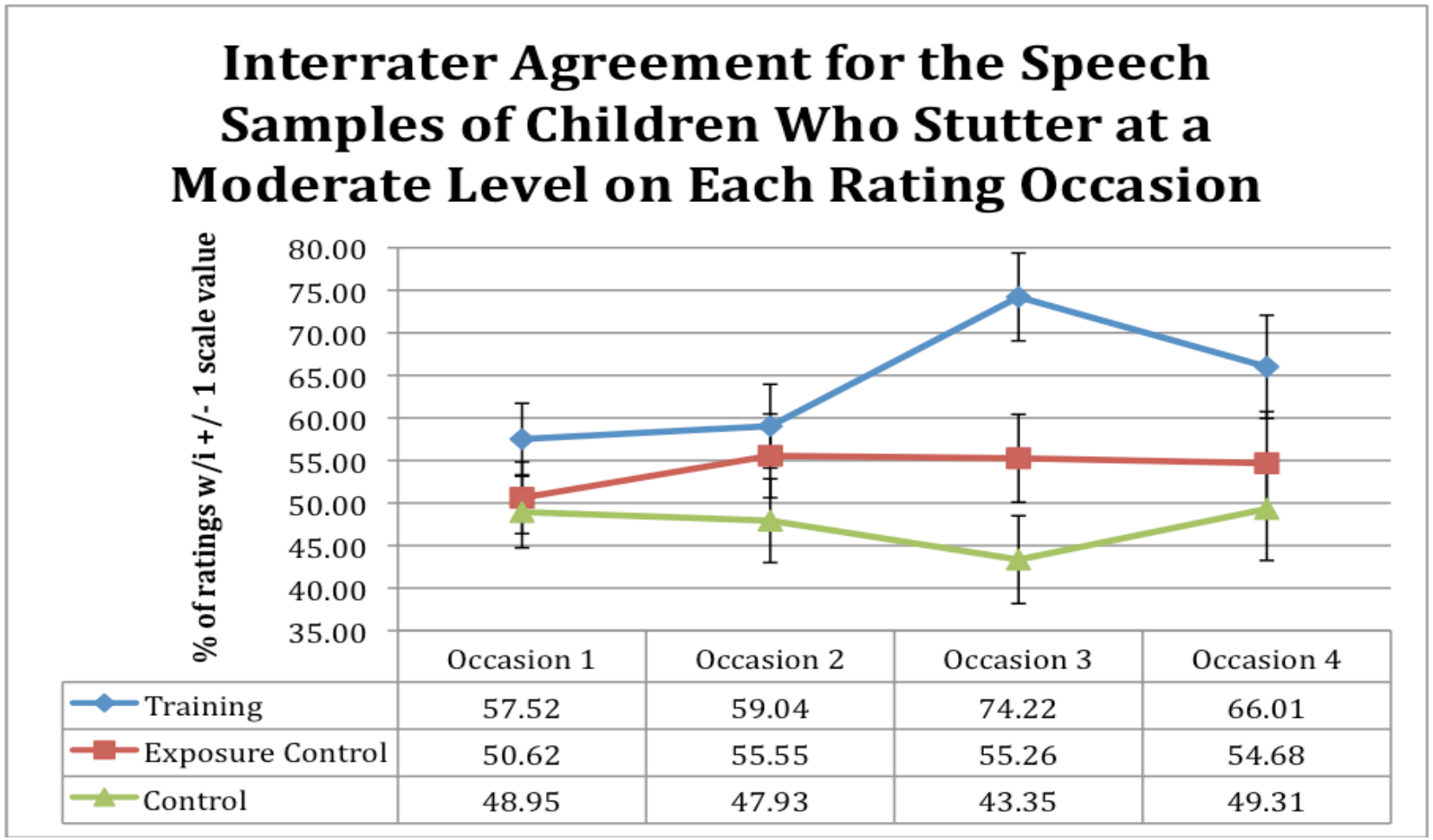



*Figure 11*. Study Two: Interrater agreement participants separated by group for each rating occasions for children who stutter at a moderate level only. Error bars for 95% confidence intervals are also presented.

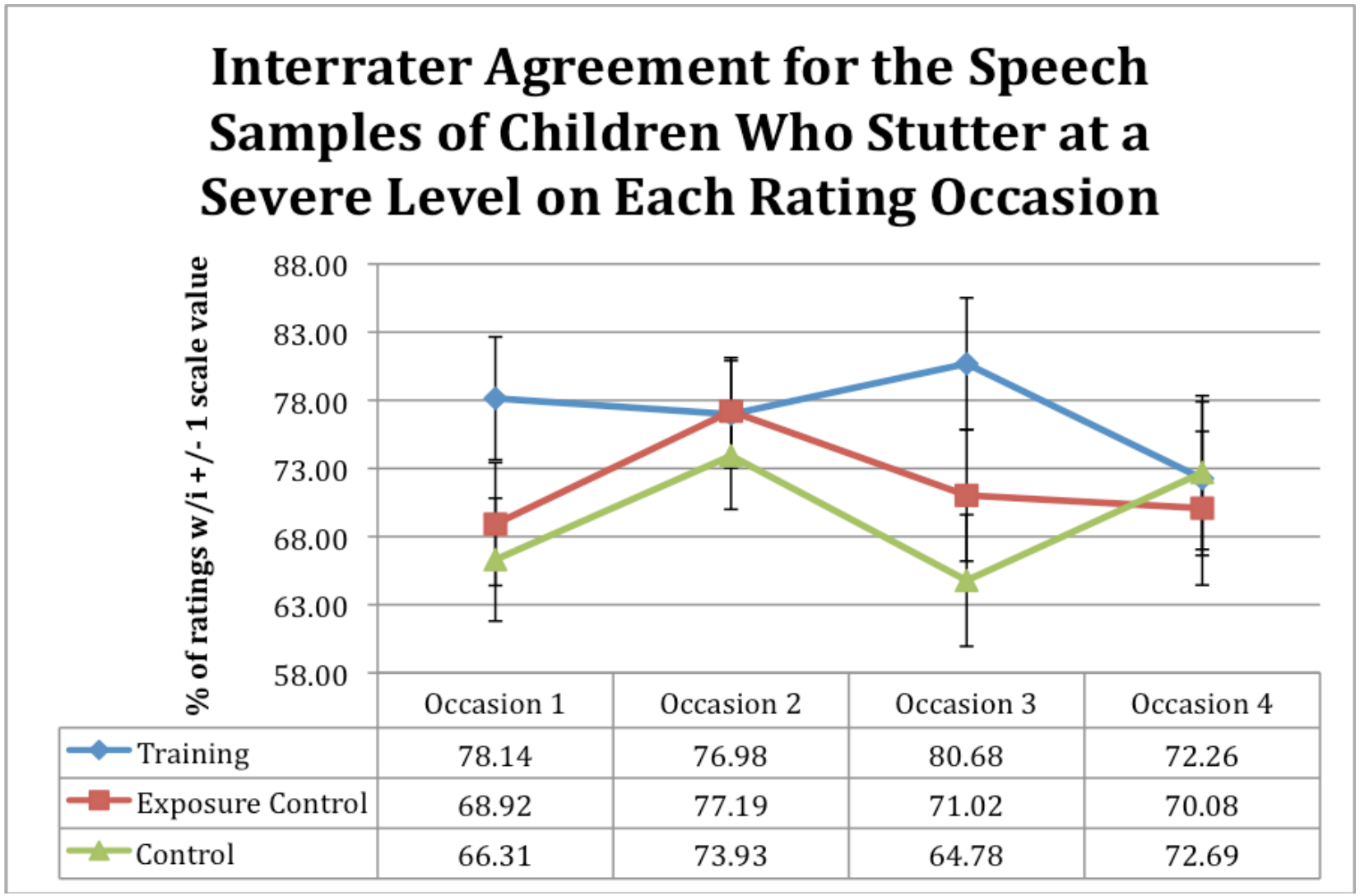
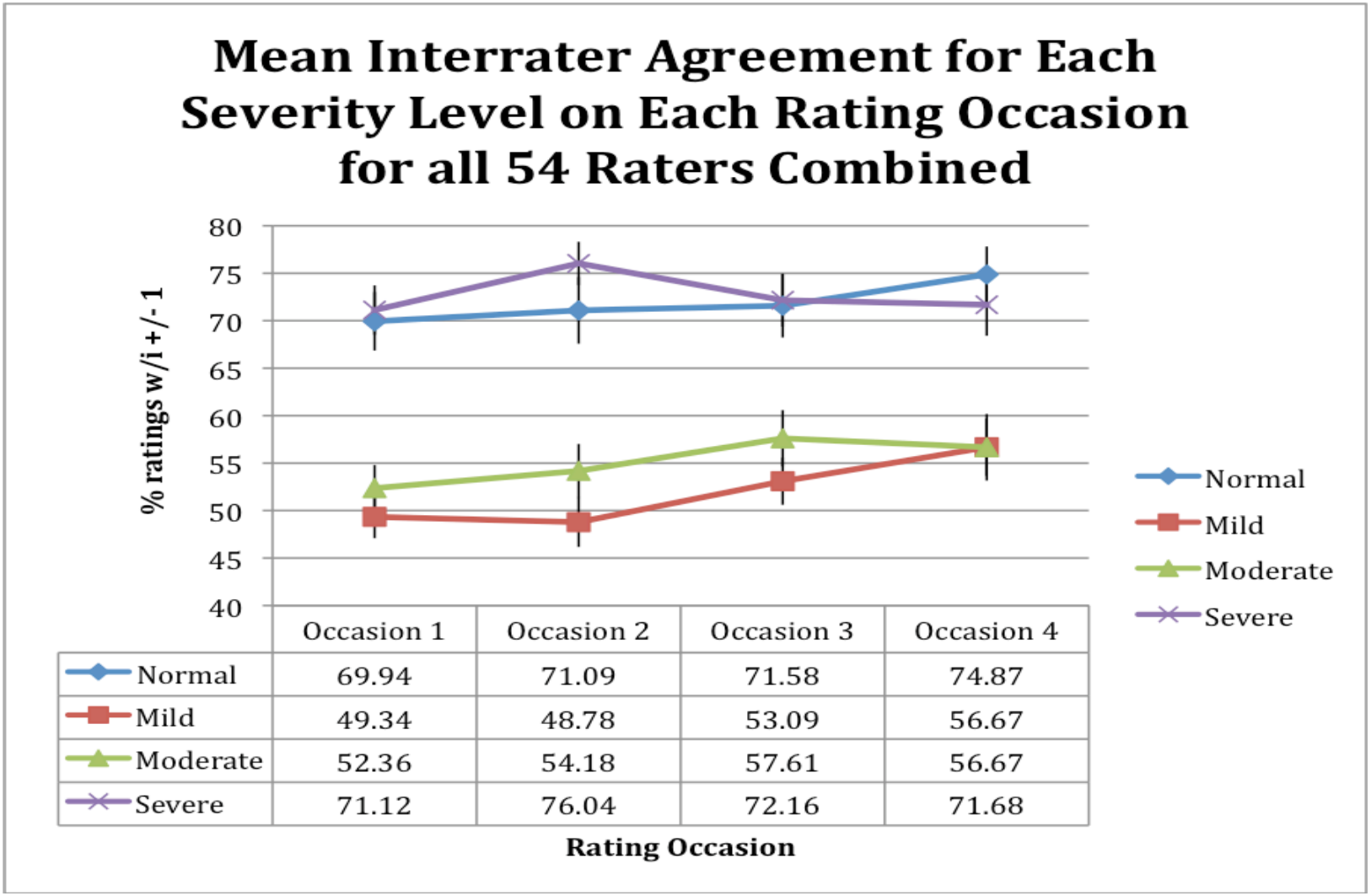



*Figure 12.* Study Two: Interrater agreement for participants separated by group for each rating occasions for children who stutter at a severe level only. Error bars for 95% confidence intervals are also presented.




| | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| Normal | 69.94 | 71.09 | 71.58 | 74.87 |
| Mild | 49.34 | 48.78 | 53.09 | 56.67 |
| Moderate | 52.36 | 54.18 | 57.61 | 56.67 |
| Severe | 71.12 | 76.04 | 72.16 | 71.68 |

*Figure 13.* Study Two: Mean interrater agreement levels for each stuttering severity level on each rating occasions for all groups combined (N=54). Error bars for 95% confidence intervals are also presented.
***Normal** speakers **significantly higher** then **mild** and **moderate** speakers; **mild** speakers **significantly lower** than **moderate** and **severe**; **moderate** speakers **significantly lower** than **severe**.

REFERENCES


Armson, J. & Kiefte, M. (2008). The effect of SpeechEasy on stuttering frequency, speech rate, and speech naturalness. *Journal of Fluency Disorders, 33,* 120-124.

Baer, D. M., Wolf, M. M. & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20,* 313-327.

Baker, L., & Pickering, J. (2010, November 19). *The effect of voice feminization therapy on listener and client perceptions of gender and vocal communication for MtF transsexuals*. Poster presented at the Annual Convention of the American Speech-Language Hearing Association, Philadelphia, PA.

Barlow, D. H. & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Benoit, K., Munson, B., Thurmes, A., Cordero, K. N., Baylis, A., & Moller, K. (2008, November). *Factors affecting speech naturalness in young adults with a history of cleft palate.* Poster session presented at the annual meeting of the American Speech Language and Hearing Association, Chicago, IL.

Block, S., Onslow, M., Packman, A., Gray, B., & Dacakis, G. (2005). Treatment of chronic stuttering: Outcomes from a student training clinic. *International Journal of Language and Communication Disorders, 40,* 455-466.

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

Bloodstein, O. (1950). A rating scale study of conditions under which stuttering is reduced or absent. *Journal of Speech and Hearing Disorders, 15,* 29-36.

Bothe, A. K. (2008). Identification of children's stuttered and nonstuttered speech by highly experienced judges: Binary judgments and comparisons with disfluency-types definitions. *Journal of Speech, Language, and Hearing Research, 51,* 867-878.

Bothe, A. K., Davidow, J. H., Bramlett, R. E., & Ingham, R. J. (2006). Stuttering treatment review 1970-2005: I. Systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. *American Journal of Speech-Language Pathology, 15,* 321-341.

Brewer, N. T. & Chapman, G. B. (2002). The fragile basic anchoring effect. *Journal of Behavioral Decision Making, 15,* 65-77.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment.* London: Sage University Press.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Conture, E. G. & Guitar, B. E. (1993). Evaluating efficacy of treatment of stuttering: School-age children. *Journal of Fluency Disorders, 18,* 253-287.

Cordes, A. K. (1993). *Studies of interjudge agreement for time-interval measurements of stuttering: Effects of interval duration and effects of training with highly-agreed or poorly-agreed exemplars.* Unpublished doctoral dissertation, University of California, Santa Barbara.

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

Cordes, A. K. (2000). Individual and consensus judgments of disfluency types in the speech of persons who stutter. *Journal of Speech, Language, and Hearing Research, 43,* 951-964.

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

Cordes, A. K. & Ingham, R. J. (1994b). Time-interval measurement of stuttering: Effects of training with highly agreed or poorly agreed exemplars. *Journal of Speech and Hearing Research, 37,* 1295-1307.

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

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

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

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

Costello, J. M. (1983). Current behavioral treatments for children. In D. Prins & R. J. Ingham (Eds.), *Treatment of stuttering in early childhood: Methods and issues* (pp. 69-111). San Diego: College Hills.

Costello, J. M. & Hurst, M. R. (1981). An analysis of the relationship among stuttering behaviors. *Journal of Speech and Hearing Research, 24,* 247-256.

Coughlin-Woods, S., Lehman, M. E., & Cooke, P. A. (2005). Ratings of speech naturalness of children ages 8-16 years. *Perceptual and Motor Skills, 100,* 295-304.

Craig, A., Hancock, K., Chang, E., McCready, C., Shepley, A., McCaul, A., et al. (1996). A controlled clinical trial for stuttering in persons aged 9 to 14 years. *Journal of Speech and Hearing Research, 39,* 808-826.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Thompson Learning.

Curlee, R. F. (1993). Evaluating treatment efficacy for adults: Assessment of stuttering disability. *Journal of Fluency Disorders, 18,* 319-331.

de Kinkelder, M. & Boelens, H. (1998). Habit-reversal treatment for children's stuttering: Assessment in three settings. *Journal of Behavior Therapy and Experimental Psychology, 29,* 261-265.

Druce, T., Debney, S., & Byrt, T. (1997). Evaluation of an intensive treatment program for stuttering in young children. *Journal of Fluency Disorders, 22,* 169-186.

Einarsdottir, J. (2009). *The identification and measurement of stuttering in preschool children*. Unpublished doctoral dissertation, University of Iceland, Reykjavik.

Einarsdottir, J. & Ingham, R. J. (2008). The effect of stuttering measurement training on judging stuttering occurrence in preschool children who stutter. *Journal of Fluency Disorders, 33,* 167-179.

Evitts, P., Webster, K., Starmer, H., Calberg, R., Miller, J., Bean, S., et al. (2010, November 19). *Listener impressions of voice, speech, and personality following supracricoid laryngectomy*. Poster presented at the Annual Convention of the American Speech-Language Hearing Association, Philadelphia, PA.

Fagel, W. P. F., van Herpt, L. W. A. C., & Boves, L. (1983). Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. *Speech Communication, 2,* 315-326.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2009). G*Power [Computer software]. Retrieved from www.psycho.uni-dusseldorf.de/abteilungen/aap/gpower3/

Few, L. R. & Lingwall, J. B. (1972). A further analysis of fluency within stuttered speech. *Journal of Speech and Hearing Research, 15,* 356-363.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage.

Finn, P. (1997). Adults recovered from stuttering without formal treatment: Perceptual assessment of speech normalcy. *Journal of Speech, Language, and Hearing Research, 40,* 821-831.

Finn, P. & Ingham, R. J. (1994). Stutterers' self-ratings of how natural speech sounds and feels. *Journal of Speech and Hearing Research, 37,* 326-340.

Finn, P., Ingham, R. J., Ambrose, N., & Yairi, E. (1997). Children recovered from stuttering without formal treatment: Perceptual assessment of speech normalcy. *Journal of Speech, Language, and Hearing Research, 40,* 867-876.

Fleiss, J. L. (1986). *Design and analysis of clinical experiments*. New York: Wiley.

Fowler, S. & Ingham, R. J. (1987). Stuttering Treatment Rating Recorder (STRR) [Computer software]. Santa Barbara: University of California, Santa Barbara.

Fox, P. T., Ingham, R. J., Ingham, J. C., Zamarripa, F., Xiong, J. H., & Lancaster, J. L. (2000). Brain correlates of stuttering and syllable production: A PET performance-correlation analysis. *Brain, 123,* 1985-2004.

Franken, M. C. (1987). Perceptual and acoustic evaluation of stuttering therapy. In H. F. M. Peters & W. Hulstijn (Eds.), *Speech motor dynamics in stuttering* (pp. 285-294). New York: Springer-Verlag.

Franken, M. C., Boves, L., Peters, H. F. M., & Webster, R. L. (1992). Perceptual evaluation of the speech before and after fluency shaping stuttering therapy. *Journal of Fluency Disorders, 17,* 223-241.

Franken, M. C., Boves, L., Peters, H. F. M., & Webster, R. L. (1995). Perceptual rating instrument for speech evaluation of stuttering treatment. *Journal of Speech and Hearing Research, 38,* 280-288.

Garcia, J.M. & Cannito, M.P. (1996). Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria. *Journal of Speech and Hearing Research, 39,* 750-760.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech, Language, and Hearing Research, 36,* 14-20.

Gooch, J. L., Hardin-Jones, M., Chapman, K. L., Trost-Cardamone, J. E., & Sussman, J. (2001). Reliability of listener transcriptions of compensatory articulations. *Cleft Palate-Craniofacial Journal, 38,* 59-67.

Gordon-Brannan, M. & Hodson, B. W. (2000). Intelligibility/severity measurements of prekindergarten children's speech. *American Journal of Speech-Language Pathology, 9,* 141-150.

Gow, M. L. & Ingham, R. J. (1992). Modifying electroglottograph-identified intervals of phonation: The effect on stuttering. *Journal of Speech and Hearing Research, 35,* 495-511.

Hancock, K., Craig, A., McCready, C., McCaul, A., Costello, D., Campbell, K. et al. (1998). Two-to-six-year controlled-trial stuttering outcomes for children and adolescents. *Journal of Speech, Language, and Hearing Research, 41,* 1242-1252.

Hearne, A., Packman, A. C., Onslow, M., & O'Brian, S. (2008). Developing treatment for adolescents who stutter: A Phase I trial of the Camperdown Program. *Language, Speech, and Hearing Services in Schools, 39,* 487-497.

Hewat, S., Onslow, M., Packman, A., & O'Brian, S. (2006). A Phase II clinical trial of self-imposed time-out treatment for stuttering in adults and adolescents. *Disability and Rehabilitation, 28,* 33-42.

Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Longman.

Hustad, K. C. & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12,* 198-208.

Ingham, J. C. & Ingham, R. J. (1987). Stuttering measurement training [videotape and manual]. Santa Barbara, CA: University of California, Santa Barbara.

Ingham, J. C. & Ingham, R. J. (2010). The Stuttering Measurement System (SMS) training workbook [Computer software manual]. Retrieved from http://www.coe.uga.edu/csse/sms/

Ingham, J. C. & Riley, G. (1998). Guidelines for documentation of treatment efficacy for young children who stutter. *Journal of Speech, Language, and Hearing Research, 41,* 753-770.

Ingham, R. J. (1984). *Stuttering and behavior therapy: Current status and experimental foundations*. San Diego: College-Hill.

Ingham, R. J. (1985). Speech naturalness and stuttering research: A review. In S. E. Gerber & G. T. Mencher (Eds.), *International perspectives on communication disorders* (pp. 168-180). Washington: Gallaudet University Press.

Ingham, R. J. (1993). Current status of stuttering and behavior modification-II. Principal issues and practices in stuttering therapy. *Journal of Fluency Disorders, 18,* 57-79.

Ingham, R. J., Bakker, K., Moglia, R., & Kilgo, M. (1999). Stuttering Measurement System (SMS) [Computer software]. University of California, Santa Barbara.

Ingham, R. J. & Cordes, A. K. (1997). Self-measurement and evaluating stuttering treatment efficacy. In R. F. Curlee & G. M. Siegel (Eds.), *Nature and treatment of stuttering: New directions* (pp. 413-437). Boston: Allyn and Bacon.

Ingham, R. J., Cordes, A. K., & Finn, P. (1993). Time-interval measurement of stuttering: Systematic replication of Ingham, Cordes, & Gow (1993). *Journal of Speech and Hearing Research, 36,* 1168-1176.

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

Ingham, R. J., Cordes, A. K., Kilgo, M., & Moglia, R. (1998). Stuttering measurement assessment and training (SMAAT) [Computer software]. Santa Barbara, CA: University of California, Santa Barbara.

Ingham, R. J. & Costello, J. (1985). Stuttering treatment outcome evaluation. In J. M. Costello (Ed.), *Speech disorders in adults* (pp. 189-223). San Diego: College-Hill Press.

Ingham, R. J., Fox, P. T., Ingham, J. C., Xiong, J., Zamarripa, F., Hardies, L. J. et al. (2004). Brain correlates of stuttering and syllable production: Gender comparison and replication. *Journal of Speech, Language, and Hearing Research, 47,* 321-341.

Ingham, R. J., Gow, M., & Costello, J. M. (1985). Stuttering and speech naturalness: Some additional data. *Journal of Speech and Hearing Disorders, 50,* 217-219.

Ingham, R. J., Ingham, J. C., Moglia, R. & Kilgo, M. (2010). Stuttering Measurement System (SMS) [Computer software]. Retrieved from http://www.coe.uga.edu/csse/sms/

Ingham, R. J., Ingham, J. C., Onslow, M., & Finn, P. (1989). Stutterers' self-ratings of speech naturalness: Assessing effects and reliability. *Journal of Speech and Hearing Research, 32,* 419-431.

Ingham, R. J., Kilgo, M., Ingham, J. C., Moglia, R., Belknap. H., & Sanchez, T. (2001). Evaluation of a stuttering treatment based on reduction of short phonation intervals. *Journal of Speech, Language, and Hearing Research, 44,* 1229-1244.

Ingham, R. J., Martin, R. R., Haroldson, S. K., Onslow, M., & Leney, M. (1985). Modification of listener-judged naturalness in the speech of stutterers. *Journal of Speech and Hearing Research, 28,* 495-504.

Ingham, R. J., Moglia, R. A., Frank, P., Ingham, J. C., & Cordes, A. K. (1997). Experimental investigation of the effects of frequency-altered feedback on the speech of adults who stutter. *Journal of Speech, Language, and Hearing Research, 40,* 361-372.

Ingham, R. J. & Onslow, M. (1985). Measurement and modification of speech naturalness during stuttering therapy. *Journal of Speech and Hearing Disorders, 50,* 261-281.

Ingham, R. J. & Packman, A. C. (1978). Perceptual assessment of normalcy of speech following stuttering therapy. *Journal of Speech and Hearing Research, 21,* 63-73.

Ingham, R. J., Sato, W., Finn, P., & Belknap, H. (2001). The modification of speech naturalness during rhythmic stimulation during rhythmic stimulation treatment of stuttering. *Journal of Speech, Language, & Hearing Research, 44,* 841-852.

Ingham, R. J., Warner, A., Byrd, A., & Cotton, J. (2006). Speech effort measurement and stuttering: Investigating the chorus reading effect. *Journal of Speech, Language, and Hearing Research, 49,* 660-670.

James, J. E. (1981a). Punishment of stuttering: Contingency and stimulus parameters. *Journal of Communication Disorders, 14,* 375-386.

James, J. E. (1981b). Self-monitoring of stuttering: Reactivity and accuracy. *Behavior Research Therapy, 19,* 291-296.

James, J. E. (1983). Parameters of the influence of self-initiated time-out from speaking on stuttering. *Journal of Communication Disorders, 16,* 123-132.

Jones, R. J. & Azrin, N. H. (1969). Behavioral engineering: Stuttering as a function of stimulus duration during speech synchronization. *Journal of Applied Behavior Analysis, 2,* 223-229.

Keintz, C. (2007, November 15). *Influence of visual information on the speech intelligibility in bilateral facial paralysis*. Poster presented at the Annual Convention of the American Speech-Language Hearing Association, Boston, MA.

Kalinowski, J., Noble, S., Armson, J., & Stuart, A. (1994). Pretreatment and posttreatment speech naturalness ratings of adults with mild and severe stuttering. *American Journal of Speech-Language Pathology, 3,* 61-66.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kearns, K. J. (1990). Reliability of procedures and measures. In L. B. Olswang, C. K. Thompson, S. F. Warren & N. J. Mingetti (Eds.), *Treatment efficacy research in communication disorders* (pp. 79-90). Rockville, MD: American Speech-Language-Hearing Foundation.

Keuning, K., Wieneke, G. H., & Dejonckere, P. H. (1999). The intrajudge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: The effect of judges and speech samples. *Cleft Palate-Craniofacial Journal, 36,* 328-333.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103-115.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512-520.

Krikorian, C. M. & Runyan, C. M. (1983). A perceptual comparison: Stuttering and nonstuttering children's nonstuttered speech. *Journal of Fluency Disorders, 8,* 283-290.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93,* 586-595.

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

Logan, K. J., Roberts, R. R., Pretto, A. P., & Morey, M. J. (2002). Speaking slowly: Effects of four self-guided training approaches on adults' speech rate and naturalness. *American Journal of Speech and Hearing Research, 11,* 163-174.

Mackey, L. S., Finn, P., & Ingham, R. J. (1997). Effect of speech dialect on speech naturalness ratings: A systematic replication of Martin, Haroldson, and Triden (1984). *Journal of Speech, Language, and Hearing Research, 40,* 349-360.

Martin, R. R. & Haroldson, S. K. (1992). Stuttering and speech naturalness: Audio and audiovisual judgments. *Journal of Speech and Hearing Research, 35,* 521-528.

Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders, 49,* 53-58.

Max, L. & Gracco, V. L. (2005). Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter. *Journal of Speech, Language, and Hearing Research, 48,* 524-542.

Maxwell, D. L. & Satake, E. (2006). *Research and statistical methods in communication science and disorders.* Canada: Thomson Delmar Learning.

McGraw, K.O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

Merenbach, A. (2007). Randomness [Computer software]. Retrieved from http://mac.wareseeker.com/Utilities/randomness-1.5.2.zip/334231

Metz, D. E., Onufrak, J. A., & Ogburn, R. S. (1979). An acoustical analysis of stutterer's speech prior to and at the termination of speech therapy. *Journal of Fluency Disorders, 4,* 249-254.

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychophysical dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55,* 516-525.

Morrow, J. R. & Jackson, A. W. (1993). How "significant" is your reliability? *Research Quarterly for Exercise and Sport, 64,* 352-355.

Nichols, A. C. (1966). Audience ratings of the "naturalness" of spoken and written sentences. *Speech Monographs, 33,* 156-159.

Nicolosi, L., Harryman, E., & Kresheck, J. (2004). *Terminology of communication disorders: Speech-language-hearing* (5th ed.). Philadelphia: Lippincott Williams & Wilkins.

O'Brian, S., Cream, A., Onslow, M., & Packman, A. (2001). A replicable, non-programmed, instrument-free method for the control of stuttering with prolonged speech. *Asia Pacific Journal of Speech, Language, and Hearing, 6,* 91-96.

O'Brian, S., Onslow, M., Cream, A., & Packman, A. (2003). The Camperdown Program: Outcomes of a new prolonged-speech treatment model. *Journal of Speech, Language, and Hearing Research, 46,* 933-946.

O'Brian, S., Packman, A., Onslow, M., Cream, A., O'Brian, N., & Bastock, K. (2003). Is listener comfort a viable construct in stuttering research? *Journal of Speech, Language, and Hearing Research, 46,* 503-509.

O'Brien, S., Packman, A., Onslow, M., & O'Brien, N. (2003). Generalizability theory II: Application to perceptual scaling of speech naturalness in adults who stutter. *Journal of Speech, Language, and Hearing Research, 46,* 718-723.

Onslow, M., Adams, R., & Ingham, R. (1992). Reliability of speech naturalness ratings of stuttered speech during treatment. *Journal of Speech and Hearing Research, 35,* 994-1001.

Onslow, M., Costa, L., Andrews, C., Harrison, E., & Packman, A. (1996). Speech outcomes of a prolonged-speech treatment for stuttering. *Journal of Speech and Hearing Research, 39,* 734-749.

Onslow, M., Costa, L., & Rue, S. (1990). Direct early intervention with stuttering: Some preliminary data. *Journal of Speech and Hearing Disorders, 55,* 405-416.

Onslow, M., Hayes, B., Hutchins, L., & Newman, D. (1992). Speech naturalness and prolonged-speech treatments for stuttering: Further variables and data. *Journal of Speech and Hearing Research, 35,* 274-282.

Onslow, M. & Ingham, R. J. (1987). Speech quality measurement and the management of stuttering. *Journal of Speech and Hearing Disorders, 52,* 2-17.

Packman, A & Onslow, M. (1998). The behavioral data language of stuttering. In A.K. Cordes & R. J. Ingham (Eds.), *Treatment efficacy for stuttering: A search for empirical biases* (pp. 27-50). San Diego, CA: Singular.

Packman, A., Onslow, M., & van Doorn, J. (1994). Prolonged speech and modification of stuttering: Perceptual, acoustic, and electroglottographic data. *Journal of Speech, Language, and Hearing Research, 37,* 724-737.

Parrish, W. M. (1951). The concept of "naturalness". *Quarterly Journal of Speech, 37,* 448-450.

Perkins, W. H. (1973). Replacement of stuttering with normal speech: I. Rationale. *Journal of Speech and Hearing Disorders, 38,* 283-294.

Perkins, W. H., Rudas, J., Johnson, L., Michael, W., & Curlee, R. F. (1974). Replacement of stuttering with normal speech: III. Clinical effectiveness. *Journal of Speech and Hearing Disorders, 4,* 416-428.

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

Qu, C., Zhou, L. & Luo, Y. (2008). Electrophysiological correlates of adjustment process in anchoring effects. *Neuroscience Letters, 445,* 199-203.

Ramig, P. R. (1984). Rate changes in the speech of stutterers after therapy. *Journal of Fluency Disorders, 9,* 285-294.

Rousseau, I., Onslow, M., Packman, A., & Jones, M. (2008). Comparisons of audio and audiovisual measures of stuttering frequency and severity in preschool-age children. *American Journal of Speech-Language Pathology, 17,* 173-178.

Runyan, C. M. & Adams, M. R. (1978). Perceptual study of the speech of "successfully therapeutized" stutterers. *Journal of Fluency Disorders, 3,* 25-39.

Runyan, C. M. & Adams, M. R. (1979). Unsophisticated judges' perceptual evaluations of the speech of "successfully treated" stutterers. *Journal of Fluency Disorders, 4,* 29-38.

Runyan, C. M., Bell, J. N., & Prosek, R. A. (1990). Speech naturalness ratings of treated stutterers. *Journal of Speech and Hearing Disorders, 55,* 434-438.

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.). *Intelligibility in speech disorders: Theory, measurement, and management* (pp. 11-34). Philadelphia: John Benjamins Publishing Company.

Schiavetti, N., Martin, R. R., Haroldson, S. K., & Metz, D. E. (1994). Psychophysical analysis of audiovisual judgments of speech naturalness of nonstutterers and stutterers. *Journal of Speech and Hearing Research, 37,* 46-52.

Schiavetti, N. & Metz, D. E. (1997). Stuttering and the measurement of speech naturalness. In R. F. Curlee & G. M. Siegel (Eds*.). Nature and treatment of stuttering: New directions* (2nd ed., pp. 398-412). Boston: Allyn and Bacon.

Schiavetti, N., Sacco, P. R., Metz, D. E., & Sitler, R. W. (1983). Direct magnitude estimation and interval scaling of stuttering severity. *Journal of Speech and Hearing Research, 26,* 568-573.

Schissel, R. J. & Flournoy, J. E. (1978). An investigation of the variability of judgments of experienced and inexperienced listeners in their use of a screening test of articulation. *Journal of Communication Disorders, 11,* 459-468.

Shriberg, L. D. (1972). Articulation judgments: Some perceptual considerations. *Journal of Speech and Hearing Research, 15,* 876-882.

Shriberg, L. D. & Kent, R. D. (1995). *Clinical Phonetics* (2nd ed.). Boston: Allyn & Bacon.

Shriberg, L. D., Kwiatkowski, J., & Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27,* 456-465.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Siegel, G. M. (1962). Experienced and inexperienced articulation examiners. *Journal of Speech and Hearing Disorders, 27,* 28-35.

Stevens, S. S. (1975). *Psychophysics.* New York: Wiley.

Stokes, T. F. & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10,* 349-367.

Stokes, T. F. & Osnes, P. G. (1989). An operant pursuit of generalization. *Behavior Therapy, 20,* 337-355.

Stuart, A. & Kalinowski, J. (2004). The perception of speech naturalness of post-therapeutic and altered auditory feedback speech of adults with mild and severe stuttering. *Folia Phoniatrica et Logopaedica, 56,* 347-357.

Stuart, A. Kalinowski, J., Rastatter, M. P., Saltuklaroglu, T., & Dayalu, V. (2004). Investigations of the impact of altered auditory feedback of people who stutter: Initial fitting and 4-month follow-up. *International Journal of Language and Communication Disorders, 39,* 93-113.

Stuart, A., Kalinowski, J., Saltuklaroglu, T., & Guntupalli, V. K. (2006). Investigations of the impact of altered auditory feedback in-the-ear devices on the speech of people who stutter: One-year follow-up. *Disability and Rehabilitation, 28,* 757-765.

Tasko, S. M., McClean, M. D., & Runyan, C. M. (2007). Speech motor correlates of treatment related changes in stuttering severity and speech naturalness. *Journal of Communication Disorders*, *40,* 42-65.

Teesson, K., Packman, A., & Onslow, M. (2003). The Lidcombe Behavioral Data Language of stuttering. *Journal of Speech, Language, and Hearing Research, 46,* 1009-1015.

Tinsley, H. E. A. & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358-376.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Van Borsel, J. & Eeckhout, H. (2008). The speech naturalness of people who stutter under delayed auditory feedback as perceived by different groups of listeners. *Journal of Fluency Disorders, 33,* 241-251.

Van Riper, C. (1973). *The treatment of stuttering.* Englewood Cliffs: New Jersey.

Wendahl, R. W. & Cole, J. (1961). Identification of stuttering during relatively fluent speech. *Journal of Speech and Hearing Research, 4,* 281-286.

Wenker, R. B., Wegener, J. G., & Hart, K. J. (1996). The impact of presentation mode and disfluency on judgments about speakers. *Journal of Fluency Disorders, 21,* 147-159.

Williams, D. E., Wark, M., & Minifie, F. D. (1963). Ratings of stuttering by audio, visual, and audiovisual cues. *Journal of Speech and Hearing Research, 6,* 91-100.

Yaruss, J. S. (1998). Real-time analysis of speech fluency: Procedures and reliability training. *American Journal of Speech-Language Pathology, 7,* 25-36.

Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders, 55,* 550-560.

Young, M. A. (1964). Identification of stutterers from recorded samples of their "fluent" speech. *Journal of Speech and Hearing Research, 7,* 291-303.

Young, M. A. (1969a). Observer agreement: Cumulative effects of rating many samples. *Journal of Speech and Hearing Research, 12,* 135-143.

Young, M. A. (1969b). Observer agreement: Cumulative effects of repeated ratings of the same samples and of knowledge of group results. *Journal of Speech and Hearing Research, 12,* 144-155.

Zyski, B. J. & Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders, 20,* 367-378.

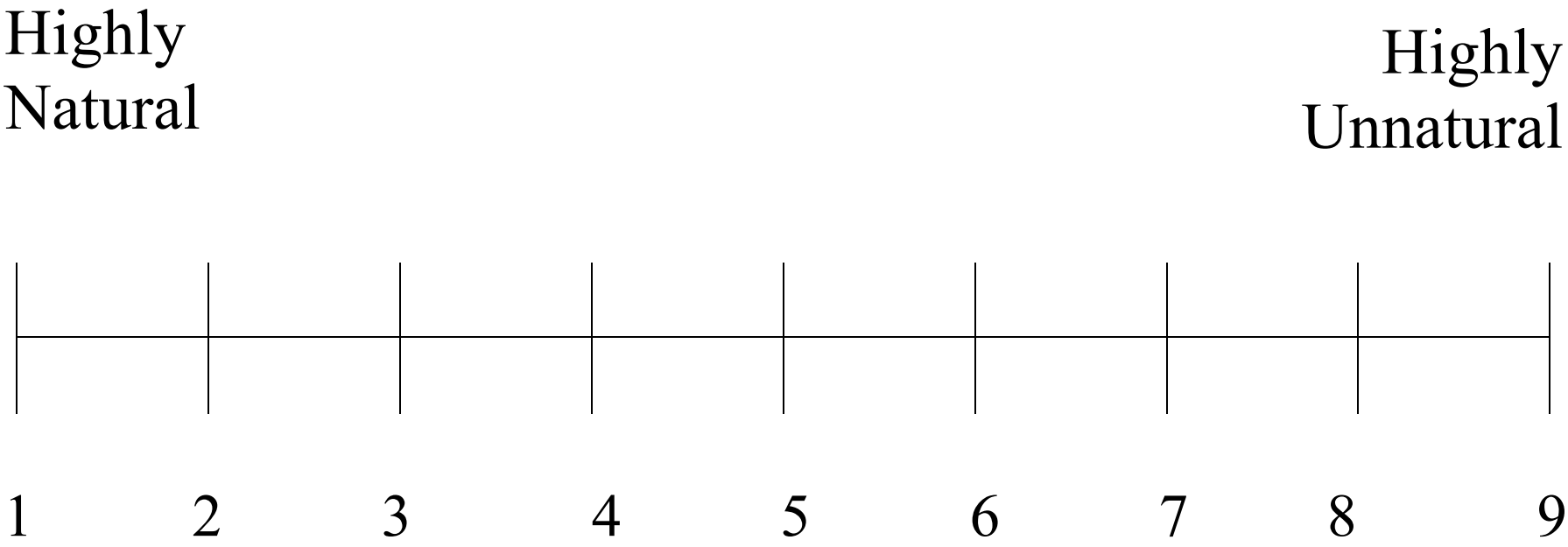
APPENDIX A

FRANKEN AND COLLEAGUES' SPEECH QUALITY STUDIES

| Study | Analysis | Factors | % variance explained | Scales in factor |
|---|---|---|---|---|
| 1987 | Factor | Speech Rate & General Quality | 64.2 | slow – quick<br>dragging – brisk<br>unnatural – natural<br>unpleasant – pleasant<br>ugly – beautiful |
| | | Articulation Quality | 15.9 | slovenly – polished<br>broad – cultured<br>ugly – beautiful<br>unpleasant – pleasant |
| | | Voice Dynamics | 7.7 | monotonous – melodious<br>expressive – expressionless |
| | | Voice Pitch & Static Voice Quality | 7.4 | low pitch – high pitch<br>deep – shrill<br>dull – clear |
| | | Potency Factor | 4.8 | weak – powerful<br>soft – loud |
| 1992 | Discriminant | Distorted Speech Dimension | 55.8 | halting – fluent<br>tense – relaxed<br>unnatural – natural |
| | | Dynamics/ Prosody Dimension | 39.6 | monotonous – melodious<br>flat – expressive<br>weak – strong accentuation |
| | | Voice Dimension | 4.6 | soft – loud<br>low pitch – high pitch |
| 1995 | Factor | Voice Dynamics | 57.4 | weak – powerful<br>weak – strong accentuation<br>flat – expressive<br>monotonous – melodious<br>soft – loud<br>slow – quick |
| | | Articulation Quality | 21.2 | fluent – halting<br>tense – relaxed<br>slurred – precise<br>slovenly – polished<br>unpleasant – pleasant<br>unnatural – natural |
| | | Pitch | 7.9 | shrill – deep<br>low pitch – high pitch |

APPENDIX B

SPEECH NATURALNESS RATING SCALE



APPENDIX C

STUDIES COMPARING THE SPEECH NATURALNESS OF PEOPLE WHO STUTTER AND PEOPLE WHO DO NOT STUTTER USING MARTIN, HAROLDSON, & TRIDEN'S (1984) 9-POINT SCALE

| Study | Speakers (N) | Listeners (N) | # of Speech Samples | Speech Sample Length | Speech samples include stutters | PWS Mean Nat. Rating | PNS Mean Nat. Rating | Raters Intrajudge Agreement | Raters Interjudge Agreement |
|---|---|---|---|---|---|---|---|---|---|
| Armson & Kiefte (2008) | 14 PWS w/ & w/o Speech-Easy | 30 naïve 15 rated reading & 15 monologue | 56 | 1-min | Yes | Read =5.3 (2.19) ReadSE=3.3 (1.46) Mono=5.5 (1.99) ReadSE=3.2 (1.4) | n/a | - | ICC Read=.982 Read SE=.959 Mono=.954 Mono SE=.941 |
| Block, Onslow, Packman, Gray, & Dacakis (2005) | 78 PWS pre\post smooth speech tx | 48 undergrads | 77 | 15-sec | No | Pre=3.8(1.3) Post=4.5 (1.3) | n/a | - | - |
| Finn (1997) | 15 PWUS 15 PNS | 10 trained grad students | 30 | 1-min | Yes (if there was any) | 2.36 (1) | 1.77 (.5) | 93.4% w/i +/- 1 | 64.7% w/i +/- 1 |
| Finn, Ingham, Ambrose, & Yari (1997) | 10 CWUS 10 CNS | 26 unsoph 14 exp SLPs | 20 | 1-min | Yes | SLPs= 3.71 (1.29) Unsoph= 3.82 (1.22) | SLPs= 3.24 (.97) Unsop =3.82 (1.05) | °SLPs= 65% Unsoph= 64% | °SLPs= 45.1% Unsoph= 40.6% |
| Hewat, Onslow, Packman, & O'Brian (2006) | 18 PWS in tx 11 PNS | 10 unspoh | 41 | 15-sec | No | Subjects= 1.5-4.5 PS tx=1.5-9 | 1-4.2 | - | °33.5%+/-0 71.4 %+/-1 90.6%+/-2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 12 PWS in prior tx | | | | | | | | |
| Ingham, Gow, & Costello (1985) | 15 PWS (tx)<br>15 PNS | 30 undergrads | 30 (1/ speaker) | 1-min | No | 4.26 | 2.39 | - | - |
| Ingham, Sato, Finn, & Belknap (2001) Study 2 | 3 PWS (data only for 2- poor rel)<br>2 PNS | 20 undergrads | 36 | 1-min | Yes | CF 2-8.8<br>HV 1.5-9 | 2-4<br>1.2-1.8 | - | Via visual inspection of SS graph |
| Ingham, Warner, Byrd, & Cotton (2006) | 12 PWS<br>12 PNS | 1 research assistant | 42 | 1-min | Yes | BR1= 5.51<br>CR1= 3.47<br>BR2= 4.89<br>CR2= 3.38<br>BR3= 5.12 | BR1= 1.44<br>CR1= 3.33<br>BR2= 1.25<br>CR2= 3.22<br>BR3= 1.28 | - | °87.2% |
| Mackey, Finn, & Ingham (1997) | 10 PWS (no dialect)<br>10 PNS (no dialect)<br>10 PNS (dialect) | 30 undergrads | 30 (10 from each group) | 1-min | Yes | 5.59 | 3.43 dial<br>1.71 no | °76% = PWS<br>80% = PNS-D<br>98% = PNS | °59% = PWS<br>56%= PNS-D<br>80% = PNS |
| Martin & Haroldson (1992) | 10 PWS<br>10 PNS | 24 undergrads no exp w/ stuttering | 10 PWS<br>10 PNS | 1-min | Yes (doesn't say, so assume) | Audio 6.04 (1.97)<br>AV 6.81 (2.01) | Audio 2.3 (1.21)<br>AV 2.27 (1.18) | °PWSA 43/92 PWSAV 55/89 PNSA 66/97 PWSAV | °PWSA 40% +/- 0; 80% +/- 1 PWSAV 41% +/- 0; 79% +/-1 PNSA |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 57/94 | 40% +/- 0; 80% +/- 1 PNSAV 40% +/- 0; 81% +/- 1 ICC PWSA= .97; PWSAV = .96; PNSA =.77; PNSAV= .74 |
| Martin, Haroldson, & Triden (1984) | 10 PWS 10 DAF PWS 10 PNS | 30 undergrads | 30 (10 from each group) | 1-min | Yes | 6.52 (2) DAF=5.84 (1.79) | 2.12 (1.17) | +/-0 PWS 45%, DAF 44%, PNS 54% +/- 1 PWS 90%, DAF 85%, PNS 89% | +/-0 PWS 32%, DAF 32%, PNS 31% +/- 1 PWS 74%, DAF 77%, PNS 75% ICC Group PWS-.98; DAF-.98; PNS-.75 ICC Indiv PWS-.74; DAF-.57; PNS-.10 |
| Metz, Schiavetti, & Sacco (1990) | 20 PWS post-tx 20 PNS | 30 SLP undergrads 15 used interval & | 40 PWS 40 PNS 1 reading | 30-sec | No | Reading 5.91 (1.6) Speech 5.92 (1.97) | Reading 2.84 (1.11) Speech | - | ICC Group Reading .984; Speech |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 15 used DME only interval presented. | 1 speech? | | | | 3.55 (1.51) | | .982 ICC Individ. Reading .793; Speech .695 |
| O'Brian, Onslow, Cream, & Packman (2003) | 16 PWS tx 18 PNS | 3 unspoh | 36 | 15-sec | No | 4.5 (1.9) | 3.6 (2.1) | 100% w/i +/- 1 | - |
| O'Brian, Packman, Onslow, & O'Brian (2003) | 10 PWS 10 PNS | 15 unspoh | 10 pre 10 post 10 control | 30-sec | Yes | Means not reported | Means not reported | - | - |
| O'Brian, Packman, Onslow, Cream, O'Brian, & Bastock (2003) | 10 PWS pre/post PS 10 PNS | 15 untrained | 40 | 30-sec | Yes | Pre 5.57 Post 2.29 | 1.4 | - | ICC .71 68% +/- 1 |
| Onslow, Hayes, Hutchines, & Newman (1992) Study 1 | 7 PWS 7 PNS | 29 undergrads | 84 | 30-sec | No | 5.49 (1.01) | 3.25 (.77) | 22/29 rated at least 9/11 sample w/i +/-1 | - |
| Runyan, Bell, & Prosek (1990) | PWS PNS # not reported | 10 SLP grads | 280 | Not given | No | 3.86 (1.7) | 2.79 (1.4) | - | Spearman Brown coeff .91 |
| Schiavetti, Martin, Haroldson, & Metz (1994) | 10 PWS 10 PNS | 40 undergrads | 10 PWS 10 PNS | 1-min | Yes doesn't say but assume | Mean not reported just ICC for interval scale | Mean not reported just ICC for | - | ICC Group .997 Individual .935 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | interval scale | | |
| Stuart & Kalinowski (2004) | 20 PWS DAF, FAF, & no DAF 5 PNS | 35 naïve | 55 | 15-sec | Yes | DAF Mild 3.3; Sev. 4.5; FAF Mild 3.1; Sev. 3.8; NAF Mild 5.3; NAF Sev. 6.2; Pre-tx Mild 3.4; Sev. 5.4 Post-tx Mild 7; Sev. 7.8 | 1.5 | Spearman Rank Corr b/t 2 ratings .78 +/- 0 37.3% +/- 1 70% | - |
| Stuart, Kalinowski, Saltuklaroglu, & Guntupalli (2006) | 9 PWS w/ & w/o SE 5 PNS | 27 undergrads | 75; 70 PWS & 5 PNS | 15-sec | Yes | Initial SE Read=4, Mono=4; SE 4-mo Read=4.5, Mono=5; SE 12-mo Read=2.5, Mono=3.5 Initial No SE Read= 6.4, Mono =7.5; No SE 12-mo Read=7, Mono= 7.75 | Reading 1.5 | - | ICC=.70 |
| Van Borsel & Eeckhout (2008) | 8 PWS DAF 8 PNS | 14 SLPs 14 PWS 14 naïve | 16 1 from each | 15-sec | No | SLPs 6.64 (1.14); PWS 5.46 | SLPs 1.84 (.83); | - | - |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | speaker | | | (1.05); Naïve 7.32 (.83) | PWS 2.34 (.85); naïve 2.76 (1.14) | | |

*Authors don't report. This is based on description of number of tapes and ratings per tape, but I'm honestly not sure.
° percent of ratings within +/- 1 scale value of each other

A=Audio
AV=Audiovisual
BR=Base Rate
CNS=Children Who Never Stuttered
CR= Chorus Reading
CWUS=Children Who Used to Stutter
DAF=Delayed Auditory Feedback
DME=Direct Magnitude Estimation
FAF=Frequency Altered Feedback
ICC=Intraclass Correlation Coefficient
PWUS=People Who Used to Stutter
SLP=Speech-Language Pathologist
SS=Single Subject Research Design
PS=Prolonged Speech
NAF= No Altered Feedback
PNS= People Who Never Stuttered
PWS=People Who Stutter

APPENDIX D

INTERRATER AGREEMENT IN STUDIES MEASURING SPEECH NATURALNESS USING MARTIN, HAROLDSON, & TRIDEN'S (1984) 9-POINT SCALE

| Study | Speakers (N) | Listeners (N) | # of Speech Samples | Speech Sample Length | Speech samples include stutters | Raters Interjudge Agreement |
|---|---|---|---|---|---|---|
| Coughlin-Woods, Lehman, & Cooke (2005) | 60 CNS & 10 CW various communication disorders | 39 naïve only 26 used in data analysis | 70 | 30-sec | n/a except for 2 children | 26/32 had 60% of samples w/i +/- 1. |
| Finn (1997) | 15 PWUS<br>15 PNS | 10 trained grad students | 30 | 1-min | Yes (if there was any) | 64.7% w/i +/- 1 |
| Finn, Ingham, Ambrose, & Yari (1997) | 10 CWUS<br>10 CNS | 26 unsoph<br>14 exp SLPs | 20 | 1-min | Yes | °SLPs= 45.1%<br>Unsoph= 40.6% |
| Hewat, Onslow, Packman, & O'Brian (2006) | 18 PWS in tx<br>11 PNS<br>12 PWS in prior tx | 10 unspoh | 41 | 15-sec | No | 33.5%+/-0<br>71.4 %+/-1<br>90.6%+/-2 |
| Ingham, Moglia, Frank, Ingham, & Cordes (1997) | 4 PWS under FAF & not | 1 clinician<br>1 rel clinician | Not reported | 30-sec | Yes | ES= 62.5%<br>FG= 80%<br>AG= 97.5%<br>EO= 87.5% |
| Ingham & Riley (1998) | 2 CWS | 2 trained raters | Not reported | 15-sec | Yes | #1=73%<br>#2=80% |
| Ingham, Warner, Byrd, & Cotton (2006) | 12 PWS<br>12 PNS | 1 res ass | 42 | 1-min | Yes | 87.2% |
| Kalinowski, Noble, Armson, & Stewart (1994) | 10 PWS pre/post tx | 64 undergrads ½ rated ½ samples | 20 total each group did 10 | 1-min | Yes | °Mi pre= 67%<br>Mi post= 59%<br>Sev pre =60% |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | Sev post =32% |
| Mackey, Finn, & Ingham (1997) | 10 PWS (no dialect)<br>10 PNS (no dialect)<br>10 PNS (dialect) | 30 undergrads | 30 (10 from each group) | 1-min | Yes | °59% = PWS<br>56%=<br>PNS-D<br>80% = PNS |
| Martin & Haroldson (1992) | 10 PWS<br>10 PNS | 24 undergrads no exp w/ stuttering | 10 PWS<br>10 PNS | 1-min | Yes (doesn't say, so assumed) | °PWSA 40% +/- 0; 80% +/- 1<br>PWSAV 41% +/- 0; 79% +/-1<br>PNSA<br>40% +/- 0; 80% +/- 1<br>PNSAV 40% +/- 0; 81% +/- 1 |
| Martin, Haroldson, & Triden (1984) | 10 PWS<br>10 DAF PWS<br>10 PNS | 30 undergrads | 30 (10 from each group) | 1-min | Yes | +/-0 PWS 32%, DAF 32%, PNS 31%<br>+/- 1<br>PWS 74%, DAF 77%, PNS 75% |
| O'Brian, Packman, Onslow, Cream, O'Brian, & Bastock (2003) | 10 PWS pre/post PS<br>10 PNS | 15 untrained | 40 | 30-sec | Yes | 68% +/- 1 |
| Onslow, Adams, & Ingham (1992) | 10 PWS | 30 soph<br>30 unsoph | 40 15-sec<br>20 30-sec<br>10 60-sec | 15, 30, 60 sec | No | °S15= 54.2<br>S30= 50.1<br>S60= 61.8<br>U15= 62.5<br>U30= 48.7<br>U60= 59.2 |
| Onslow, Costa, Andrews, Harrison, & Packman (1996) | 12 PWS | 1 S & 1 S for reliability | *30 per rating point | 60 sec. | Yes | 62.9% +/-1<br>97.1 %+/-2<br>100% +/-3 |
| Packman, Onslow, van Doorn (1994) | 24 PWS | 10 clinicians | 24 | 30-sec | No | 79% the same or w/i +/- 1 scale value |

*Authors don't report. This is based on description of number of tapes and ratings per tape, but I'm honestly not sure.

° percent of ratings within +/- 1 scale value of each other

A=Audio
AV=Audiovisual
BR=Base Rate
CR= Chorus Reading
DAF=Delayed Auditory Feedback
FAF=Frequency Altered Feedback
CNS=Children Who Never Stuttered
CWUS=Children Who Used to Stutter
CW=Children With
Mi=Mild Stuttering Severity
PNS=People Who Never Stuttered
PS=Prolonged Speech
PWS=People Who Stutter
PWUS=People Who Used To Stutter
S=Sophisticated
Sev=Severe Stuttering Severity
SLP=Speech-Language Pathologist
U=Unsophisticated

APPENDIX E

INTRARATER AGREEMENT IN STUDIES MEASURING SPEECH NATURALNESS USING MARTIN, HAROLDSON, & TRIDEN'S (1984) 9-POINT SCALE

| Study | Speakers (N) | Listeners (N) | # of Speech Samples | Speech Sample Length | Speech samples include stutters | % of Samples Re-Rated | Time Between Ratings | Raters Intrajudge Agreement |
|---|---|---|---|---|---|---|---|---|
| Coughlin-Woods, Lehman, & Cooke (2005) | 60 CNS & 10 CW various comm. dis. | 39 naïve only 26 used in data analysis | 70 | 30-sec | n/a except for 2 children | 20 | Same session | 32/39 had at least 79% of their ratings w/i +/-1 |
| Finn (1997) | 15 PWUS<br>15 PNS | 10 trained grad students | 30 | 1-min | Yes (if there was any) | 100 | 1 week | 93.4% w/i +/- 1 |
| Finn & Ingham (1994) | 12 PWS | Same 12 PWS | Online | 30-sec | Yes | 100 | Same session<br><br>At least 1 week | Online 27% +/- 0; 62% +/- 1<br>offline 41% +/-0;<br>76% +/- 1 |
| Finn, Ingham, Ambrose, & Yari (1997) | 10 CWUS<br>10 CNS | 26 unsoph<br>14 exp SLPs | 20 | 1-min | Yes | 100 | 1 week | °SLPs= 65%<br>Unsoph= 64% |
| Kalinowski, Noble, Armson, & Stewart (1994) | 10 PWS pre/post tx | 64 undergrads ½ rated ½ samples | 20 total each group did 10 | 1-min | Yes | 100 | 1 week | °Mi pre =89%<br>Mi post =86%<br>Sev pre =83%<br>Sev post =89% |
| Mackey, Finn, & Ingham (1997) | 10 PWS (no dialect)<br>10 PNS (no | 30 undergrads | 30 (10 from each | 1-min | Yes | 100 | 1-3 weeks | °76% = PWS<br>80% = PNS-D<br>98% = PNS |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | dialect)<br>10 PNS<br>(dialect) | | group) | | | | | |
| Martin & Haroldson (1992) | 10 PWS<br>10 PNS | 24 undergrads no exp w/ stuttering | 10 PWS<br>10 PNS | 1-min | Yes (doesn't say, so assumed) | 25 | at least 2 weeks later | °PWSA 43/92<br>PWSAV 55/89<br>PNSA 66/97<br>PNSAV 57/94 |
| Martin, Haroldson, & Triden (1984) | 10 PWS<br>10 DAF PWS<br>10 PNS | 30 undergrads | 30 (10 from each group) | 1-min | Yes | 100 | 1-3 weeks | +/-0 PWS 45%, DAF 44%, PNS 54%<br>+/- 1 PWS 90%, DAF 85%, PNS 89% |
| O'Brian, Onslow, Cream, & Packman (2003) | 16 PWS tx<br>18 PNS | 3 U | 36 | 15-sec | No | 16.6<br>5.5 | Same session | 100% w/i +/- 1 |
| Onslow, Adams, & Ingham (1992) | 10 PWS | 30 S<br>30 U | 40 15-sec 20 30-sec 10 60-sec | 15, 30, 60 sec | No | 100 | 3 weeks | °S15= 65.6<br>S30= 75.2<br>S60= 72.3<br>U15= 68.7<br>U30= 69.1<br>U60= 72.4 |
| Onslow, Costa, Andrews, Harrison, & Packman (1996) | 12 PWS | 1 S & 1 S for reliability | *30 per rating point | 60 sec. | Yes | Not stated | 1 week | 94.3% were +/- 1<br>100% were +/- 2 |
| Onslow, Hayes, Hutchines, & Newman (1992) Study 1 | 7 PWS<br>7 PNS | 29 undergrads | 84 | 30-sec | No | 14.2 | Same session | 22/29 rated at least 9/11 sample w/i +/-1 |
| Onslow, Hayes, Hutchines, & Newman (1992) Study 2 | 36 PWS post-tx | 15 listeners no other details listed | 36 | 30-sec | No | 22.2 | Same session | Only 10/15 raters had 6/8 samples w/i+/-1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stuart & Kalinowski (2004) | 20 PWS DAF, FAF, & no DAF<br>5 PNS | 35 naïve | 55 | 15-sec | Yes | 100 | Same session | Spearman Rank Corr b/t 2 ratings .78<br>+/- 0 37.3% +/- 1 70% |

*Authors don't report. This is based on description of number of tapes and ratings per tape, but I'm honestly not sure.
° percent of ratings within +/- 1 scale value of each other

A=Audio
AV=Audiovisual
DAF=Delayed Auditory Feedback
FAF=Frequency Altered Feedback
CNS=Children Who Never Stuttered
CWUS=Children Who Used to Stutter
CW=Children With
D=Dialect
Exp=Experienced
Mi=Mild Stuttering Severity
PNS=People Who Never Stuttered
PWS=People Who Stutter
PWUS=People Who Used To Stutter
SLP=Speech-Language Pathologist
tx=Treatment
U=Unsophisticated
S=Sophisticated
Sev = Severe Stuttering Severity

# APPENDIX F

## STUDIES REPORTING RELIABILITY WHEN USING MARTIN ET AL.'S (1984) SCALE TO MEASURE SPEECH NATURALNESS

Armson & Kiefte, 2008
Coughlin-Woods et al., 2005
Finn, 1997
Finn & Ingham, 1994
Finn et al., 1997
Gow & Ingham, 1992
Hewat et al., 2006
Ingham et al., 1989
Ingham et al., 1997
Ingham et al., 2006
Ingham, Kilgo, Ingham, Moglia, Belnap, & Sanchez, 2001
Ingham, Martin et al., 1985
Ingham & Onslow, 1985
Ingham & Riley, 1998
Ingham, Sato et al., 2001
Kalinowski et al., 1994
Mackey et al., 1997
Martin & Haroldson, 1992
Martin et al., 1984
Metz et al., 1990
O'Brian, Onslow et al., 2003
O'Brian, Packman, Onslow, Cream et al., 2003
Onslow, Adams, & Ingham, 1992
Onslow et al., 1996
Onslow, Hayes et al., 1992
Packman et al., 1994
Runyan et al., 1990
Schiavetti et al., 1994
Stuart & Kalinowski, 2004
Stuart et al., 2004
Stuart et al., 2006
Tasko et al., 2007

APPENDIX G

STUDIES REPORTING RELIABILITY WITH INTRACLASS CORRELATION COEFFICIENTS (ICC)

| Study | Speakers (N) | Listeners (N) | # of Speech Samples | Speech Sample Length | Speech samples include stutters | Specify type of ICC calculated? | ICC Interjudge Agreement |
|---|---|---|---|---|---|---|---|
| Armson & Kiefte (2008) | 14 PWS w/ & w/o SE | 30 naïve 15 rated reading & 15 monologue | 56 | 1-min | Yes | No | Reading=.982<br>Reading SE=.959<br>Monologue=.954<br>Monologue SE=.941 |
| Hewat, Onslow, Packman, & O'Brian (2006) | 18 PWS in tx<br>11 PNS<br>12 PWS in prior tx | 10 unspoh | 41 | 15-sec | No | Yes (2,1)∇ | .84 |
| Kalinowski, Noble, Armson, & Stewart (1994) | 10 PWS pre/post tx | 64 undergrads ½ rated ½ samples | 20 total each group did 10 | 1-min | Yes | No | Mild pre=.93<br>Mild post=.99<br>Severe pre=.97<br>Severe post=.89 |
| Martin & Haroldson (1992) | 10 PWS<br>10 PNS | 24 undergrads no exp w/ stuttering | 10 PWS<br>10 PNS | 1-min | Yes (doesn't say, so assumed) | No | Group/Individual<br>PWSA=.97/.78<br>PWSAV=.96/.81<br>PNSA=.77/.24<br>PNSAV=.74/.16 |
| Martin, Haroldson, & Triden (1984) | 10 PWS<br>10 DAF PWS<br>10 PNS | 30 undergrads | 30 (10 from each group) | 1-min | Yes | No | Group/Individual<br>PWS=.98/.74<br>DAF=.98/.57<br>PNS=.75/.10 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Metz, Schiavetti, & Sacco (1990) | 20 PWS<br>post-tx<br>20 PNS | 30 SLP undergrads 15 used interval & 15 DME; interval only presented. | 40 PWS<br>40 PNS | 30-sec | No | No | Group/Individual<br>Reading=.984/.793<br>Speech=.982/.695 |
| O'Brian, Packman, Onslow, Cream, O'Brian, & Bastock (2003) | 10 PWS<br>pre/post PS<br>10 PNS | 15 untrained | 40 | 30-sec | Yes | Yes (2,1)$^{\nabla}$ | .71 |
| Onslow, Adams, & Ingham (1992) | 10 PWS | 30 soph<br>30 unsoph | 40 15-sec 20 30-sec 10 60-sec | 15, 30, 60 sec | No | Yes (2,1)$^{\nabla}$ | S15=.5 U15=.5<br>S30=.43 U30=.34<br>S60=.64 U60=.51 |
| Packman, Onslow, van Doorn (1994) | 24 PWS | 10 clinicians | 24 | 30-sec | No | Yes (2,1)$^{\nabla}$ | .91 |
| Schiavetti, Martin, Haroldson, & Metz (1994) | 10 PWS<br>10 PNS | 40 undergrads | 10 PWS<br>10 PNS | 1-min | Yes doesn't say but assume | No | Group=.997<br>Individual=.935 |
| Stuart & Kalinowski (2004) | 20 PWS<br>DAF, FAF, & no DAF<br>5 PNS | 35 naïve | 55 | 15-sec | Yes | Yes (2,1)$^{\nabla}$ | .71 |
| Stuart, Kalinowski, Rastatter, Saltuklaroglu, & Dayalu | 8 PWS w/ & w/o & 4 mo post SE | 15 undergrads | 48 | 15-sec | No | Yes (2,1)$^{\nabla}$ | .73 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (2004) | | | | | | | |
| Stuart, Kalinowski, Saltuklaroglu, & Guntupalli (2006) | 9 PWS w/ & w/o SE<br>5 PNS | 27 undergrads | 75; 70 PWS & 5 PNS | 15-sec | Yes | Yes (2,1)[V] | .70 |
| Tasko, McClean, & Runyan (2007) | 35 PWS pre/post | 3 groups of grad students judge 1/3 samples | 70 1 pre & 1 post | 1-min | Yes | Yes (2, k)[V] | .98, .97, .95 for 3 listener groups |

*Authors don't report. This is based on description of number of tapes and ratings per tape, but I'm honestly not sure.

° percent of ratings within +/- 1 scale value of each other

[V] Refers to six forms of ICC provided by Shrout & Fleiss (1979). The first parameter in the brackets refers to the model (one-way or two-way ANOVA) and the second parameter refers to the unit of generalization either a single judge or a group of k judges (Onslow, Adams, & Ingham, 1992). An ICC (2,1) form would be a 2-way ANOVA generalizing to a single judge.

A=Audio
AV=Audiovisual
BR=Base Rate
CR= Chorus Reading
DAF=Delayed Auditory Feedback
DME=Direct Magnitude Estimation
FAF=Frequency Altered Feedback
ICC=Intraclass Correlation Coefficient
PNS=People Who Never Stuttered
PS=Prolonged Speech
PWS=People Who Stutter
S=Sophisticated
SE=SpeechEasy
SLP=Speech-Language Pathologist
U=Unsophisticated

Participant #:_____ Session #: _____ DVD #: _____

APPENDIX H

DATA COLLECTION FORM FOR STUDY ONE*

| Sample # | Naturalness Rating | Sample # | Naturalness Rating |
|---|---|---|---|
| 1. | _____ | 16. | _____ |
| 2. | _____ | 17. | _____ |
| 3. | _____ | 18. | _____ |
| 4. | _____ | 19. | _____ |
| 5. | _____ | 20. | _____ |
| 6. | _____ | 21. | _____ |
| 7. | _____ | 22. | _____ |
| 8. | _____ | 23. | _____ |
| 9. | _____ | 24. | _____ |
| 10. | _____ | 25. | _____ |
| 11. | _____ | 26. | _____ |
| 12. | _____ | 27. | _____ |
| 13. | _____ | 28. | _____ |
| 14. | _____ | 29. | _____ |
| 15. | _____ | 30. | _____ |

Participant #:_____ Session #: _____ DVD #: _____

| Sample # | Naturalness Rating | Sample # | Naturalness Rating |
|---|---|---|---|
| 31. | _____ | 47. | _____ |
| 32. | _____ | 48. | _____ |
| 33. | _____ | | |
| 34. | _____ | | |
| 35. | _____ | | |
| 36. | _____ | | |
| 37. | _____ | | |
| 38. | _____ | | |
| 39. | _____ | | |
| 40. | _____ | | |
| 41. | _____ | | |
| 42. | _____ | | |
| 43. | _____ | | |
| 44. | _____ | | |
| 45. | _____ | | |
| 46. | _____ | | |

* Title not included on copies given to participants.

**24 samples rated twice during the same session

Participant #:_____

APPENDIX I
STUDY ONE INTAKE QUESTIONNAIRE*

1. Name:
2. Phone:
3. Age:
4. Gender:
5. Native English Speaker: (If no, what is your native language)
6. E-mail:
7. Fluent in other languages? Which?
8. Occupation:
9. Have you ever been hospitalized with a head injury, stroke, psychiatric related illness or brain tumor? Have you ever had brain surgery?
10. Have you ever been diagnosed with a speech or language disability, learning disability, dyslexia, or central auditory processing disorder? If so, what? When? If yes to any of the above, are you currently receiving any special help for your disability or disorder? Please describe.
11. What is your highest level of education?
12. Have you had any formal clinical or academic training in stuttering? What?
13. Do you personally know anyone who stutters?
14. Have you ever stuttered yourself?
15. Have you ever had speech therapy? Is so, for what reason?
16. Do you have any hearing loss that you know of? If so, what type of loss (congenital or acquired)? Hearing aids? Do you have any vision loss? Is it corrected with glasses or contacts?

* Title not included on copies given to participants.

APPENDIX J

PARTICIPANT'S PROMISE NOT TO DISCUSS STUDY*

As a participant in this study it is very important that you make your judgments independently, without trying to see how other judges are making their decisions. If you agree not to discuss any part of this study with anyone until after the study is COMPLETELY over, please initial below. It is VERY important to the study's results that you not discuss any aspect of your participation in this project, ESPECIALLY with other people also participating in the study (not even "did you see the cute little boy in the Spiderman shirt?").

I promise not to discuss any aspect of this study including the speech samples I am viewing with anyone.

| _____ | _____ | _____ |
|---|---|---|
| Participant | Participant # | Researcher |

* Title not included on copies given to participants.

APPENDIX K

STUDY ONE INSTRUCTIONS FOR TASK ONE*

Thank you for agreeing to participate in my study. Please read all of the instructions carefully before you begin the task.

You will be asked to rate the speech naturalness of 48 clips of children who stutter. The clips are on the enclosed CD-ROM. When you are ready to begin, insert the CD into a computer and click on the only file on the CD to begin. It will play on QuickTime or iTunes and may play on Windows Media Player. I couldn't get mine to work, but I have a very old PC and probably an outdated player so it may have been a problem with the computer and not the CD. *If you have any difficulty opening the file please contact me and I'll get you another one ASAP.* **DO NOT** use one of your classmates' CDs as each of you has a unique CD…your peer's CD *isn't* the same as yours.

Four sheets are enclosed behind this letter. The first is the instructions to rate speech naturalness. Please read them thoroughly before you begin the task. The second sheet is the 9-point scale you will use to rate the speech naturalness of children who stutter. Is it provided as a way to remind you what the numbers stand for as you rate speech naturalness. ***Please have it so you can see it as you rate the speech samples.***

The last two sheets are your data collection sheets. As you watch the videos, please write your naturalness rating on the blank corresponding to the video sample you just watched.

When you are ready to begin, insert the CD and start the video. Please watch it all the way through without stopping and rate all of the samples in one sitting. It will take you 32 minutes to complete the entire task. **Do not rewatch any of the samples.** Rate them based on your first viewing of the video. **Please use headphones as you watch the video and turn the sound up to 100% volume because the sound quality is not great and will be difficult to hear without headphones and without the volume turned all the way up.**

*One final reminder, PLEASE do not discuss your task with any of your classmates. Everyone's task is not the same therefore please do not discuss it with anyone related to the study. Once you finish your second task you'll be cleared to ask me any questions you have and talk about the entire study….but not until you are completely finished with the study.*

When you complete the task please place all of the materials back into the envelope and return it to my mailbox in the CMSD office on the 5th floor of Aderhold Hall. It's marked "Robin Edge" and is on the right hand side of the mailboxes. **Please complete the task and return the materials by Thursday March 25th.** Your next step will be ready to be picked up at that time as well.

Please e-mail me with any questions you may have robinbramlett@bellsouth.net.

Thanks again, Robin

* Title not included on copies given to participants.

Instructions to Rate Speech Naturalness

We are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you feel the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate. The next page contains a visual of the scale. Please have it available to refer to throughout the task.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

Please play the samples without pause or repetition.

APPENDIX L
STUDY ONE INSTRUCTIONS FOR TASK TWO*
EXPERIMENTAL GROUP SECOND TASK INSTRUCTIONS*

During this session you will be trained to measure speech naturalness, complete a Criterion Test to assess the effectiveness of your training, and rate speech samples of children who stutter. Pleas read the following directions carefully before you begin.

1. Read the handout "Training Instructions" completely.
2. Open the file on the CD entitled "SMS Training"
3. As you watch the speech samples in this video rate their speech naturalness following the "Training Instructions" directions. Record your ratings on the "Part Four: Rating Naturalness Data Recording Sheets."
4. When you finish sample 55-3, read the "Criterion Test" handout completely.
5. Open the file on the CD entitled "SMS CT."
6. As you watch the speech samples in this video rate their speech naturalness following the "Criterion Test" directions. Record your ratings on the "Part Five: Criterion Test Data Recording Sheets."
7. Read the instructions titled "Posttraining Instructions to Rate Speech Naturalness."
8. Open the file on the CD with "dissi" in the title.
9. Record your ratings on the last data collection form.

* Title not included on copies given to participants.

EXPERIMENTAL GROUP TRAINING INSTRUCTIONS & DATA COLLECTION FORMS*

In this session you will be trained to rate speech naturalness using a training program developed by Janis and Roger Ingham and colleagues (Ingham, Bakker, Moglia, & Kilgo, 1999). First you will be given some information about speech naturalness and then you will watch speech samples of people who stutter and people who do not stutter in order to rate their naturalness.

Although speech rate and stuttering frequency are important characteristics of speech production, it has become clear that they do not provide a complete picture of a person's speech – and especially the posttreatment speech of people who stutter. For example, speech following treatment may be stutter free and within normal speech rate ranges, yet still sound stilted or artificial or cautious- i.e. unnatural (Runyan & Adams, 1978; 1979). Or, a child learning to change stuttered speech to fluent speech may attempt to rely on sing-song speech or unusual animation to produce fluent utterances. These aspects of speech naturalness are not necessarily captured by speech rate or stuttering frequency data.

The concept of speech naturalness encompasses many aspects of speech beyond rate and fluency, including prosody (or the melody of speech determined mostly by modifications of pitch, quality, strength, and duration; Nicolosi, Harryman, & Kreschеck, 2004), vocal quality, placement and length of pauses, expressiveness, articulation (or speech sound production), syntax (or the way in which words are put together in a sentence to convey meaning; Nicolosi et al.), and semantics (or the meaning of language; Nicolosi et al.), discourse conventions (or the rules for the expression or exchange of ideas; Nicolosi et al.), and other less well-defined features. Because of this multiplicity of characteristics, only some of which may be relevant to a particular person's speech, we have selected as our measure of naturalness a perceptual rating

scale developed by Martin, Haroldson, & Triden (1984). It is simple to use and has been shown to be highly reliable (cf, Ingham, 1985). This is the same method you used in the previous rating task of this study. Using this system, the judge merely listens to a sample of speech and then rates its naturalness according to a 1-9-point scale where "1" means the speech sounded "highly natural" and "9" indicates the speech sounded "highly unnatural". Ratings between "1" and "9" indicate perceived degrees of naturalness in between those extremes. Martin et al. (1984) found that speech naturalness ratings for normal speaking adult talkers averaged 2.3. A rating of "3" is typically considered the uppermost boundary of "normal naturalness" (Ingham, 1985; Ingham, Gow, & Costello, 1985; Ingham & Onslow, 1985; Schiavetti & Metz, 1997).

This training program will present speech samples of nonstutterers and stutterers. The samples are one minute in length and present speech that illustrates various components and degrees of naturalness. Following is a list of directions explaining how to get started rating naturalness. Read through the list before performing any of the operations. Then return to #1 and carry out the directions.

1. You will use the file entitled "SMS" on the enclosed CD to access each sample. Prepare to listen to Sample 40.

2. Before rating, play a brief bit of the sample so that you are familiar with the speaker's style of speech and have a general idea of its naturalness.

3. Now you are ready to rate the naturalness of the samples that follow in Steps 1-3. All samples are one minute in duration. Step 1 will present three samples of nonstuttering speakers; Step 2 contains 3 samples of stuttering speakers; Step 3 contains samples with varying degrees of naturalness. Your job is to listen to each sample and concentrate on its naturalness, having in mind a rating between 1 and 9.

4. When you have completed the naturalness rating(s) for a given sample, turn to the data summary page in front of you and record your rating in box "1" beside the correct sample number. Then compare your ratings to the ones shown there.

**5.** If your rating *is not within* the acceptable range, listen to the sample a second time, keeping in mind the naturalness rating assigned to that sample on the data sheet. This should serve as an exemplar of that particular rating scale value. Rerate the sample and repeat this process until your rating is within the Target Range.

6. When your rating *is within* the Target Range, continue to the next sample. Stop this process when you have completed Sample 55.

***** Title not included on copies given to participants.

Participant #:_____

PART FOUR: RATING NATURALNESS

DATA RECORDING SHEETS

Steps 1-5, Samples 40-55

Step 1: Nonstuttering Speakers

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 40 | | | | | | 2-4 |
| Sample 41 | | | | | | 3-5 |
| Sample 42 | | | | | | 3-5 |

Step 2: Stuttering Speakers

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 43 | | | | | | 5-7 |
| Sample 44 | | | | | | 4-6 |
| Sample 45 | | | | | | 6-8 |

Step 3: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 46 | | | | | | 5-7 |
| Sample 47 | | | | | | 3-5 |
| Sample 48 | | | | | | 6-8 |

Participant #:_____

Step 4: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br><br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 49 | | | | | | 4-6 |
| Sample 50 | | | | | | 6-8 |
| Sample 51 | | | | | | 6-8 |

Step 5: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br><br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 52-1 | | | | | | 4-6 |
| Sample 52-2 | | | | | | 4-6 |
| Sample 53-1 | | | | | | 7-9 |
| Sample 53-2 | | | | | | 7-9 |
| Sample 54-1 | | | | | | 4-7 |
| Sample 54-2 | | | | | | 4-7 |
| Sample 55-1 | | | | | | 3-5 |
| Sample 55-2 | | | | | | 3-5 |
| Sample 55-3 | | | | | | 3-5 |

EXPERIMENTAL GROUP CRITERION TEST INSTRUCTIONS & DATA COLLECTION FORMS*

THE CRITERION TEST

Now you are ready for The Criterion Test, which is included for the purpose of evaluating whether you have acquired the ability to measure the speech naturalness produced by a variety of speakers who stutter. This part of the training program contains 24 1-minute samples not presented before. The task is exactly the same as the one you just completed, i.e., that you rate naturalness at one-minute intervals for each speaker while watching and listening to the speech samples. For the purpose of the Criterion Test, however, you will have only one opportunity to view and record the speech naturalness of each speaker.

What follows is a list of instructions explaining how to take the Criterion Test. Review them first, then return to #1 and begin.

1. Listen to a bit of each speech sample before you begin recoding data so that you are familiar with each speaker's speech pattern. Return to Sample 56, and begin.

2. When you have reached the end of the first sample turn to the Criterion Test data sheet (at the end of these instructions). Record your data in the spaces beside each sample number.

* Title not included on copies given to participants.

Participant #:_____

PART FIVE: CRITERION TEST

DATA RECORDING SHEETS

Samples 56-63

| SAMPLE | YOUR DATA |
|---|---|
| 56-1 | |
| 56-2 | |
| 56-3 | |
| | |
| 57-1 | |
| 57-2 | |
| 57-3 | |
| | |
| 58-1 | |
| 58-2 | |
| 58-3 | |
| | |
| 59-1 | |
| 59-2 | |
| 59-3 | |
| | |
| 60-1 | |
| 60-2 | |
| 60-3 | |

Participant #:_____

| | |
|---|---|
| 61-1 | |
| 61-2 | |
| 61-3 | |
| | |
| 62-1 | |
| 62-2 | |
| 62-3 | |
| | |
| 63-1 | |
| 63-2 | |
| 63-3 | |

When you finish number 63-3, read the instructions on the next page for your last task. You may take a break between this task and the next one if you like.

POSTTRAINING INSTRUCTIONS TO RATE SPEECH NATURALNESS*

As in your previous sessions, we are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you feel the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate. The next page contains a visual of the scale. Please have it available to refer to throughout the task.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

You may recognize some of the video samples from your previous sessions. During today's session you should try to make your judgments in a manner that is consistent with the way that you were judging speech naturalness at the END of your training session. Try to make your judgments today in a manner that you believe would be "correct" according to what you learned in your training session.

Please play the samples without pause or repetition.

* Title not included on copies given to participants.

EXPOSURE CONTROL GROUP'S SECOND TASK INSTRUCTIONS*

During this session you will rate speech samples of adults and children who stutter. Please read the following directions carefully before you begin.

1. Read the handout "Instructions to Rate Speech Naturalness of SMS samples" completely.
2. Open the file on the CD entitled "SMS"
3. As you watch the speech samples in this video rate their speech naturalness following the "Instructions to Rate Speech Naturalness of SMS samples" directions. Record your ratings on the "Part Four: Rating Naturalness Data Recording Sheets."
4. After you finish sample 63-3, read the "Instructions to Rate Speech Naturalness" completely.
5. Open the file on the CD with "dissi" in the title.
6. Record your ratings on the last data collection form.

* Title not included on copies given to participants.

EXPOSURE CONTROL GROUP'S INSTRUCTIONS TO RATE SPEECH NATURALNESS OF *SMS* SAMPLES*

Instructions to Rate Speech Naturalness of *SMS* Samples

1. In this session you will rate the speech naturalness of people who stutter and people who do not stutter. The samples are one minute in length. You will use the file entitled "SMS" on the enclosed CD to access each sample. Prepare to listen to sample 40.

2. Now you are ready to rate the naturalness of the samples that follow. Listen to each sample and concentrate on its naturalness, having in mind a rating between 1 and 9. At the end of each sample when you have determined the naturalness rating for a given sample, turn to the data summary page attached and record your rating in the box beside the correct sample number.

3. When you finish all of the samples, read the enclosed instructions for the next task.

* Title not included on copies given to participants.

Participant #:_____

PART FOUR: RATING NATURALNESS

DATA RECORDING SHEETS

| SAMPLE | YOUR DATA |
|---|---|
| Sample 40 | |
| Sample 41 | |
| Sample 42 | |
| Sample 43 | |
| Sample 44 | |
| Sample 45 | |

| SAMPLE | YOUR DATA |
|---|---|
| Sample 46 | |
| Sample 47 | |
| Sample 48 | |
| Sample 49 | |
| Sample 50 | |
| Sample 51 | |

| SAMPLE | YOUR DATA |
|---|---|
| Sample 52-1 | |
| Sample 52-2 | |
| Sample 53-1 | |
| Sample 53-2 | |
| Sample 54-1 | |
| Sample 54-2 | |
| Sample 55-1 | |
| Sample 55-2 | |
| Sample 56-1 | |
| Sample 56-2 | |
| Sample 56-3 | |
| Sample 57-1 | |
| Sample 57-2 | |

Participant #:_____

| SAMPLE | YOUR DATA |
|---|---|
| Sample 57-3 | |
| Sample 58-1 | |
| Sample 58-2 | |
| Sample 58-3 | |
| Sample 59-1 | |
| Sample 59-2 | |
| Sample 59-3 | |
| Sample 60-1 | |
| Sample 60-2 | |
| Sample 60-3 | |
| Sample 61-1 | |
| Sample 61-2 | |
| Sample 61-3 | |
| Sample 62-1 | |
| Sample 62-2 | |
| Sample 62-3 | |
| Sample 63-1 | |
| Sample 63-2 | |

CONTROL GROUP'S INSTRUCTIONS TO RATE SPEECH NATURALNESS ON OCCASION2*

As in your previous session, we are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you feel the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate. The next page contains a visual of the scale. Please have it available to refer to throughout the task.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

You may recognize some of the video samples from your previous session. Please do not attempt to remember what your rating was in the prior session; instead, rate how natural you feel the sample is today.

Please play the samples without pause or repetition (based on Martin et al., 1984).

* Title not included on copies given to participants.

APPENDIX M
STUDY ONE POST STUDY QUESTIONNAIRE*

Please answer the following questions providing as much detail as possible regarding your experience participating in the naturalness study. Please be honest. Nothing you say positive or negative will affect your extra credit and any feedback I receive will help me with the next step in my dissertation process.

1. Were you able to finish both steps of this study? If no, please explain.

2. Did you complete both of the steps independently?

3. What criteria did you use to determine what naturalness was?

4. Did you compare the children's speech to other children's speech or to adults' speech?

5. As best you can remember, how many days passed between your completion of Step 1 and Step 2?

6. If your videos included adults, did you have any trouble rating the speech naturalness of adults and then going back to rate the speech naturalness of children? If so, please explain.

7. On a scale of 1 to 9 with "1" meaning "not at all" and "9" meaning "a lot," please write the number that best represents how much you have discussed any part of this study with someone other than the researcher.

8. Did you have any difficulty completing either of the steps? Is so, what? Please be as detailed as possible.

9. If you were conducting the same study, would you have done anything differently?

10. Any other comments? Participant #: _____

* Title not included on copies given to participants.

APPENDIX N

STUDY ONE: MEAN SPEECH NATURALNESS SCORES BY GROUP FOR EACH SPEECH SAMPLE FOR EACH RATING OCCASION

Training Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| 4.1 | 3.00 | 3.20 | 4.60 | 4.67 |
| 5.2 | 5.53 | 6.27 | 6.80 | 6.67 |
| 6.1 | 5.86 | 6.53 | 6.87 | 6.53 |
| 6.4 | 6.13 | 6.80 | 7.13 | 7.47 |
| 6.5 | 7.20 | 6.90 | 7.07 | 7.60 |
| 7.2 | 7.00 | 7.40 | 7.27 | 7.00 |
| 7.3 | 7.33 | 7.60 | 7.73 | 7.53 |
| 8.1 | 7.73 | 8.20 | 8.20 | 8.20 |
| 8.2 | 8.20 | 8.47 | 8.20 | 8.13 |
| 8.3 | 6.40 | 7.13 | 7.53 | 7.73 |
| 8.4 | 7.10 | 8.27 | 7.67 | 7.87 |
| 8.5 | 7.30 | 7.87 | 7.67 | 7.67 |
| 8.6 | 8.40 | 8.53 | 8.07 | 8.20 |
| 9.2 | 3.93 | 4.80 | 5.87 | 6.07 |
| 9.3 | 4.73 | 5.07 | 5.80 | 5.93 |
| 11.3 | 3.13 | 2.80 | 5.13 | 5.40 |
| 11.4 | 1.87 | 2.33 | 4.27 | 4.87 |
| 11.5 | 2.40 | 2.33 | 4.53 | 4.80 |
| 13.1 | 6.87 | 8.00 | 8.00 | 7.87 |
| 13.2 | 7.73 | 7.53 | 7.80 | 7.93 |
| 14.1 | 3.13 | 3.93 | 4.67 | 5.20 |
| 14.4 | 5.33 | 5.40 | 6.27 | 5.87 |
| 14.5 | 3.80 | 3.73 | 5.07 | 5.13 |
| 14.6 | 3.87 | 4.40 | 5.53 | 5.33 |

Exposure Control Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| 4.1 | 3.50 | 3.64 | 3.64 | 3.43 |
| 5.2 | 5.50 | 5.93 | 5.36 | 5.71 |
| 6.1 | 4.21 | 5.64 | 5.43 | 5.00 |
| 6.4 | 6.07 | 6.86 | 6.36 | 6.86 |
| 6.5 | 7.21 | 7.00 | 7.00 | 6.93 |
| 7.2 | 6.21 | 6.43 | 5.71 | 5.86 |
| 7.3 | 6.71 | 7.14 | 6.86 | 6.86 |
| 8.1 | 7.64 | 8.07 | 7.86 | 7.86 |
| 8.2 | 7.86 | 8.29 | 8.00 | 8.29 |

| | | | | |
|---|---|---|---|---|
| 8.3 | 5.67 | 7.50 | 7.36 | 7.21 |
| 8.4 | 6.93 | 8.07 | 7.64 | 7.93 |
| 8.5 | 7.79 | 7.58 | 7.43 | 7.71 |
| 8.6 | 8.29 | 8.29 | 8.07 | 8.07 |
| 9.2 | 4.29 | 5.29 | 4.14 | 4.79 |
| 9.3 | 5.14 | 5.43 | 4.36 | 4.79 |
| 11.3 | 3.79 | 3.29 | 3.64 | 3.71 |
| 11.4 | 3.00 | 3.00 | 3.00 | 2.79 |
| 11.5 | 2.71 | 2.93 | 3.07 | 3.14 |
| 13.1 | 6.57 | 7.50 | 6.86 | 7.07 |
| 13.2 | 7.14 | 7.14 | 7.21 | 6.86 |
| 14.1 | 4.71 | 4.21 | 3.64 | 4.07 |
| 14.4 | 5.57 | 5.07 | 5.07 | 4.86 |
| 14.5 | 3.86 | 4.14 | 3.50 | 3.57 |
| 14.6 | 3.50 | 3.71 | 3.86 | 3.07 |

Control Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| 4.1 | 3.43 | 4.07 | 3.71 | 3.79 |
| 5.2 | 6.00 | 6.00 | 5.43 | 5.79 |
| 6.1 | 6.00 | 5.93 | 6.29 | 6.07 |
| 6.4 | 6.21 | 6.57 | 6.64 | 6.93 |
| 6.5 | 7.57 | 7.50 | 9.36 | 7.21 |
| 7.2 | 6.36 | 6.64 | 6.29 | 6.29 |
| 7.3 | 7.00 | 6.93 | 7.14 | 6.93 |
| 8.1 | 7.71 | 8.21 | 8.00 | 9.36 |
| 8.2 | 8.29 | 8.36 | 7.93 | 8.21 |
| 8.3 | 6.79 | 7.29 | 7.64 | 7.21 |
| 8.4 | 7.00 | 7.79 | 7.86 | 7.86 |
| 8.5 | 7.64 | 7.57 | 7.64 | 7.29 |
| 8.6 | 8.50 | 8.36 | 8.36 | 8.64 |
| 9.2 | 5.50 | 5.36 | 4.71 | 5.21 |
| 9.3 | 5.14 | 5.79 | 5.64 | 5.50 |
| 11.3 | 3.21 | 3.43 | 3.79 | 4.14 |
| 11.4 | 2.43 | 3.14 | 3.07 | 3.57 |
| 11.5 | 2.71 | 3.43 | 3.50 | 3.43 |
| 13.1 | 6.36 | 7.07 | 6.57 | 7.29 |
| 13.2 | 6.93 | 7.14 | 6.93 | 7.21 |
| 14.1 | 4.86 | 4.07 | 4.07 | 4.93 |
| 14.4 | 6.07 | 5.86 | 5.36 | 5.07 |
| 14.5 | 4.07 | 4.14 | 4.21 | 4.57 |
| 14.6 | 4.00 | 3.71 | 4.21 | 4.29 |

APPENDIX O

STUDY ONE: INTRARATER AGREEMENT OF SPEECH NATURALNESS - THE PERCENTAGE OF RERATINGS WITHIN +/- 1 OF ORIGINAL VALUE.

| Judge | Group | % occasion 1 versus occasion 2 | % occasion 3 versus occasion 4 |
|---|---|---|---|
| 1 | 1 | 70.83% | 91.67% |
| 2 | 2 | 70.83% | 83.33% |
| 3 | 1 | 79.17% | 100% |
| 4 | 3 | 79.17% | 62.5% |
| 5 | 1 | 83.33% | 100% |
| 6 | 3 | 95.83% | 100% |
| 7 | 1 | 50% | 83.33% |
| 8 | 1 | 66.67% | 100% |
| 9 | 2 | 79.17% | 87.5% |
| 10 | 1 | 87.5% | 91.67% |
| 11 | 1 | 79.17% | 75% |
| 12 | 1 | 83.33% | 95.83% |
| 13 | 2 | 58.33% | 79.17% |
| 14 | 3 | 66.67% | 91.67% |
| 15 | 3 | 83.33% | 66.67% |
| 16 | 2 | 75% | 75% |
| 17 | 3 | 75% | 79.17% |
| 18 | 1 | 70.83% | 95.83% |
| 19 | 3 | 79.17% | 79.17% |
| 20 | 3 | 87.5% | 70.83% |
| 21 | 3 | 58.33% | 79.17% |
| 22 | 2 | 54.17% | 58.33% |
| 23 | 2 | 75% | 87.5% |
| 24 | 1 | 91.67% | 91.67% |
| 25 | 2 | 75% | 79.17% |
| 26 | 3 | 87.5% | 83.33% |
| 27 | 2 | 66.67% | 75% |
| 28 | 1 | 50% | 95.83% |
| 29 | 3 | 87.5% | 91.67% |
| 30 | 2 | 83.33% | 75% |
| 31 | 2 | 58.33% | 79.19% |
| 32 | 2 | 87.5% | 79.17% |

| | | | |
|---|---|---|---|
| 33 | 3 | 75% | 62.5% |
| 34 | 1 | 70.83% | 100% |
| 35 | 2 | 87.5% | 87.5% |
| 36 | 3 | 75% | 91.67% |
| 37 | 1 | 79.17% | 83.33% |
| 38 | 2 | 95.83% | 91.67% |
| 39 | 3 | 66.67% | 91.67% |
| 40 | 1 | 70.83% | 91.67% |
| 41 | 2 | 70.83% | 87.5% |
| 42 | 3 | 66.67% | 91.67% |
| 43 | 1 | 83.33% | 95.83% |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group

APPENDIX P

STUDY ONE: INTERRATER AGREEMENT OF SPEECH NATURALNESS - THE PERCENTAGE OF RATINGS FOR EACH JUDGE WITHIN +/- 1 OF EVERY OTHER JUDGE IN HER GROUP ON EACH OCCASION.

| Judge | Group | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|---|
| 1 | 1 | 56.7% | 60.8% | 57.9% | 61.7% |
| 2 | 2 | 52.6% | 52.7% | 52.4% | 43.2% |
| 3 | 1 | 57.5% | 59.0% | 57.7% | 63.1% |
| 4 | 3 | 48.9% | 40.5% | 52.8% | 56.2% |
| 5 | 1 | 56.0% | 57.9% | 59.5% | 60.9% |
| 6 | 3 | 49.7% | 50.0% | 42.4% | 46.4% |
| 7 | 1 | 47.6% | 47.6% | 51.8% | 50.0% |
| 8 | 1 | 59.6% | 59.5% | 56.5% | 60.8% |
| 9 | 2 | 48.6% | 48.6% | 56.4% | 63.9% |
| 10 | 1 | 55.0% | 55.1% | 58.8% | 57.8% |
| 11 | 1 | 48.5% | 41.3% | 50.0% | 51.5% |
| 12 | 1 | 41.8% | 42.0% | 58.5% | 60.8% |
| 13 | 2 | 54.3% | 53.8% | 43.3% | 58.5% |
| 14 | 3 | 43.2% | 38.8% | 43.9% | 36.6% |
| 15 | 3 | 46.5% | 53.6% | 50.2% | 50.5% |
| 16 | 2 | 52.2% | 54.9% | 54.8% | 60.1% |
| 17 | 3 | 52.7% | 39.7% | 45.6% | 46.3% |
| 18 | 1 | 49.4% | 48.4% | 54.7% | 52.9% |
| 19 | 3 | 51.7% | 60.5% | 53.6% | 55.1% |
| 20 | 3 | 39.9% | 47.8% | 54.6% | 56.5% |
| 21 | 3 | 49.8% | 51.5% | 57.5% | 53.6% |
| 22 | 2 | 49.9% | 50.2% | 45.2% | 50.2% |
| 23 | 2 | 40.9% | 43.0% | 41.3% | 49.1% |
| 24 | 1 | 59.0% | 58.7% | 54.5% | 56.7% |
| 25 | 2 | 39.4% | 36.2% | 52.7% | 54.7% |
| 26 | 3 | 49.6% | 48.7% | 56.6% | 58.5% |
| 27 | 2 | 39.7% | 44.1% | 45.9% | 42.5% |
| 28 | 1 | 56.6% | 54.7% | 49.5% | 53.4% |
| 29 | 3 | 50.8% | 50.0% | 59.4% | 57.4% |
| 30 | 2 | 54.1% | 54.9% | 56.9% | 52.6% |
| 31 | 2 | 58.6% | 50.0% | 56.1% | 57.3% |
| 32 | 2 | 45.3% | 47.1% | 50.3% | 53.5% |
| 33 | 3 | 55.1% | 46.8% | 54.4% | 51.6% |
| 34 | 1 | 52.0% | 50.3% | 55.5% | 60.2% |
| 35 | 2 | 42.1% | 44.5% | 57.0% | 60.3% |
| 36 | 3 | 55.2% | 51.9% | 61.8% | 64.6% |
| 37 | 1 | 52.6% | 49.8% | 52.9% | 58.4% |
| 38 | 2 | 50.1% | 51.8% | 50.7% | 47.9% |
| 39 | 3 | 56.9% | 54.9% | 54.5% | 58.5% |

| | | | | | |
|---|---|---|---|---|---|
| 40 | 1 | 57.5% | 58.4% | 49.2% | 53.3% |
| 41 | 2 | 42.2% | 38.3% | 48.0% | 52.5% |
| 42 | 3 | 25.2% | 19.6% | 39.2% | 29.9% |
| 43 | 1 | 52.4% | 52.5% | 42.6% | 39.9% |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group

APPENDIX Q

STUDY TWO: PRESTUDY INTAKE QUESTIONNAIRE

1. Name: 2. Email:

3. Age: 4. Gender: 5. Phone:

6. Is English the first language you learned to speak? *If no, what is your native language?*

7. Are you fluent in other languages? *If yes, which?*

8. Do you have any hearing loss that you know of? *If so, what type of loss?*
   Do you have hearing aids?

9. Do you have any vision loss? *If yes, is it corrected with glasses or contacts?*

10. Have you ever been hospitalized with a head injury, stroke, psychiatric related illness or brain tumor? Have you ever had brain surgery? *If yes, list when.*

11. Have you ever been diagnosed with a speech or language disability, learning disability, dyslexia, or central auditory processing disorder? *If so, what? When? If yes to any of the above, are you currently receiving any special help for your disability or disorder? Please describe.*

12. Have you ever stuttered? *If yes, please give as many details as possible (i.e. when).*

13. Have you had any formal clinical or academic training in stuttering? *If yes, what?*

14. Does anyone in your immediate family or a close friend stutter?

    *If yes, who? How long have you known him or her?*

15. Have you ever had a teacher who stuttered? *If yes what year? How long?*

*16.* Have you ever had speech therapy? *If so, when and for what reason?*

*17.* Do you have any experience with 3-6 year old children? *If yes, what?*

18. What year are you at UGA?

19. How would you prefer to be reminded of your appointments? Phone or email?

*Qualify for study? Y / N* *Group Assignment:* _____
*If yes, date and time scheduled:* _____
*(For researcher use only)*

APPENDIX R
STUDY TWO CONSENT FORM


University of Georgia
Institutional Review Board
Approved: 8-31-11
Expires: 1-30-12

Consent for Participation in The Speech Naturalness of Children

I, _____, agree to participate in a research study titled "The Speech Naturalness of Children as Perceived by Trained and Untrained Listeners," which is being conducted by Robin Bramlett from the Department of Communication Disorders & Special Education (770-630-4852) under the direction of Dr. Anne Bothe in the Department of Communication Sciences and Special Education (706-542-0436). I understand that my participation is voluntary. I can refuse to participate or stop taking part at anytime without giving any reason, and without penalty or loss of benefits to which I am otherwise entitled. I can ask to have all information about me returned to me, removed from the research records, or destroyed.

The following have been explained to me and I understand them:

Item #1: REASON/PURPOSE. The purpose of this investigation is to assess the effect naturalness training has on the measurement of speech naturalness of children who stutter as rated by normal adult listeners.

Item # 2: BENEFITS. My participation in this research will provide information on the perceived speech naturalness of children that may help speech-language pathologists better evaluate them before and after they receive stuttering treatment. I will benefit from the knowledge that I have participated in research that is groundbreaking in the speech field and will further the field's knowledge base. I will benefit from the free vision and hearing screenings that will check the status of my hearing and vision. If any any vision or hearing problems are identified, I will be referred to and counseled to visit appropriate professionals for confirmation and treatment if needed.

Item # 3: PROCEDURES. The entire study will take place in the University of Georgia Speech and Hearing Clinic in Aderhold Hall. If I volunteer to take part in this study, I will be expected to come to Aderhold Hall four times over a 4-6 week period and will be asked to do the following things:

1. Complete a brief vision and hearing screening during the first visit.
2. Rate the naturalness of 36 30-second speech samples of children who stutter using a 9-point scale on four separate occassions. Each rating session will take approximately 40 minutes and will occur 7-10 days after the previous session.
3. I may be asked to participate in a 40-80 minute naturalness training session to coincide with one of the four rating sessions (i.e. I will complete the naturalness training and immediately complete one of the four naturalness rating sessions).

I understand that I will be unable to participate in this study if I do not pass the hearing and vision screenings administered today.

Item # 4: DISCOMFORTS OR STRESSES. There are no foreseeable discomforts or stresses for me.

Item # 5: RISKS. There are no foreseeable risks for my participation in this study.

Item # 6: CONFIDENTIALITY OF DATA. The only people who will know that I am a research participant are members of the research team. My confidentiality will be strictly maintained as my name will not be entered into any data system nor will my name or any information from me or about me be shared in any individually identifiable form. Only a subject number will identify all information about me. There is no need to be able to connect my personal information to the information gathered in this study, so there will be no key connecting my identifying information to the subject number. This informed consent form that includes my name will be kept in a separate locked office, away from all descriptive information and away from the results of this study. The information collected during the telephone interview will be transferred to an electronic database using only my assigned subject number. After it is entered into the computer, the original screening form, containing information that could identify me will be destroyed. All data collected during the telephone screening process will be immediately destroyed if I fail the vision or hearing screening and can not participate in the study.

Item # 7: FURTHER QUESTIONS. The researcher will answer any further questions about the research, now or during the course of the project. I can contact Robin Bramlett via phone at (770) 630-4852 or Email at robinbramlett@bellsouth.net or Dr. Bothe at (706) 542-0436 or abothe@uga.edu during business hours (9:00am-6:00pm) if I have any questions.

Item # 8: FINAL AGREEMENT & CONSENT FORM COPY. I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form to keep.

Signature of Researcher Date

Robin Bramlett, CCC-SLP
robinbramlett@bellsouth.net

Signature of Participant Date

Additional questions or problems regarding your rights as a research participant should be addressed to The Chairperson, Institutional Review Board, University of Georgia, 612 Boyd Graduate Studies Research Center, Athens, Georgia 30602-7411; Telephone (706) 542-3199; E-Mail Address IRB@uga.edu.

University of Georgia
Institutional Review Board
Approved: 8-31-11
Expires: 1-30-12

Participant #:_____

APPENDIX S

STUDY TWO: POST STUDY QUESTIONNAIRE

Please answer the following questions providing as much detail as possible regarding your experience participating in the naturalness study. Please be honest. Nothing you say positive or negative will affect your extra credit and any feedback I receive will help me with the next step in my dissertation process.

1. What criteria did you use to determine what naturalness was?

2. Did you compare the children's speech to other children's speech or to adults' speech?

3. If your videos included adults, did you have any trouble rating the speech naturalness of adults and then going back to rate the speech naturalness of children? If so, please explain.

4. Did you have any difficulty completing any of the steps? Is so, what? Please be as detailed as possible.

5. If you were conducting the same study, would you have done anything differently?

6. Any other comments?

## APPENDIX T

## FIRST PRETRAINING RATING SESSION INSTRUCTIONS TO PARTICIPANTS*

We are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you think the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

The samples will be played without pause and no sample will be repeated (based on Martin et al., 1984).

* Title not included on copies given to participants.

Participant #:_____ Session #:_____ DVD #:_____

APPENDIX U

SPEECH NATURALNESS DATA COLLECTION FORM*

| Sample # | Naturalness Rating | Sample # | Naturalness Rating |
|---|---|---|---|
| 1. | _____ | 21. | _____ |
| 2. | _____ | 22. | _____ |
| 3. | _____ | 23. | _____ |
| 4. | _____ | 24. | _____ |
| 5. | _____ | 25. | _____ |
| 6. | _____ | 26. | _____ |
| 7. | _____ | 27. | _____ |
| 8. | _____ | 28. | _____ |
| 9. | _____ | 29. | _____ |
| 10. | _____ | 30. | _____ |
| 11. | _____ | 31. | _____ |
| 12. | _____ | 32. | _____ |
| 13. | _____ | 33. | _____ |
| 14. | _____ | 34. | _____ |
| 15. | _____ | 35. | _____ |
| 16. | _____ | 36. | _____ |
| 17. | _____ | | |
| 18. | _____ | | |
| 19. | _____ | | |
| 20. | _____ | | |

*Title not included on copies given to participants

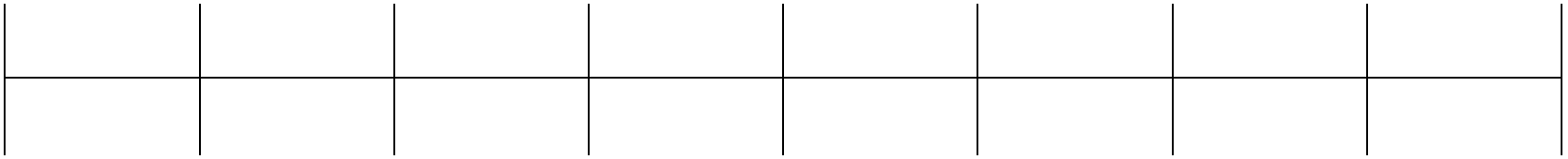
Participant #: _____ Session #:_____

APPENDIX V

DISCUSSION OF STUDY SCALE*

On a 1 to 9 scale with "1" meaning "Not At All" and "9" meaning "A Lot" please circle the number that best represents how much you have discussed any part of this study with someone other than the researcher between your last session and now.

Not At All A Lot



1 2 3 4 5 6 7 8 9

*Title not included on copies given to participants.

## APPENDIX W

## SECOND PRETRAINING RATING SESSION (FOR ALL GROUPS), THIRD & FOURTH PRETRAINING RATING SESSION (FOR CONTROL GROUPS ONLY) INSTRUCTIONS TO PARTICIPANTS*

As in your previous session, we are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you think the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you now. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

You may recognize some of the video samples from your previous sessions. Please do not attempt to remember what your rating was in prior sessions; instead, rate how natural you think the sample sounds now.

The samples will be played without pause and samples will not be replayed (based on Martin et al., 1984).

* Title not included on copies given to participants.

APPENDIX X

*SMS* TRAINING INSTRUCTIONS

*(FOR EXPERIMENTAL GROUP ONLY)

*To be read by experimenter with participant given a copy to read along.* (Bolded information are changes or additions from the original *SMS* workbook pp.74-76. This information was not bolded on the copy used by participants). Otherwise these are taken direct from the *SMS* training workbook (Ingham & Ingham, 2010).

**In this session you will be trained to rate speech naturalness using a training program developed by Janis and Roger Ingham and colleagues (Ingham, Bakker, Moglia, & Kilgo, 1999). First you will be given some information about speech naturalness and then you will watch speech samples of people who stutter and people who do not stutter in order to rate their naturalness.**

Although speech rate and stuttering frequency are important characteristics of speech production, it has become clear that they do not provide a complete picture of a person's speech – and especially the posttreatment speech of people who stutter. For example, speech following treatment may be stutter free and within normal speech rate ranges, yet still sound stilted or artificial or cautious- i.e. unnatural (Runyan & Adams, 1978; 1979). Or, a child learning to change stuttered speech to fluent speech may attempt to rely on sing-song speech or unusual animation to produce fluent utterances. These aspects of speech naturalness are not necessarily captured by speech rate or stuttering frequency data.

The concept of speech naturalness encompasses many aspects of speech beyond rate and fluency, including prosody **(which is the melody of speech determined mostly by modifications of vocal pitch, quality, strength, and duration; Nicolosi, Harryman, & Krescheck, 2004)**, vocal quality, placement and length of pauses, expressiveness, articulation **(which is speech sound production)**, syntax **(the way in which words are put together in a sentence to convey meaning; Nicolosi et al.)**, and semantics **(the meaning of language; Nicolosi et al.)**, discourse conventions **(the rules for the expression or exchange of ideas; Nicolosi et al.)**, and other less well-defined features. Because of this multiplicity of characteristics, only some of which may be relevant to a particular person's speech, we have selected as our measure of naturalness a perceptual rating scale developed by Martin, Haroldson,

& Triden (1984). It is simple to use and has been shown to be highly reliable (cf, Ingham, 1985). **This is the same method you have used in previous sessions of this study. Using this system**, the judge merely listens to a sample of speech and then rates its naturalness according to a 1-9-point scale where "1" means the speech sounded "highly natural" and "9" indicates the speech sounded "highly unnatural". Ratings between "1" and "9" indicate perceived degrees of naturalness in between those extremes. Martin et al. (1984) found that speech naturalness ratings for normal speaking adult talkers averaged 2.3. A rating of "3" is typically considered the uppermost boundary of "normal naturalness" (Ingham, 1985; Ingham, Gow, & Costello, 1985; Ingham & Onslow, 1985; Schiavetti & Metz, 1997).

This training program will present speech samples of nonstutterers and stutterers. The samples are one and two minutes in length and present speech that illustrates various components and degrees of naturalness. Following is a list of directions explaining how to get started rating naturalness. Read through the list before performing any of the operations. Then return to #1 and carry out the directions. *(#1 was deleted because it is about setting up the computer and the experimenter completed that step prior to the session starting.)*

1. **You will access each sample from the window currently open on the right side of your computer screen. You will record your naturalness ratings using the SMS program currently open on the left side of your computer screen**. Prepare to listen to Sample 40.

2. **Double click the sample you need to open in quick time.**

3. Before rating, play a brief bit of the sample so that you are familiar with the speaker's style of speech and have a general idea of its naturalness.

4. Now you are ready to rate the naturalness of the samples that follow in Steps 1-3. All samples are one minute in duration. Step 1 will present three samples of nonstuttering speakers; Step 2 contains 3 samples of stuttering speakers; Step 3 contains samples with varying degrees of naturalness.

5. **When you are ready to begin, start the sample. When the sample begins, immediately depress the computer's space bar to start the SMS timer.** Your job is to listen to each sample and concentrate on its naturalness, having in mind a rating between 1 and 9. When the naturalness tone sounds and the "NA" signal appears in the bottom left corner of the

screen, use the number row on the computer keyboard to indicate your rating. You must press a number key within 5 seconds of the signal. At the completion of the sample the computer will automatically display your naturalness rating on the data summary screen. For samples that require multiple naturalness ratings, the cumulative averaged rating will be displayed in the NAT box during data collection, and the final averaged naturalness score will appear on the data summary screen at the end of the sample. **Ignore any other information that appears on the blue screen.**

*6.* When you have completed the naturalness rating(s) for a given sample, turn to the data summary page in front of you **and record your rating in box "1" beside the correct sample number**. Then compare your ratings to the ones shown there. *(The explanation about how the standard naturalness ratings were obtained from the naïve undergrads was omitted from here).*

**7.** If your rating *is not within* the acceptable range, listen to the sample a second time, keeping in mind the naturalness rating assigned to that sample on the data sheet. This should serve as an exemplar of that particular rating scale value. **Rerate the sample and repeat this process until your rating is within the Target Range.**

8. When your rating *is within* the Target Range, continue to the next sample. Stop this process when you have completed **Sample 55**.

9. **When you are ready to begin, start the sample. When the sample begins, immediately depress the computer's space bar to start the SMS timer.**

Participant #: _____

PART FOUR: RATING NATURALNESS

DATA RECORDING SHEETS for experimental group

Steps 1-5, Samples 40-55

Step 1: Nonstuttering Speakers

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 40 | | | | | | 2-4 |
| Sample 41 | | | | | | 3-5 |
| Sample 42 | | | | | | 3-5 |

Step 2: Stuttering Speakers

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 43 | | | | | | 5-7 |
| Sample 44 | | | | | | 4-6 |
| Sample 45 | | | | | | 6-8 |

Step 3: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 46 | | | | | | 5-7 |
| Sample 47 | | | | | | 3-5 |
| Sample 48 | | | | | | 6-8 |

Participant #: _____

Step 4: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **One-Minute Samples**<br><br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 49 | | | | | | 4-6 |
| Sample 50 | | | | | | 6-8 |
| Sample 51 | | | | | | 6-8 |

Step 5: Varied Naturalness

| | 1 | 2 | 3 | 4 | 5 | **Two-Minute Samples**<br><br>TARGET RANGE |
|---|---|---|---|---|---|---|
| Sample 52 | | | | | | 4-6 |
| Sample 53 | | | | | | 7-9 |
| Sample 54 | | | | | | 4-7 |
| Sample 55 | | | | | | 3-5 |

**STOP HERE AND NOTIFY THE RESEARCHER YOU ARE READY TO MOVE ON.**

APPENDIX Y

CRITERION TEST INSTRUCTIONS (FOR EXPERIMENTAL GROUP ONLY)*

The Criterion Test

Now you are ready for The Criterion Test, which is included for the purpose of evaluating whether you have acquired the ability to make online measurement of speech naturalness produced by a variety of speakers who stutter. This part of the training program contains eight 3-minute samples not presented before. The task is exactly the same as the one you just completed, i.e., that you rate naturalness at one-minute intervals for each speaker while watching and listening to the speech samples. For the purpose of the Criterion Test, however, you will have only one opportunity to view and record the speech parameters of each speaker.

What follows is a list of instructions explaining how to take the Criterion Test. Review them first, then return to #1 and begin.

1. Listen to a bit of each speech sample before you begin recoding data so that you are familiar with each speaker's speech pattern. Return to Sample 56, and begin.

2. **When you are ready to begin, start the sample. When the sample begins, immediately depress the computer's space bar to start the SMS timer.**

3. When you have reached the end of the first sample turn to the Criterion Test data sheet (at the end of these instructions). Record your data in the spaces beside each sample number and proceed to the next sample. Continue in this fashion until you have successfully rated the speech naturalness of all 8 speech samples. When you have finished rating sample 63, please give your data sheet to the researcher.

*[After the participant has rated the speech naturalness of all 8 Criterion Test samples the researcher will assess whether the participant's ratings fall into the* SMS *target ranges. If the participant meets the criterion of scoring three consecutive samples within the target naturalness range he or she will complete the first posttraining rating task. If the participant does not meet this criterion, he or she will retrain using the* SMS *system and retake the Criterion Test. If after two training trials the participant still does not score three consecutive samples within the target naturalness range, he or she will continue in the study, but his or her data will be separated from the participants who pass the Criterion Test and they will not be counted in the group sample.]*

* Title not included on copies given to participants.

Participant #: _____

PART FIVE: CRITERION TEST for experimental group

DATA RECORDING SHEET

Samples 56-63

| SAMPLE | YOUR DATA |
|---|---|
| 56 | |
| | |
| 57 | |
| | |
| 58 | |
| | |
| 59 | |
| | |
| 60 | |
| | |
| 61 | |
| | |
| 62 | |
| | |
| 63 | |

# APPENDIX Z

## POSTTRAINING RATING SESSIONS (FOR EXPERIMENTAL GROUP) INSTRUCTIONS TO PARTICIPANTS*

As in your previous sessions, we are studying what makes speech sound natural or unnatural. In order to study this, we are asking you to rate the speech naturalness of some samples of children speaking. When we begin you will hear a number of 30-second speech samples. A sample number will introduce each one.

Your task is to rate the naturalness of each speech sample after viewing the entire 30-second clip. You will be prompted to "Record Naturalness Rating" after each sample is finished. Please record your judgment for each sample after the sample ends, during the silent period, rather than during the sample.

If the speech sample sounds highly natural to you, write a 1 on the data collection form beside the matching sample number. If you think the sample sounds highly unnatural, write a 9 beside the sample number. If the sample sounds somewhere between highly natural and highly unnatural, write the appropriate number on the form. Do not hesitate to use the ends of the scale (1 or 9) when appropriate.

"Naturalness" will not be defined for you. Please make your ratings based on how natural or unnatural the speech sounds to you. When making your ratings, please compare the children's speech to age appropriate peers rather than to adult speakers.

You may recognize some of the video samples from your previous sessions. During today's session you should try to make your judgments in a manner that is consistent with the way that you were judging speech naturalness at the END of your training session. Try to make your judgments today in a manner that you believe would be "correct" according to what you learned in your training session.

The samples will be played without pause and samples will not be replayed (based on Martin et al., 1984).

* Title not included on copies given to participants.

APPENDIX AA

INSTRUCTIONS TO RATE SPEECH NATURALNESS OF *SMS* SAMPLES (FOR EXPOSURE CONTROL GROUP ONLY)*

For this session you will use a computer program to rate the speech naturalness of various speech samples using the 9-point scale you have used in previous sessions before rating samples the way we have done in past sessions. For these first ratings you will rate the speech naturalness of 1, 2 & 3 minute speech samples every 60 seconds. You will use a computer program called the "SMS" to do so. This computer program will prompt you to rate speech naturalness every 60 seconds.

1. You will access each sample from the window currently open on the right side of your computer screen (The samples will be in a folder using the same numbers as your data collection form). You will record your naturalness ratings using the *SMS* program currently open on the left side of your computer screen. Prepare to listen to Sample 40.

2. Double click the sample you need to open it in quick time.

3. Before rating, play a brief bit of the sample so that you are familiar with the speaker's style of speech and have a general idea of its naturalness.

**4.** When you are ready to begin, start the sample. When the sample begins, immediately depress the computer's space bar to start the SMS timer. Your job is to listen to each sample and concentrate on its naturalness, having in mind a rating between 1 and 9. When the naturalness tone sounds and the "NA" signal appears in the bottom left corner of the screen, use the number row on the computer keyboard to indicate your rating. You must press a number key within 5 seconds of the signal. At the completion of the sample the computer will automatically display your naturalness rating on the data summary screen. For samples that require multiple naturalness ratings, the cumulative averaged rating will be displayed in the NAT box during data collection, and the final averaged naturalness score will appear on the data summary screen at the end of the sample. Ignore any other information that appears on the blue screen.

5. When you have completed the naturalness rating(s) for a given sample, turn to the data summary page in front of you and record your rating in the box beside the correct sample number. You will then begin the next sample.

6. **When you are ready to begin, start the sample. When the sample begins, immediately depress the computer's space bar to start the SMS timer.**

   *After the control participants complete speech sample number 55 they will immediately rate the Criterion Test samples 56-63.*

***** Title not included on copies given to participants.

Participant #: _____

PART FOUR: RATING NATURALNESS

DATA RECORDING SHEETS for exposure control group*

| Speech Sample | Naturalness Rating |
|---|---|
| 40 | |
| 41 | |
| 42 | |
| 43 | |
| 44 | |
| 45 | |
| 46 | |
| 47 | |
| 48 | |
| 49 | |
| 50 | |
| 51 | |

| Speech Sample | Naturalness Rating |
|---|---|
| 52 | |
| 53 | |
| 54 | |
| 55 | |
| 56 | |
| 57 | |
| 58 | |
| 59 | |
| 60 | |
| 61 | |
| 62 | |
| 63 | |

* Title not included on copies given to participants.

APPENDIX AB

STUDY TWO: MEAN SPEECH NATURALNESS SCORE BY GROUP FOR EACH SPEECH SAMPLE FOR EACH RATING OCCASION

Training Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| NK 1.2 | 2.83 | 2.22 | 3.39 | 3.06 |
| NK 1.3 | 1.72 | 2.0 | 2.78 | 2.5 |
| NK 1.4 | 2.22 | 2.28 | 3.0 | 2.78 |
| NK 1.5 | 2.72 | 1.78 | 3.11 | 2.56 |
| NK 1.6 | 1.94 | 1.55 | 2.72 | 2.5 |
| NK 2.1 | 2.11 | 1.83 | 2.94 | 2.78 |
| NK 2.2 | 1.39 | 1.56 | 2.33 | 2.11 |
| NK 4.1 | 1.78 | 1.56 | 2.33 | 2.28 |
| NK 4.2 | 1.33 | 1.22 | 2.11 | 1.94 |
| 4.1 | 1.78 | 1.56 | 2.33 | 2.28 |
| 4.3 | 4.28 | 4.11 | 4.89 | 4.44 |
| 5.2 | 7.06 | 7.06 | 6.78 | 6.89 |
| 6.1 | 5.67 | 7.56 | 7.56 | 7.39 |
| 6.4 | 7.22 | 7.11 | 7.78 | 7.83 |
| 6.5 | 7.50 | 8.28 | 7.94 | 7.50 |
| 7.2 | 7.50 | 7.56 | 7.28 | 7.06 |
| 7.3 | 8.33 | 7.83 | 8.00 | 7.39 |
| 8.1 | 8.56 | 8.50 | 8.44 | 7.94 |
| 8.2 | 8.72 | 8.83 | 8.56 | 8.33 |
| 8.3 | 7.72 | 7.78 | 7.97 | 7.56 |
| 8.4 | 8.44 | 8.44 | 8.06 | 7.89 |
| 8.5 | 7.94 | 7.72 | 8.00 | 7.61 |
| 8.6 | 8.67 | 8.72 | 8.61 | 8.17 |
| 9.2 | 4.83 | 5.89 | 6.67 | 5.33 |
| 9.3 | 5.72 | 5.72 | 6.22 | 5.89 |
| 11.1 | 1.72 | 2.22 | 4.00 | 3.83 |
| 11.2 | 2.44 | 2.67 | 4.44 | 3.89 |
| 11.3 | 3.67 | 3.11 | 4.61 | 4.22 |
| 11.4 | 2.33 | 2.33 | 3.83 | 3.83 |
| 11.5 | 3.11 | 2.89 | 4.33 | 4.06 |
| 13.1 | 7.56 | 7.28 | 7.50 | 7.56 |
| 13.2 | 8.11 | 8.06 | 7.89 | 7.56 |
| 14.1 | 5.33 | 6.00 | 6.50 | 6.11 |
| 14.4 | 6.89 | 6.94 | 7.28 | 6.83 |
| 14.5 | 4.94 | 5.28 | 6.50 | 5.72 |
| 14.6 | 6.00 | 6.83 | 6.47 | 6.17 |

Exposure Control Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| NK 1.2 | 3.11 | 2.72 | 2.56 | 2.00 |
| NK 1.3 | 1.83 | 2.06 | 1.67 | 1.56 |
| NK 1.4 | 2.22 | 2.50 | 2.00 | 1.89 |
| NK 1.5 | 2.06 | 1.78 | 1.89 | 1.72 |
| NK 1.6 | 1.61 | 1.56 | 1.56 | 1.39 |
| NK 2.1 | 2.50 | 2.28 | 2.39 | 2.00 |
| NK 2.2 | 1.50 | 1.61 | 1.33 | 1.17 |
| NK4.1 | 1.89 | 1.89 | 1.72 | 1.28 |
| NK 4.2 | 1.39 | 1.17 | 1.44 | 1.28 |
| 4.1 | 4.89 | 4.39 | 4.78 | 4.56 |
| 4.3 | 3.78 | 4.22 | 4.83 | 4.11 |
| 5.2 | 5.56 | 6.61 | 6.22 | 6.44 |
| 6.1 | 5.67 | 6.94 | 6.67 | 6.78 |
| 6.4 | 6.72 | 7.28 | 7.50 | 7.33 |
| 6.5 | 7.11 | 8.00 | 7.33 | 6.83 |
| 7.2 | 7.06 | 7.17 | 7.06 | 7.06 |
| 7.3 | 7.94 | 7.72 | 7.89 | 7.83 |
| 8.1 | 8.33 | 8.56 | 8.50 | 8.33 |
| 8.2 | 8.50 | 8.67 | 8.56 | 8.50 |
| 8.3 | 7.50 | 7.39 | 7.78 | 7.78 |
| 8.4 | 7.89 | 8.22 | 8.17 | 8.22 |
| 8.5 | 7.56 | 8.06 | 8.17 | 7.89 |
| 8.6 | 8.39 | 8.67 | 8.39 | 8.50 |
| 9.2 | 5.00 | 5.67 | 6.33 | 4.72 |
| 9.3 | 5.17 | 5.78 | 5.61 | 5.06 |
| 11.1 | 1.61 | 2.22 | 2.94 | 2.83 |
| 11.2 | 2.89 | 2.67 | 3.33 | 3.11 |
| 11.3 | 3.61 | 3.39 | 4.00 | 4.06 |
| 11.4 | 2.44 | 3.06 | 3.22 | 2.72 |
| 11.5 | 3.50 | 3.33 | 4.28 | 3.28 |
| 13.1 | 7.56 | 7.83 | 8.00 | 7.78 |
| 13.2 | 8.06 | 7.94 | 8.22 | 7.78 |
| 14.1 | 6.11 | 6.06 | 6.00 | 6.06 |
| 14.4 | 6.72 | 7.17 | 7.33 | 6.78 |
| 14.5 | 5.78 | 6.22 | 6.39 | 5.72 |
| 14.6 | 5.61 | 6.11 | 6.22 | 5.83 |

Control Group

| Speech Sample | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|
| NK 1.2 | 3.28 | 3.00 | 2.44 | 2.17 |
| NK 1.3 | 1.89 | 2.39 | 1.89 | 2.06 |
| NK 1.4 | 2.33 | 2.67 | 1.78 | 1.78 |
| NK 1.5 | 2.44 | 2.28 | 2.28 | 1.89 |
| NK 1.6 | 2.33 | 2.06 | 1.61 | 1.72 |
| NK 2.1 | 2.17 | 2.22 | 2.06 | 1.89 |
| NK 2.2 | 1.39 | 1.56 | 1.56 | 1.22 |
| NK4.1 | 1.67 | 1.89 | 1.83 | 1.39 |
| NK 4.2 | 1.33 | 1.22 | 1.28 | 1.00 |
| 4.1 | 4.00 | 4.11 | 3.67 | 3.44 |
| 4.3 | 3.83 | 3.78 | 3.50 | 3.22 |
| 5.2 | 6.33 | 6.61 | 6.56 | 6.83 |
| 6.1 | 5.39 | 6.56 | 6.61 | 7.00 |
| 6.4 | 6.72 | 7.22 | 6.94 | 7.06 |
| 6.5 | 7.22 | 7.28 | 6.83 | 7.33 |
| 7.2 | 7.06 | 6.83 | 7.17 | 6.83 |
| 7.3 | 8.17 | 7.78 | 7.44 | 7.83 |
| 8.1 | 8.33 | 8.67 | 8.28 | 8.50 |
| 8.2 | 8.67 | 8.83 | 8.61 | 8.83 |
| 8.3 | 7.61 | 8.00 | 7.56 | 7.94 |
| 8.4 | 8.28 | 8.44 | 8.22 | 8.50 |
| 8.5 | 7.67 | 7.94 | 7.78 | 8.06 |
| 8.6 | 8.67 | 8.78 | 8.33 | 8.67 |
| 9.2 | 4.50 | 4.94 | 5.44 | 4.83 |
| 9.3 | 4.83 | 5.17 | 5.11 | 4.72 |
| 11.1 | 1.67 | 2.17 | 2.39 | 2.22 |
| 11.2 | 2.56 | 2.61 | 2.17 | 2.39 |
| 11.3 | 3.50 | 3.00 | 2.78 | 2.56 |
| 11.4 | 2.33 | 2.56 | 2.61 | 2.33 |
| 11.5 | 3.11 | 2.83 | 2.67 | 2.56 |
| 13.1 | 6.94 | 6.83 | 6.72 | 7.61 |
| 13.2 | 7.22 | 7.39 | 8.00 | 7.61 |
| 14.1 | 5.28 | 5.44 | 4.67 | 4.50 |
| 14.4 | 6.28 | 5.94 | 6.06 | 5.44 |
| 14.5 | 4.44 | 4.94 | 5.22 | 5.22 |
| 14.6 | 5.00 | 5.78 | 4.94 | 5.11 |

APPENDIX AC
STUDY TWO: SPEECH NATURALNESS RATINGS FOR ALL SPEECH SAMPLES BY ALL JUDGES ON ALL OCCASIONS BY GROUP

Experimental Group

| Speech Sample | NK 1.2 | | | | NK 1.3 | | | | NK 1.4 | | | | NK 1.5 | | | | NK 1.6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 4 | 3 | 4 | 3 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 4 | 1 | 3 | 2 | 3 | 1 | 2 | 2 |
| 4 | 3 | 3 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 2 | 4 | 2 | 2 | 2 | 1 | 2 | 1 | 2 |
| 7 | 1 | 2 | 3 | 4 | 1 | 1 | 3 | 3 | 1 | 2 | 4 | 3 | 2 | 1 | 4 | 4 | 1 | 1 | 4 | 2 |
| 11 | 3 | 2 | 3 | 5 | 3 | 2 | 2 | 2 | 2 | 1 | 2 | 4 | 4 | 3 | 4 | 3 | 1 | 2 | 2 | 3 |
| 15 | 4 | 1 | 4 | 2 | 3 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 2 | 3 | 1 | 1 | 2 |
| 18 | 1 | 1 | 3 | 3 | 1 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 1 | 1 | 3 | 3 | 1 | 1 | 2 | 2 |
| 21 | 2 | 2 | 3 | 3 | 1 | 1 | 3 | 3 | 2 | 1 | 4 | 4 | 2 | 2 | 4 | 3 | 2 | 1 | 4 | 4 |
| 24 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 27 | 5 | 1 | 4 | 2 | 2 | 2 | 5 | 2 | 2 | 2 | 4 | 1 | 3 | 1 | 4 | 2 | 1 | 1 | 3 | 1 |
| 30 | 2 | 2 | 5 | 3 | 1 | 1 | 4 | 2 | 1 | 1 | 2 | 2 | 3 | 1 | 2 | 1 | 1 | 1 | 3 | 2 |
| 34 | 3 | 2 | 5 | 4 | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 1 | 3 | 3 |
| 37 | 1 | 2 | 3 | 3 | 2 | 1 | 3 | 3 | 2 | 2 | 3 | 3 | 1 | 2 | 3 | 3 | 2 | 1 | 3 | 3 |
| 40 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 2 | 2 | 3 | 2 | 3 | 2 |
| 43 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 |
| 46 | 5 | 5 | 5 | 5 | 3 | 5 | 4 | 5 | 4 | 5 | 5 | 5 | 5 | 4 | 5 | 5 | 4 | 4 | 5 | 5 |
| 49 | 4 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 1 | 2 | 1 | 1 | 1 | 2 | 2 |
| 52 | 4 | 4 | 4 | 4 | 2 | 3 | 3 | 4 | 2 | 3 | 4 | 4 | 2 | 2 | 4 | 4 | 3 | 2 | 5 | 4 |
| 55 | 3 | 4 | 4 | 3 | 2 | 4 | 3 | 4 | 5 | 4 | 3 | 3 | 4 | 2 | 4 | 3 | 3 | 3 | 3 | 2 |
| Speech Sample | NK 2.1 | | | | NK 2.2 | | | | NK 4.1 | | | | NK 4.2 | | | | 4.1 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 5 | 4 | 5 | 3 |
| 4 | 2 | 3 | 2 | 2 | 2 | 4 | 1 | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 5 | 5 | 4 | 4 |
| 7 | 3 | 3 | 5 | 5 | 1 | 2 | 4 | 2 | 1 | 1 | 4 | 3 | 1 | 1 | 3 | 3 | 1 | 2 | 3 | 3 |
| 11 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 1 | 1 | 3 | 2 | 2 | 5 | 5 | 3 |
| 15 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 5 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 7 | 5 | 5 | 4 |
| 18 | 3 | 2 | 2 | 4 | 3 | 1 | 5 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 8 | 3 | 6 | 6 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 1 | 1 | 3 | 3 | 1 | 1 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 1 | 3 | 3 | 7 | 3 | 6 | 4 |
| 24 | 2 | 1 | 3 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 5 | 4 | 6 |
| 27 | 4 | 1 | 4 | 2 | 3 | 2 | 3 | 2 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 1 | 5 | 5 | 5 | 3 |
| 30 | 3 | 2 | 4 | 4 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 7 | 6 | 6 | 4 |
| 34 | 2 | 2 | 5 | 4 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 3 | 6 | 5 | 4 |
| 37 | 1 | 1 | 3 | 3 | 1 | 1 | 3 | 2 | 1 | 1 | 3 | 3 | 1 | 1 | 3 | 3 | 3 | 2 | 4 | 3 |
| 40 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 6 | 5 | 3 |
| 43 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 5 | 3 | 6 | 5 |
| 46 | 3 | 2 | 3 | 4 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 4 | 3 | 3 | 4 | 3 | 7 | 6 | 6 | 5 |
| 49 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7 | 4 | 5 | 5 |
| 52 | 1 | 4 | 4 | 4 | 1 | 1 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 1 | 2 | 3 | 5 | 8 | 6 | 6 |
| 55 | 5 | 3 | 4 | 4 | 1 | 4 | 4 | 5 | 3 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 5 | 6 | 5 | 5 |
| Speech Sample | 4.3 | | | | 5.2 | | | | 6.1 | | | | 6.4 | | | | 6.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 6 | 5 | 6 | 3 | 6 | 7 | 8 | 7 | 7 | 9 | 9 | 9 | 8 | 7 | 9 | 9 | 9 | 9 | 8 | 9 |
| 4 | 5 | 5 | 3 | 4 | 7 | 8 | 5 | 7 | 6 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 8 |
| 7 | 2 | 2 | 4 | 3 | 7 | 7 | 6 | 6 | 8 | 8 | 7 | 7 | 6 | 8 | 7 | 8 | 8 | 8 | 7 | 8 |
| 11 | 3 | 4 | 4 | 4 | 5 | 3 | 5 | 4 | 5 | 7 | 7 | 6 | 8 | 6 | 9 | 7 | 5 | 8 | 8 | 4 |
| 15 | 6 | 4 | 5 | 4 | 8 | 9 | 7 | 7 | 7 | 8 | 8 | 8 | 9 | 8 | 7 | 8 | 9 | 9 | 8 | 8 |
| 18 | 8 | 6 | 6 | 5 | 9 | 9 | 7 | 9 | 6 | 8 | 7 | 7 | 9 | 8 | 7 | 9 | 8 | 8 | 8 | 8 |
| 21 | 5 | 4 | 5 | 5 | 8 | 4 | 6 | 8 | 7 | 7 | 8 | 8 | 8 | 7 | 7 | 8 | 9 | 9 | 8 | 8 |
| 24 | 1 | 1 | 4 | 5 | 4 | 4 | 4 | 6 | 6 | 8 | 7 | 8 | 6 | 4 | 7 | 8 | 9 | 7 | 8 | 8 |
| 27 | 5 | 4 | 5 | 3 | 7 | 8 | 9 | 6 | 7 | 9 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 7 | 8 | 9 |
| 30 | 5 | 6 | 6 | 5 | 9 | 7 | 9 | 9 | 7 | 8 | 8 | 8 | 8 | 8 | 9 | 8 | 9 | 9 | 9 | 8 |
| 34 | 3 | 4 | 4 | 6 | 7 | 9 | 7 | 7 | 6 | 5 | 7 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 8 | 7 |
| 37 | 3 | 1 | 4 | 3 | 8 | 7 | 5 | 6 | 1 | 4 | 5 | 4 | 4 | 3 | 6 | 5 | 4 | 6 | 6 | 5 |
| 40 | 3 | 3 | 3 | 3 | 8 | 9 | 7 | 7 | 4 | 7 | 8 | 8 | 7 | 8 | 8 | 8 | 6 | 8 | 8 | 8 |
| 43 | 2 | 2 | 5 | 5 | 5 | 7 | 8 | 7 | 2 | 7 | 8 | 7 | 5 | 5 | 8 | 7 | 6 | 8 | 7 | 7 |
| 46 | 5 | 6 | 7 | 6 | 7 | 8 | 7 | 7 | 5 | 7 | 8 | 7 | 6 | 8 | 8 | 8 | 6 | 9 | 9 | 7 |
| 49 | 3 | 4 | 5 | 5 | 8 | 7 | 6 | 6 | 6 | 8 | 7 | 6 | 7 | 6 | 6 | 5 | 8 | 9 | 8 | 6 |
| 52 | 6 | 7 | 7 | 6 | 6 | 6 | 8 | 7 | 8 | 9 | 8 | 7 | 9 | 9 | 9 | 8 | 7 | 9 | 8 | 8 |
| 55 | 6 | 6 | 5 | 5 | 8 | 8 | 8 | 8 | 4 | 8 | 7 | 9 | 7 | 8 | 7 | 9 | 5 | 8 | 8 | 9 |
| Speech Sample | 7.2 | | | | 7.3 | | | | 8.1 | | | | 8.2 | | | | 8.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 6 | 6 | 8 | 6 | 8 | 8 | 9 | 7 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 8 | 9 | 8 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 7 | 7 | 7 | 7 | 9 | 7 | 8 | 7 | 8 | 8 | 8 | 8 | 8 | 9 | 9 | 8 | 7 | 8 | 8 | 8 |
| 7 | 9 | 8 | 7 | 7 | 9 | 9 | 7 | 8 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 6 | 8 | 7 | 7 |
| 11 | 6 | 5 | 6 | 7 | 9 | 6 | 8 | 6 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 8 | 8 | 7 | 8 |
| 15 | 8 | 7 | 6 | 6 | 9 | 7 | 7 | 7 | 9 | 8 | 7 | 7 | 9 | 9 | 7 | 7 | 8 | 7 | 8 | 6 |
| 18 | 9 | 9 | 8 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 8 | 7 | 9 | 9 | 7 | 8 | 9 | 7 | 8 | 7 |
| 21 | 8 | 7 | 7 | 7 | 9 | 8 | 8 | 7 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 8 | 9 | 9 |
| 24 | 6 | 5 | 6 | 7 | 8 | 7 | 8 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 8 | 6 | 7 | 7 | 9 |
| 27 | 7 | 9 | 8 | 7 | 8 | 8 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 7 | 8 | 8 |
| 30 | 7 | 8 | 8 | 7 | 8 | 8 | 8 | 7 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 7 |
| 34 | 8 | 8 | 7 | 8 | 8 | 9 | 9 | 9 | 8 | 8 | 9 | 8 | 8 | 9 | 9 | 9 | 6 | 7 | 7 | 7 |
| 37 | 9 | 9 | 7 | 5 | 6 | 9 | 7 | 6 | 9 | 9 | 8 | 6 | 9 | 9 | 8 | 7 | 9 | 8 | 8 | 5 |
| 40 | 7 | 7 | 9 | 8 | 8 | 7 | 9 | 8 | 7 | 6 | 8 | 7 | 8 | 8 | 7 | 8 | 7 | 7 | 7.5 | 7 |
| 43 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 8 | 9 | 8 | 7 | 7 | 8 | 7 |
| 46 | 7 | 7 | 7 | 6 | 8 | 6 | 7 | 7 | 8 | 9 | 9 | 7 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 8 |
| 49 | 8 | 9 | 8 | 7 | 9 | 9 | 7 | 6 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 9 |
| 52 | 7 | 8 | 6 | 7 | 8 | 6 | 8 | 8 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 8 | 8 | 9 | 8 | 7 |
| 55 | 8 | 8 | 8 | 8 | 8 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| Speech Sample | 8.4 | | | | 8.5 | | | | 8.6 | | | | 9.2 | | | | 9.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 2 | 6 | 8 | 6 | 6 | 4 | 7 | 8 |
| 4 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 9 | 9 | 8 | 8 | 4 | 6 | 7 | 6 | 6 | 8 | 5 | 7 |
| 7 | 9 | 9 | 8 | 9 | 9 | 7 | 7 | 7 | 9 | 9 | 9 | 8 | 7 | 5 | 4 | 5 | 5 | 7 | 5 | 5 |
| 11 | 8 | 9 | 8 | 9 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 4 | 5 | 7 | 4 | 3 | 2 | 5 | 4 |
| 15 | 9 | 8 | 8 | 7 | 8 | 7 | 7 | 6 | 9 | 8 | 7 | 8 | 7 | 6 | 6 | 5 | 7 | 3 | 6 | 5 |
| 18 | 9 | 8 | 8 | 7 | 9 | 8 | 7 | 7 | 9 | 9 | 9 | 8 | 7 | 7 | 9 | 5 | 9 | 7 | 7 | 6 |
| 21 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 6 | 7 | 7 | 6 | 6 | 7 | 7 |
| 24 | 8 | 9 | 7 | 9 | 8 | 9 | 7 | 7 | 9 | 9 | 9 | 9 | 1 | 1 | 6 | 3 | 1 | 3 | 3 | 4 |
| 27 | 9 | 9 | 8 | 9 | 8 | 6 | 9 | 9 | 9 | 9 | 9 | 9 | 4 | 9 | 7 | 8 | 7 | 9 | 9 | 7 |
| 30 | 8 | 9 | 7 | 6 | 8 | 8 | 9 | 7 | 9 | 9 | 9 | 9 | 7 | 8 | 7 | 6 | 6 | 8 | 8 | 8 |
| 34 | 7 | 7 | 8 | 8 | 6 | 7 | 8 | 7 | 8 | 8 | 9 | 7 | 3 | 5 | 7 | 5 | 4 | 5 | 5 | 5 |
| 37 | 8 | 9 | 8 | 6 | 6 | 9 | 8 | 6 | 9 | 9 | 9 | 7 | 8 | 8 | 6 | 3 | 8 | 7 | 6 | 5 |
| 40 | 7 | 6 | 8 | 7 | 6 | 7 | 8 | 7 | 6 | 8 | 8 | 7 | 5 | 6 | 7 | 6 | 6 | 7 | 7 | 7 |
| 43 | 8 | 8 | 8 | 7 | 7 | 7 | 8 | 8 | 8 | 7 | 8 | 8 | 2 | 5 | 6 | 5 | 7 | 6 | 7 | 6 |
| 46 | 8 | 9 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 9 | 8 | 8 | 7 | 7 | 8 | 7 | 6 | 7 | 8 | 6 |
| 49 | 9 | 8 | 8 | 8 | 8 | 7 | 8 | 9 | 9 | 9 | 9 | 9 | 3 | 4 | 6 | 6 | 4 | 3 | 5 | 4 |
| 52 | 9 | 9 | 8 | 7 | 9 | 8 | 8 | 8 | 9 | 9 | 8 | 7 | 7 | 6 | 7 | 6 | 7 | 7 | 6 | 6 |

| 55 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 2 | 6 | 5 | 3 | 5 | 4 | 6 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Speech Sample | 11.1 | | | | 11.2 | | | | 11.3 | | | | 11.4 | | | | 11.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 2 | 2 | 3 | 4 | 2 | 2 | 4 | 5 | 5 | 2 | 6 | 4 | 2 | 3 | 4 | 4 | 3 | 3 | 6 | 3 |
| 4 | 2 | 4 | 4 | 5 | 4 | 7 | 4 | 4 | 4 | 5 | 3 | 4 | 3 | 7 | 2 | 4 | 5 | 4 | 4 | 5 |
| 7 | 1 | 1 | 4 | 4 | 3 | 3 | 5 | 5 | 5 | 4 | 4 | 4 | 2 | 3 | 4 | 4 | 3 | 5 | 4 | 4 |
| 11 | 2 | 3 | 3 | 3 | 3 | 2 | 3 | 4 | 3 | 2 | 5 | 4 | 3 | 1 | 3 | 3 | 3 | 3 | 4 | 2 |
| 15 | 2 | 1 | 3 | 4 | 4 | 3 | 5 | 4 | 6 | 5 | 5 | 5 | 3 | 2 | 3 | 3 | 6 | 5 | 4 | 4 |
| 18 | 1 | 1 | 4 | 3 | 2 | 1 | 5 | 4 | 7 | 2 | 4 | 5 | 2 | 1 | 3 | 4 | 4 | 2 | 4 | 5 |
| 21 | 1 | 1 | 4 | 3 | 2 | 2 | 4 | 3 | 3 | 3 | 4 | 4 | 3 | 1 | 4 | 4 | 4 | 2 | 4 | 4 |
| 24 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 3 | 4 | 1 | 2 | 4 | 1 | 1 | 2 | 4 | 1 | 1 | 2 | 3 |
| 27 | 2 | 2 | 7 | 4 | 2 | 1 | 8 | 3 | 4 | 1 | 6 | 4 | 5 | 2 | 7 | 4 | 3 | 2 | 8 | 5 |
| 30 | 3 | 4 | 6 | 4 | 3 | 3 | 5 | 5 | 2 | 3 | 6 | 3 | 2 | 3 | 5 | 5 | 3 | 3 | 5 | 4 |
| 34 | 3 | 3 | 5 | 5 | 3 | 2 | 6 | 5 | 3 | 5 | 6 | 6 | 3 | 3 | 5 | 6 | 2 | 3 | 6 | 5 |
| 37 | 1 | 1 | 4 | 3 | 1 | 1 | 4 | 3 | 2 | 2 | 4 | 3 | 1 | 1 | 4 | 3 | 1 | 2 | 4 | 3 |
| 40 | 1 | 1 | 4 | 4 | 1 | 1 | 4 | 3 | 2 | 2 | 3 | 4 | 1 | 1 | 3 | 3 | 1 | 1 | 4 | 5 |
| 43 | 1 | 1 | 2 | 4 | 1 | 1 | 2 | 3 | 2 | 2 | 6 | 4 | 1 | 1 | 4 | 3 | 2 | 1 | 3 | 3 |
| 46 | 3 | 3 | 5 | 5 | 3 | 4 | 5 | 4 | 4 | 4 | 5 | 4 | 3 | 3 | 4 | 4 | 4 | 4 | 5 | 5 |
| 49 | 2 | 2 | 2 | 2 | 1 | 3 | 3 | 2 | 1 | 3 | 3 | 3 | 1 | 2 | 3 | 2 | 3 | 3 | 2 | 3 |
| 52 | 2 | 6 | 5 | 5 | 4 | 6 | 5 | 5 | 5 | 5 | 6 | 6 | 3 | 4 | 5 | 5 | 4 | 5 | 5 | 6 |
| 55 | 1 | 3 | 5 | 4 | 4 | 5 | 6 | 5 | 4 | 5 | 5 | 5 | 3 | 3 | 4 | 4 | 4 | 3 | 4 | 4 |
| Speech Sample | 13.1 | | | | 13.2 | | | | 14.1 | | | | 14.4 | | | | 14.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 8 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 7 | 6 | 6 | 7 | 3 |
| 4 | 8 | 9 | 7 | 8 | 8 | 9 | 8 | 8 | 6 | 6 | 7 | 6 | 8 | 8 | 8 | 7 | 7 | 6 | 6 | 7 |
| 7 | 8 | 7 | 6 | 6 | 9 | 8 | 7 | 8 | 6 | 6 | 6 | 6 | 8 | 8 | 6 | 7 | 5 | 6 | 6 | 6 |
| 11 | 6 | 6 | 7 | 5 | 7 | 7 | 5 | 7 | 7 | 6 | 5 | 6 | 7 | 5 | 6 | 5 | 3 | 4 | 6 | 4 |
| 15 | 8 | 6 | 7 | 8 | 9 | 8 | 8 | 7 | 6 | 7 | 7 | 6 | 7 | 6 | 7 | 7 | 7 | 6 | 7 | 5 |
| 18 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 3 | 3 | 6 | 6 | 9 | 9 | 7 | 6 | 6 | 3 | 6 | 7 |
| 21 | 8 | 7 | 7 | 7 | 8 | 7 | 8 | 8 | 7 | 7 | 5 | 7 | 8 | 8 | 7 | 7 | 8 | 7 | 6 | 6 |
| 24 | 6 | 3 | 6 | 7 | 7 | 3 | 7 | 6 | 7 | 6 | 6 | 6 | 7 | 4 | 8 | 7 | 3 | 1 | 5 | 4 |
| 27 | 9 | 6 | 9 | 8 | 9 | 9 | 8 | 8 | 5 | 6 | 8 | 8 | 7 | 8 | 9 | 8 | 5 | 7 | 9 | 6 |
| 30 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 7 | 5 | 5 | 5 | 4 | 7 | 6 | 6 | 5 | 6 | 4 | 5 | 5 |
| 34 | 8 | 8 | 8 | 8 | 7 | 9 | 8 | 9 | 4 | 5 | 6 | 5 | 5 | 7 | 7 | 8 | 4 | 4 | 6 | 5 |
| 37 | 7 | 9 | 7 | 6 | 8 | 9 | 6 | 6 | 3 | 5 | 5 | 4 | 7 | 5 | 6 | 5 | 3 | 2 | 5 | 4 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | 7 | 8 | 8 | 9 | 8 | 9 | 9 | 8 | 4 | 4 | 7 | 7 | 3 | 5 | 7 | 7 | 4 | 5 | 6 | 7 |
| 43 | 8 | 8 | 8 | 8 | 7 | 8 | 8 | 7 | 3 | 5 | 7 | 7 | 5 | 6 | 7 | 7 | 3 | 6 | 7 | 6 |
| 46 | 7 | 8 | 8 | 8 | 8 | 8 | 9 | 7 | 6 | 8 | 8 | 6 | 6 | 6 | 8 | 7 | 4 | 6 | 7 | 7 |
| 49 | 7 | 5 | 8 | 7 | 9 | 8 | 8 | 8 | 6 | 7 | 7 | 5 | 8 | 9 | 8 | 7 | 5 | 6 | 7 | 5 |
| 52 | 8 | 9 | 8 | 7 | 8 | 9 | 9 | 7 | 7 | 9 | 8 | 6 | 8 | 8 | 8 | 7 | 6 | 9 | 9 | 8 |
| 55 | 6 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 5 | 7 | 7 | 8 | 7 | 9 | 8 | 9 | 4 | 7 | 7 | 8 |

| Speech Sample | 14.6 | | | |
|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 |
| 1 | 8 | 7 | 8 | 7 |
| 4 | 6 | 7 | 7 | 6 |
| 7 | 4 | 7 | 7 | 7 |
| 11 | 4 | 6 | 3 | 4 |
| 15 | 8 | 7 | 6 | 6 |
| 18 | 9 | 9 | 6 | 6 |
| 21 | 8 | 7 | 6 | 7 |
| 24 | 5 | 7 | 6 | 7 |
| 27 | 6 | 8 | 8 | 7 |
| 30 | 4 | 6 | 6 | 4 |
| 34 | 6 | 6 | 6 | 5 |
| 37 | 7 | 3 | 6 | 4 |
| 40 | 4 | 7 | 6.5 | 7 |
| 43 | 5 | 6 | 6 | 7 |
| 46 | 6 | 7 | 7 | 7 |
| 49 | 5 | 8 | 7 | 5 |
| 52 | 7 | 7 | 7 | 7 |
| 55 | 6 | 8 | 8 | 8 |

Exposure Control Group

| Speech Sample | NK 1.2 | | | | NK 1.3 | | | | NK 1.4 | | | | NK 1.5 | | | | NK 1.6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 3 | 2 | 3 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| 8 | 3 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| 12 | 3 | 3 | 3 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 2 | 2 | 4 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| 19 | 4 | 4 | 3 | 4 | 1 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 2 | 2 | 3 | 4 | 1 | 2 | 2 | 2 |
| 22 | 5 | 4 | 3 | 2 | 3 | 3 | 1 | 3 | 5 | 3 | 3 | 2 | 4 | 4 | 2 | 4 | 2 | 2 | 1 | 1 |
| 25 | 7 | 6 | 4 | 3 | 5 | 5 | 3 | 3 | 2 | 5 | 2 | 2 | 4 | 1 | 2 | 1 | 2 | 2 | 3 | 2 |
| 28 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 32 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 35 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 3 | 3 | 2 |
| 38 | 3 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 41 | 2 | 2 | 2 | 3 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 3 | 2 |
| 44 | 3 | 4 | 3 | 2 | 3 | 2 | 2 | 1 | 2 | 3 | 3 | 3 | 4 | 2 | 4 | 2 | 1 | 2 | 1 | 1 |
| 47 | 3 | 2 | 2 | 1 | 1 | 3 | 1 | 1 | 3 | 3 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| 50 | 2 | 2 | 2 | 3 | 2 | 3 | 1 | 1 | 2 | 4 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 |
| 53 | 5 | 2 | 2 | 2 | 1 | 3 | 1 | 1 | 5 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 56 | 3 | 4 | 4 | 3 | 3 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 2 | 3 | 3 |
| Speech Sample | NK 2.1 | | | | NK 2.2 | | | | NK 4.1 | | | | NK 4.2 | | | | 4.1 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8 | 7 | 7 | 8 |
| 5 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 7 | 4 | 7 | 7 |
| 8 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 5 | 4 | 5 |
| 12 | 4 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 4 | 3 | 3 |
| 16 | 2 | 4 | 1 | 3 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 1 | 3 | 4 |
| 19 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 4 | 3 | 4 |
| 22 | 4 | 3 | 5 | 2 | 5 | 1 | 1 | 1 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 6 | 4 | 6 | 6 |
| 25 | 4 | 4 | 4 | 4 | 3 | 2 | 2 | 3 | 7 | 4 | 3 | 2 | 2 | 1 | 2 | 2 | 6 | 3 | 5 | 6 |
| 28 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 | 6 | 4 | 4 |
| 32 | 2 | 3 | 2 | 2 | 1 | 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 35 | 4 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 3 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 6 | 4 | 5 | 3 |
| 38 | 3 | 3 | 4 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 2 | 1 | 3 | 3 | 1 | 1 | 4 | 4 | 5 | 4 |
| 41 | 3 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 6 | 6 | 7 | 6 |
| 44 | 4 | 4 | 6 | 3 | 2 | 1 | 4 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 4 | 1 | 4 | 5 | 5 | 4 |
| 47 | 3 | 2 | 3 | 1 | 1 | 2 | 1 | 1 | 4 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 4 | 4 | 4 | 2 |
| 50 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 3 | 6 | 5 | 5 |
| 53 | 3 | 4 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 6 | 7 | 7 | 6 |
| 56 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 4 | 5 | 3 |
| Speech Sample | 4.3 | | | | 5.2 | | | | 6.1 | | | | 6.4 | | | | 6.5 | | | |

| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6 | 7 | 7 | 5 | 2 | 5 | 5 | 8 | 6 | 8 | 8 | 7 | 4 | 7 | 8 | 8 | 5 | 8 | 8 | 8 |
| 5 | 6 | 5 | 8 | 6 | 7 | 7 | 7 | 7 | 4 | 5 | 7 | 7 | 7 | 7 | 8 | 8 | 6 | 6 | 8 | 7 |
| 8 | 5 | 6 | 4 | 5 | 5 | 8 | 7 | 7 | 3 | 5 | 5 | 6 | 6 | 7 | 7 | 5 | 7 | 9 | 5 | 8 |
| 12 | 2 | 3 | 2 | 3 | 5 | 5 | 5 | 6 | 3 | 7 | 6 | 5 | 6 | 6 | 6 | 7 | 6 | 8 | 6 | 7 |
| 16 | 2 | 2 | 4 | 1 | 6 | 7 | 8 | 8 | 3 | 8 | 9 | 8 | 7 | 7 | 9 | 8 | 8 | 8 | 9 | 6 |
| 19 | 4 | 3 | 5 | 4 | 3 | 8 | 4 | 5 | 3 | 6 | 5 | 5 | 4 | 8 | 6 | 6 | 2 | 8 | 7 | 5 |
| 22 | 3 | 2 | 6 | 6 | 7 | 6 | 5 | 7 | 7 | 6 | 6 | 7 | 8 | 6 | 7 | 9 | 9 | 9 | 8 | 9 |
| 25 | 3 | 3 | 5 | 4 | 8 | 8 | 8 | 9 | 8 | 8 | 8 | 8 | 7 | 8 | 8 | 8 | 8 | 9 | 9 | 8 |
| 28 | 6 | 4 | 3 | 3 | 5 | 5 | 4 | 5 | 5 | 6 | 4 | 4 | 7 | 8 | 4 | 5 | 8 | 5 | 3 | 4 |
| 32 | 2 | 2 | 1 | 2 | 9 | 9 | 4 | 6 | 8 | 9 | 7 | 8 | 9 | 8 | 8 | 7 | 9 | 9 | 9 | 8 |
| 35 | 4 | 4 | 4 | 3 | 5 | 5 | 4 | 3 | 4 | 4 | 4 | 3 | 7 | 5 | 6 | 5 | 6 | 6 | 6 | 4 |
| 38 | 3 | 6 | 3 | 4 | 6 | 6 | 7 | 4 | 7 | 6 | 8 | 8 | 7 | 6 | 8 | 8 | 7 | 8 | 5 | 7 |
| 41 | 4 | 5 | 6 | 7 | 4 | 6 | 8 | 8 | 7 | 8 | 6 | 8 | 5 | 9 | 8 | 9 | 7 | 9 | 7 | 8 |
| 44 | 3 | 4 | 6 | 4 | 5 | 6 | 7 | 7 | 6 | 8 | 7 | 7 | 7 | 8 | 8 | 7 | 8 | 8 | 8 | 6 |
| 47 | 1 | 3 | 4 | 2 | 7 | 7 | 7 | 6 | 7 | 7 | 8 | 7 | 8 | 8 | 9 | 9 | 8 | 8 | 8 | 7 |
| 50 | 4 | 7 | 6 | 6 | 3 | 7 | 7 | 8 | 7 | 9 | 6 | 8 | 8 | 8 | 8 | 9 | 7 | 9 | 9 | 8 |
| 53 | 6 | 6 | 7 | 5 | 7 | 7 | 6 | 6 | 7 | 8 | 8 | 8 | 7 | 8 | 8 | 7 | 9 | 9 | 8 | 7 |
| 56 | 4 | 4 | 6 | 4 | 6 | 7 | 9 | 6 | 7 | 7 | 8 | 8 | 7 | 7 | 9 | 7 | 8 | 8 | 9 | 6 |
| Speech Sample | 7.2 | | | | 7.3 | | | | 8.1 | | | | 8.2 | | | | 8.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 6 | 8 | 8 | 6 | 8 | 8 | 9 | 9 | 8 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 8 | 8 | 9 | 9 |
| 5 | 6 | 7 | 7 | 8 | 8 | 7 | 7 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 8 | 8 | 9 |
| 8 | 8 | 9 | 7 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 7 |
| 12 | 6 | 5 | 5 | 6 | 3 | 7 | 7 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 |
| 16 | 7 | 6 | 8 | 7 | 7 | 7 | 9 | 7 | 8 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 9 |
| 19 | 4 | 9 | 8 | 9 | 8 | 8 | 8 | 6 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 8 | 7 | 7 | 8 |
| 22 | 9 | 7 | 6 | 7 | 9 | 7 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 8 |
| 25 | 8 | 8 | 8 | 8 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 8 | 9 | 8 |
| 28 | 7 | 4 | 3 | 5 | 9 | 5 | 5 | 5 | 9 | 8 | 7 | 7 | 9 | 9 | 8 | 8 | 7 | 7 | 6 | 6 |
| 32 | 9 | 9 | 8 | 7 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 7 | 8 | 8 |
| 35 | 5 | 5 | 5 | 4 | 7 | 5 | 4 | 6 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 6 | 5 | 6 | 6 | 6 |
| 38 | 7 | 6 | 7 | 7 | 7 | 8 | 9 | 8 | 7 | 8 | 8 | 7 | 7 | 7 | 9 | 8 | 6 | 6 | 8 | 8 |
| 41 | 6 | 8 | 8 | 7 | 8 | 9 | 8 | 8 | 6 | 9 | 8 | 9 | 7 | 9 | 9 | 8 | 7 | 7 | 7 | 7 |
| 44 | 9 | 8 | 7 | 7 | 9 | 9 | 9 | 8 | 8 | 7 | 8 | 7 | 8 | 9 | 8 | 8 | 6 | 7 | 7 | 7 |
| 47 | 6 | 6 | 8 | 8 | 9 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 8 | 8 | 8 | 9 | 7 | 8 | 8 | 9 | 8 | 9 | 9 | 9 | 8 | 8 | 8 | 9 | 8 | 6 | 7 | 9 |
| 53 | 8 | 8 | 7 | 6 | 9 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 7 |
| 56 | 8 | 8 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 8 | 9 | 8 |
| Speech Sample | 8.4 | | | | 8.5 | | | | 8.6 | | | | 9.2 | | | | 9.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 9 | 8 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 2 | 2 | 6 | 2 | 2 | 2 | 1 | 2 |
| 5 | 8 | 8 | 9 | 9 | 7 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 5 | 3 | 7 | 6 | 7 | 4 | 7 | 4 |
| 8 | 9 | 8 | 8 | 7 | 6 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 6 | 7 | 4 | 4 | 4 | 6 | 5 | 5 |
| 12 | 6 | 8 | 8 | 8 | 7 | 8 | 6 | 8 | 9 | 9 | 9 | 9 | 7 | 6 | 6 | 5 | 6 | 7 | 6 | 6 |
| 16 | 8 | 8 | 9 | 9 | 8 | 9 | 9 | 8 | 8 | 8 | 9 | 9 | 5 | 5 | 7 | 5 | 5 | 4 | 8 | 5 |
| 19 | 9 | 9 | 8 | 9 | 7 | 8 | 6 | 8 | 9 | 9 | 9 | 9 | 2 | 6 | 3 | 7 | 3 | 5 | 5 | 3 |
| 22 | 9 | 8 | 7 | 8 | 8 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 5 | 5 | 7 | 5 | 5 | 6 | 5 | 7 |
| 25 | 8 | 8 | 9 | 9 | 8 | 7 | 9 | 9 | 8 | 9 | 9 | 9 | 8 | 7 | 7 | 7 | 8 | 9 | 7 | 6 |
| 28 | 7 | 7 | 7 | 7 | 8 | 7 | 7 | 7 | 9 | 9 | 7 | 8 | 6 | 4 | 3 | 3 | 4 | 5 | 3 | 3 |
| 32 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 2 | 7 | 7 | 2 | 8 | 8 | 6 | 3 |
| 35 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 5 | 8 | 7 | 8 | 6 | 4 | 4 | 5 | 4 | 3 | 6 | 4 | 4 |
| 38 | 5 | 8 | 8 | 7 | 5 | 7 | 8 | 7 | 5 | 7 | 4 | 7 | 7 | 7 | 8 | 6 | 6 | 7 | 4 | 7 |
| 41 | 7 | 9 | 7 | 9 | 7 | 8 | 8 | 8 | 6 | 9 | 8 | 8 | 5 | 7 | 8 | 6 | 6 | 6 | 7 | 6 |
| 44 | 8 | 8 | 9 | 8 | 8 | 7 | 8 | 7 | 9 | 8 | 7 | 8 | 3 | 5 | 8 | 5 | 3 | 5 | 7 | 7 |
| 47 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 5 | 5 | 8 | 2 | 6 | 6 | 7 | 3 |
| 50 | 7 | 8 | 8 | 9 | 6 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 4 | 8 | 5 | 7 | 6 | 7 | 6 | 7 |
| 53 | 8 | 9 | 8 | 7 | 9 | 9 | 8 | 6 | 9 | 9 | 9 | 8 | 7 | 8 | 7 | 6 | 6 | 4 | 6 | 7 |
| 56 | 8 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 6 | 8 | 3 | 5 | 7 | 7 | 6 |
| Speech Sample | 11.1 | | | | 11.2 | | | | 11.3 | | | | 11.4 | | | | 11.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 1 | 1 | 3 | 2 | 1 | 1 | 3 | 2 | 1 | 1 | 2 | 6 | 1 | 1 | 2 | 1 | 1 | 1 | 6 | 2 |
| 5 | 3 | 2 | 4 | 4 | 2 | 2 | 3 | 5 | 4 | 3 | 4 | 5 | 4 | 3 | 5 | 5 | 4 | 2 | 4 | 5 |
| 8 | 1 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 2 | 2 | 3 | 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 |
| 12 | 1 | 2 | 4 | 4 | 4 | 4 | 5 | 3 | 3 | 3 | 4 | 5 | 2 | 2 | 4 | 3 | 5 | 5 | 6 | 4 |
| 16 | 1 | 2 | 5 | 5 | 3 | 2 | 7 | 5 | 3 | 4 | 5 | 5 | 2 | 2 | 4 | 4 | 3 | 5 | 6 | 5 |
| 19 | 1 | 2 | 2 | 2 | 1 | 3 | 2 | 2 | 2 | 3 | 3 | 2 | 1 | 4 | 5 | 2 | 3 | 4 | 4 | 2 |
| 22 | 3 | 3 | 2 | 1 | 5 | 2 | 2 | 3 | 6 | 5 | 5 | 2 | 4 | 2 | 2 | 4 | 6 | 4 | 5 | 2 |
| 25 | 3 | 5 | 5 | 5 | 5 | 6 | 5 | 4 | 6 | 5 | 4 | 6 | 3 | 6 | 6 | 5 | 4 | 2 | 5 | 6 |
| 28 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 2 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 3 | 2 | 2 |
| 32 | 1 | 3 | 2 | 2 | 4 | 4 | 2 | 2 | 6 | 5 | 3 | 6 | 3 | 7 | 3 | 2 | 6 | 5 | 3 | 4 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 35 | 2 | 2 | 3 | 2 | 3 | 2 | 3 | 3 | 4 | 3 | 3 | 3 | 4 | 3 | 3 | 1 | 2 | 3 | 3 | 2 |
| 38 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 3 | 2 | 8 | 1 | 2 | 1 | 2 | 1 | 2 | 3 | 4 | 2 |
| 41 | 1 | 3 | 3 | 3 | 1 | 3 | 4 | 4 | 2 | 3 | 3 | 3 | 1 | 4 | 5 | 4 | 2 | 4 | 4 | 4 |
| 44 | 2 | 1 | 3 | 3 | 3 | 2 | 4 | 5 | 4 | 5 | 5 | 4 | 3 | 5 | 4 | 3 | 3 | 3 | 4 | 3 |
| 47 | 1 | 1 | 2 | 3 | 2 | 3 | 3 | 3 | 5 | 2 | 6 | 5 | 1 | 1 | 1 | 2 | 6 | 3 | 5 | 3 |
| 50 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 3 | 1 | 2 | 2 | 4 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 2 |
| 53 | 3 | 3 | 3 | 3 | 6 | 4 | 4 | 3 | 5 | 5 | 4 | 5 | 3 | 3 | 3 | 3 | 6 | 4 | 5 | 4 |
| 56 | 1 | 3 | 4 | 3 | 4 | 3 | 4 | 4 | 4 | 5 | 5 | 4 | 4 | 4 | 4 | 3 | 5 | 5 | 7 | 4 |
| Speech Sample | 13.1 | | | | 13.2 | | | | 14.1 | | | | 14.4 | | | | 14.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 8 | 6 | 7 | 7 | 8 | 8 | 7 | 8 | 9 | 6 | 3 |
| 5 | 8 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 5 | 4 | 6 | 4 | 7 | 6 | 7 | 5 | 5 | 4 | 6 | 6 |
| 8 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 8 | 6 | 4 | 4 | 4 | 5 | 4 | 6 | 5 | 4 | 6 | 5 | 5 |
| 12 | 6 | 7 | 8 | 8 | 8 | 6 | 8 | 8 | 6 | 5 | 5 | 5 | 8 | 8 | 7 | 7 | 6 | 7 | 4 | 4 |
| 16 | 6 | 7 | 9 | 8 | 8 | 7 | 8 | 8 | 6 | 7 | 7 | 7 | 8 | 7 | 8 | 8 | 5 | 6 | 7 | 5 |
| 19 | 7 | 9 | 7 | 6 | 7 | 8 | 8 | 7 | 3 | 7 | 5 | 7 | 5 | 9 | 7 | 4 | 3 | 5 | 4 | 6 |
| 22 | 8 | 7 | 9 | 9 | 8 | 8 | 9 | 7 | 8 | 8 | 9 | 9 | 9 | 7 | 9 | 8 | 9 | 9 | 9 | 9 |
| 25 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 7 | 8 | 9 | 9 | 8 | 9 | 8 | 9 | 9 |
| 28 | 9 | 8 | 7 | 8 | 9 | 6 | 7 | 7 | 4 | 5 | 4 | 4 | 6 | 6 | 5 | 4 | 6 | 6 | 5 | 5 |
| 32 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 7 | 8 | 8 | 5 | 6 | 8 | 9 | 7 | 9 | 7 | 8 | 6 | 7 |
| 35 | 7 | 6 | 5 | 4 | 6 | 7 | 5 | 5 | 6 | 6 | 4 | 5 | 5 | 4 | 6 | 5 | 5 | 7 | 5 | 5 |
| 38 | 8 | 8 | 9 | 8 | 8 | 8 | 9 | 9 | 6 | 6 | 5 | 2 | 4 | 6 | 7 | 6 | 4 | 6 | 7 | 4 |
| 41 | 5 | 8 | 7 | 7 | 7 | 9 | 8 | 9 | 4 | 8 | 7 | 6 | 6 | 8 | 8 | 9 | 6 | 6 | 7 | 7 |
| 44 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 3 | 7 | 6 | 6 | 7 | 7 | 6 | 6 | 6 | 6 | 5 |
| 47 | 8 | 7 | 9 | 9 | 9 | 8 | 9 | 8 | 8 | 5 | 7 | 7 | 8 | 7 | 8 | 7 | 8 | 3 | 8 | 6 |
| 50 | 8 | 8 | 7 | 8 | 8 | 8 | 9 | 9 | 4 | 6 | 4 | 8 | 7 | 8 | 7 | 8 | 3 | 4 | 5 | 6 |
| 53 | 8 | 7 | 7 | 7 | 8 | 7 | 7 | 7 | 6 | 5 | 7 | 7 | 8 | 8 | 7 | 8 | 7 | 8 | 8 | 6 |
| 56 | 7 | 7 | 9 | 8 | 8 | 9 | 9 | 7 | 8 | 5 | 7 | 8 | 6 | 8 | 9 | 8 | 3 | 4 | 8 | 5 |

| Speech Sample | 14.6 | | | |
|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 |
| 2 | 1 | 3 | 6 | 8 |
| 5 | 5 | 6 | 6 | 4 |
| 8 | 5 | 5 | 5 | 5 |
| 12 | 4 | 7 | 4 | 5 |
| 16 | 6 | 5 | 6 | 6 |

| | | | | |
|---|---|---|---|---|
| 19 | 7 | 6 | 7 | 7 |
| 22 | 8 | 7 | 7 | 6 |
| 25 | 6 | 8 | 9 | 7 |
| 28 | 5 | 6 | 4 | 4 |
| 32 | 8 | 8 | 5 | 7 |
| 35 | 5 | 5 | 6 | 5 |
| 38 | 5 | 5 | 5 | 4 |
| 41 | 5 | 7 | 9 | 8 |
| 44 | 4 | 7 | 7 | 5 |
| 47 | 8 | 6 | 5 | 7 |
| 50 | 7 | 5 | 6 | 5 |
| 53 | 7 | 8 | 7 | 7 |
| 56 | 5 | 6 | 8 | 5 |

Control Group

| Speech Sample | NK 1.2 | | | | NK 1.3 | | | | NK 1.4 | | | | NK 1.5 | | | | NK 1.6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 7 | 8 | 6 | 5 | 4 | 4 | 4 | 4 | 6 | 5 | 3 | 2 | 2 | 6 | 4 | 5 | 5 | 3 | 1 | 2 |
| 6 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 |
| 10 | 2 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 |
| 14 | 4 | 5 | 4 | 4 | 3 | 5 | 3 | 4 | 5 | 5 | 3 | 4 | 6 | 6 | 5 | 4 | 4 | 4 | 3 | 4 |
| 17 | 6 | 3 | 3 | 1 | 3 | 2 | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 4 | 2 | 3 | 4 | 2 | 1 | 1 |
| 20 | 6 | 5 | 2 | 2 | 2 | 4 | 2 | 4 | 3 | 4 | 2 | 2 | 5 | 4 | 3 | 2 | 3 | 3 | 2 | 3 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 26 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 29 | 3 | 3 | 3 | 2 | 2 | 4 | 2 | 3 | 3 | 4 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 3 | 2 | 2 |
| 33 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 36 | 4 | 6 | 5 | 3 | 4 | 3 | 4 | 2 | 3 | 4 | 6 | 2 | 2 | 3 | 4 | 1 | 3 | 4 | 3 | 3 |
| 39 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 42 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 45 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 48 | 4 | 3 | 2 | 3 | 1 | 2 | 1 | 3 | 2 | 2 | 3 | 3 | 3 | 1 | 2 | 3 | 2 | 2 | 1 | 1 |
| 51 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |
| 54 | 4 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| 57 | 6 | 5 | 6 | 6 | 2 | 5 | 6 | 5 | 4 | 7 | 1 | 5 | 6 | 3 | 8 | 4 | 5 | 4 | 5 | 5 |

| Speech Sample | NK 2.1 | | | | NK 2.2 | | | | NK 4.1 | | | | NK 4.2 | | | | 4.1 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 7 | 7 | 6 | 7 |
| 6 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 4 | 3 | 4 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 4 | 4 | 4 |
| 14 | 3 | 3 | 2 | 2 | 4 | 3 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 4 | 4 | 5 | 5 |
| 17 | 2 | 5 | 2 | 5 | 1 | 5 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 4 | 3 | 2 |
| 20 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 5 | 5 | 5 |
| 23 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |
| 26 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 4 | 1 | 1 | 1 | 1 | 2 | 1 | 3 | 4 | 2 | 1 |
| 29 | 2 | 3 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 1 | 6 | 6 | 5 | 4 |
| 33 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 4 | 5 |
| 36 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 1 | 3 | 2 | 4 | 2 | 1 | 1 | 1 | 1 | 5 | 5 | 5 | 3 |
| 39 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 |
| 42 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 |
| 45 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 3 |
| 48 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 6 | 4 | 3 | 4 |
| 51 | 2 | 2 | 5 | 3 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 5 | 4 | 3 | 2 |
| 54 | 3 | 5 | 7 | 3 | 1 | 2 | 3 | 2 | 4 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 4 | 3 | 2 |
| 57 | 4 | 3 | 2 | 2 | 2 | 3 | 9 | 2 | 2 | 2 | 9 | 1 | 3 | 1 | 1 | 1 | 5 | 7 | 7 | 5 |
| Speech Sample | 4.3 | | | | 5.2 | | | | 6.1 | | | | 6.4 | | | | 6.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 7 | 7 | 5 | 7 | 8 | 8 | 8 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 9 | 8 | 8 | 8 | 9 |
| 6 | 4 | 3 | 2 | 4 | 7 | 6 | 7 | 7 | 4 | 7 | 8 | 7 | 6 | 6 | 7 | 7 | 7 | 7 | 8 | 7 |
| 10 | 5 | 5 | 4 | 5 | 6 | 6 | 7 | 4 | 6 | 4 | 7 | 3 | 7 | 8 | 7 | 4 | 9 | 5 | 7 | 5 |
| 14 | 4 | 4 | 5 | 4 | 7 | 6 | 7 | 6 | 6 | 7 | 6 | 7 | 6 | 7 | 6 | 8 | 8 | 7 | 8 | 8 |
| 17 | 3 | 4 | 3 | 2 | 8 | 8 | 7 | 7 | 7 | 8 | 6 | 8 | 8 | 9 | 8 | 9 | 7 | 7 | 9 | 9 |
| 20 | 2 | 5 | 6 | 4 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 7 | 5 | 6 | 6 | 5 | 5 | 7 | 5 | 4 |
| 23 | 2 | 1 | 1 | 2 | 6 | 8 | 9 | 8 | 4 | 7 | 7 | 9 | 5 | 8 | 9 | 7 | 5 | 7 | 6 | 8 |
| 26 | 4 | 4 | 2 | 1 | 5 | 5 | 5 | 5 | 7 | 5 | 5 | 4 | 7 | 7 | 6 | 5 | 7 | 7 | 7 | 7 |
| 29 | 3 | 4 | 5 | 5 | 5 | 7 | 8 | 7 | 8 | 6 | 5 | 6 | 6 | 7 | 5 | 6 | 6 | 8 | 5 | 7 |
| 33 | 3 | 5 | 3 | 3 | 4 | 5 | 6 | 7 | 4 | 7 | 6 | 7 | 8 | 7 | 8 | 9 | 7 | 9 | 9 | 9 |
| 36 | 5 | 5 | 3 | 2 | 6 | 7 | 6 | 8 | 6 | 9 | 8 | 8 | 7 | 9 | 9 | 7 | 8 | 9 | 9 | 8 |
| 39 | 3 | 1 | 1 | 1 | 9 | 9 | 9 | 9 | 3 | 9 | 9 | 9 | 9 | 9 | 9 | 5 | 8 | 9 | 9 | 9 |
| 42 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 4 | 5 | 5 | 4 | 4 | 5 | 7 | 7 | 6 | 1 | 5 |

| 45 | 2 | 3 | 4 | 3 | 4 | 4 | 2 | 4 | 3 | 3 | 4 | 4 | 5 | 6 | 4 | 5 | 6 | 4 | 3 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 48 | 6 | 3 | 4 | 3 | 9 | 9 | 8 | 9 | 5 | 7 | 7 | 8 | 8 | 6 | 6 | 9 | 8 | 8 | 7 | 7 |
| 51 | 5 | 3 | 4 | 2 | 8 | 6 | 7 | 8 | 7 | 7 | 7 | 9 | 7 | 7 | 8 | 8 | 7 | 7 | 8 | 8 |
| 54 | 3 | 2 | 3 | 2 | 9 | 9 | 8 | 8 | 7 | 8 | 8 | 9 | 9 | 7 | 6 | 9 | 9 | 9 | 6 | 8 |
| 57 | 6 | 7 | 6 | 6 | 6 | 9 | 7 | 9 | 5 | 6 | 7 | 7 | 5 | 8 | 8 | 8 | 8 | 7 | 8 | 7 |
| Speech Sample | 7.2 | | | | 7.3 | | | | 8.1 | | | | 8.2 | | | | 8.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 8 | 9 | 9 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| 6 | 8 | 7 | 7 | 7 | 8 | 8 | 8 | 7 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 7 | 9 | 8 |
| 10 | 5 | 4 | 8 | 6 | 9 | 5 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 |
| 14 | 8 | 7 | 7 | 6 | 9 | 7 | 6 | 8 | 9 | 9 | 7 | 8 | 9 | 9 | 9 | 8 | 7 | 8 | 7 | 7 |
| 17 | 8 | 8 | 6 | 7 | 8 | 9 | 7 | 8 | 8 | 7 | 8 | 8 | 9 | 9 | 8 | 9 | 7 | 8 | 7 | 8 |
| 20 | 8 | 7 | 5 | 5 | 8 | 8 | 4 | 6 | 7 | 9 | 6 | 7 | 8 | 8 | 7 | 9 | 5 | 8 | 6 | 6 |
| 23 | 8 | 4 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 7 | 9 | 9 |
| 26 | 6 | 6 | 6 | 7 | 7 | 6 | 7 | 8 | 8 | 7 | 6 | 8 | 8 | 8 | 9 | 8 | 7 | 7 | 4 | 7 |
| 29 | 7 | 7 | 6 | 6 | 9 | 7 | 6 | 7 | 8 | 9 | 7 | 8 | 9 | 9 | 8 | 8 | 6 | 8 | 7 | 6 |
| 33 | 3 | 5 | 6 | 6 | 7 | 7 | 6 | 6 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 |
| 36 | 8 | 8 | 8 | 7 | 8 | 8 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| 39 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| 42 | 7 | 7 | 8 | 6 | 7 | 8 | 7 | 9 | 6 | 9 | 9 | 9 | 8 | 8 | 9 | 9 | 7 | 8 | 6 | 8 |
| 45 | 5 | 5 | 5 | 4 | 6 | 9 | 6 | 7 | 8 | 9 | 9 | 8 | 8 | 9 | 8 | 9 | 6 | 7 | 8 | 8 |
| 48 | 8 | 7 | 6 | 6 | 9 | 7 | 8 | 8 | 9 | 8 | 8 | 8 | 9 | 9 | 8 | 9 | 8 | 8 | 7 | 7 |
| 51 | 5 | 6 | 7 | 7 | 7 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 8 | 6 | 8 |
| 54 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 6 | 8 | 8 | 8 | 8 | 9 | 8 | 9 | 6 | 8 | 6 | 7 |
| 57 | 7 | 8 | 9 | 9 | 9 | 8 | 9 | 6 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| Speech Sample | 8.4 | | | | 8.5 | | | | 8.6 | | | | 9.2 | | | | 9.3 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 6 | 7 | 7 | 7 | 6 | 4 | 3 | 4 |
| 6 | 9 | 9 | 8 | 9 | 7 | 8 | 9 | 8 | 8 | 8 | 8 | 9 | 3 | 8 | 5 | 6 | 5 | 9 | 6 | 4 |
| 10 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 4 | 3 | 6 | 3 | 4 | 3 | 5 | 2 |
| 14 | 9 | 8 | 9 | 8 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 6 | 6 | 6 | 3 | 4 | 5 | 5 | 4 |
| 17 | 8 | 7 | 9 | 9 | 7 | 7 | 6 | 7 | 8 | 8 | 8 | 9 | 1 | 5 | 2 | 4 | 6 | 6 | 6 | 6 |
| 20 | 8 | 9 | 6 | 7 | 7 | 7 | 5 | 7 | 7 | 9 | 5 | 7 | 2 | 3 | 3 | 4 | 3 | 3 | 2 | 3 |
| 23 | 8 | 9 | 9 | 9 | 6 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 7 | 7 | 7 | 8 | 5 | 9 | 8 |
| 26 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 8 | 9 | 9 | 9 | 5 | 5 | 6 | 4 | 4 | 5 | 6 | 7 |

| 29 | 7 | 8 | 8 | 8 | 8 | 8 | 7 | 7 | 9 | 9 | 8 | 8 | 5 | 7 | 6 | 5 | 6 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 8 | 9 | 9 | 9 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 3 | 5 | 4 | 5 | 2 | 5 | 2 | 5 |
| 36 | 9 | 9 | 9 | 9 | 9 | 9 | 8 | 9 | 9 | 8 | 9 | 9 | 3 | 8 | 7 | 6 | 3 | 7 | 6 | 6 |
| 39 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 5 | 1 | 9 | 5 | 7 | 9 | 9 | 5 |
| 42 | 7 | 9 | 9 | 9 | 6 | 7 | 7 | 8 | 9 | 9 | 9 | 9 | 3 | 2 | 4 | 2 | 3 | 2 | 3 | 2 |
| 45 | 8 | 8 | 8 | 8 | 5 | 7 | 7 | 9 | 8 | 9 | 9 | 8 | 2 | 2 | 1 | 3 | 3 | 3 | 2 | 1 |
| 48 | 8 | 7 | 8 | 8 | 8 | 8 | 7 | 7 | 9 | 8 | 8 | 9 | 7 | 4 | 5 | 5 | 6 | 5 | 3 | 4 |
| 51 | 9 | 9 | 9 | 9 | 8 | 7 | 7 | 6 | 9 | 9 | 9 | 9 | 4 | 2 | 5 | 5 | 5 | 2 | 6 | 4 |
| 54 | 8 | 8 | 9 | 8 | 7 | 8 | 8 | 9 | 9 | 9 | 9 | 8 | 7 | 7 | 7 | 6 | 8 | 6 | 7 | 5 |
| 57 | 9 | 9 | 4 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 5 | 9 | 8 | 7 | 8 | 7 | 4 | 8 | 5 | 7 |
| Speech Sample | 11.1 | | | | 11.2 | | | | 11.3 | | | | 11.4 | | | | 11.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 6 | 3 | 3 | 3 | 1 | 1 | 1 | 3 | 5 | 3 | 2 | 3 |
| 6 | 1 | 2 | 2 | 2 | 1 | 3 | 2 | 3 | 1 | 3 | 2 | 3 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 3 |
| 10 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 |
| 14 | 3 | 4 | 3 | 4 | 6 | 4 | 3 | 3 | 4 | 5 | 3 | 3 | 3 | 5 | 4 | 3 | 5 | 6 | 5 | 5 |
| 17 | 2 | 4 | 5 | 2 | 2 | 3 | 2 | 2 | 1 | 4 | 2 | 2 | 3 | 5 | 4 | 1 | 1 | 3 | 2 | 1 |
| 20 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 |
| 23 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 1 | 2 | 1 | 2 | 2 | 4 | 1 | 2 | 2 |
| 26 | 2 | 3 | 3 | 3 | 4 | 3 | 3 | 2 | 5 | 2 | 3 | 2 | 4 | 4 | 2 | 3 | 4 | 2 | 7 | 3 |
| 29 | 5 | 3 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 4 | 5 | 4 | 3 | 4 |
| 33 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 36 | 2 | 2 | 2 | 1 | 2 | 4 | 3 | 2 | 2 | 6 | 4 | 3 | 5 | 4 | 6 | 3 | 2 | 7 | 4 | 2 |
| 39 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 42 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 |
| 45 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| 48 | 3 | 1 | 2 | 2 | 1 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 1 | 2 | 3 | 2 | 2 | 2 | 3 |
| 51 | 1 | 1 | 3 | 3 | 2 | 2 | 4 | 4 | 3 | 3 | 4 | 2 | 2 | 2 | 4 | 4 | 2 | 2 | 2 | 2 |
| 54 | 1 | 7 | 6 | 3 | 8 | 7 | 4 | 4 | 7 | 6 | 4 | 6 | 2 | 7 | 5 | 4 | 9 | 5 | 4 | 5 |
| 57 | 1 | 2 | 4 | 8 | 6 | 4 | 4 | 6 | 8 | 5 | 8 | 6 | 5 | 6 | 6 | 4 | 6 | 6 | 6 | 5 |
| Speech Sample | 13.1 | | | | 13.2 | | | | 14.1 | | | | 14.4 | | | | 14.5 | | | |
| Judge | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 9 | 9 | 8 | 9 | 9 | 9 | 9 | 9 | 7 | 4 | 3 | 3 | 7 | 5 | 4 | 4 | 2 | 3 | 3 | 3 |
| 6 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 6 | 5 | 6 | 4 | 6 | 6 | 6 | 4 | 4 | 5 | 4 | 5 |
| 10 | 7 | 6 | 6 | 7 | 7 | 7 | 5 | 7 | 3 | 3 | 3 | 2 | 5 | 3 | 7 | 2 | 3 | 3 | 3 | 2 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 8 | 8 | 7 | 7 | 8 | 8 | 8 | 6 | 6 | 6 | 5 | 5 | 7 | 8 | 7 | 6 | 6 | 6 | 5 | 6 |
| 17 | 7 | 8 | 7 | 9 | 7 | 9 | 9 | 9 | 5 | 3 | 2 | 1 | 7 | 6 | 5 | 5 | 4 | 3 | 3 | 3 |
| 20 | 8 | 7 | 8 | 9 | 8 | 8 | 8 | 9 | 3 | 4 | 2 | 5 | 4 | 5 | 5 | 5 | 4 | 3 | 3 | 4 |
| 23 | 4 | 3 | 3 | 3 | 4 | 2 | 6 | 3 | 6 | 5 | 4 | 5 | 6 | 6 | 8 | 6 | 6 | 4 | 6 | 6 |
| 26 | 8 | 8 | 8 | 8 | 8 | 7 | 8 | 9 | 6 | 4 | 4 | 5 | 6 | 5 | 6 | 4 | 4 | 5 | 5 | 5 |
| 29 | 8 | 7 | 7 | 8 | 7 | 9 | 8 | 8 | 7 | 8 | 7 | 6 | 6 | 7 | 7 | 6 | 6 | 6 | 7 | 7 |
| 33 | 7 | 6 | 6 | 7 | 9 | 8 | 7 | 9 | 3 | 7 | 6 | 4 | 3 | 6 | 6 | 6 | 2 | 4 | 5 | 5 |
| 36 | 9 | 4 | 9 | 8 | 9 | 9 | 9 | 8 | 5 | 5 | 6 | 5 | 6 | 7 | 8 | 7 | 6 | 6 | 7 | 6 |
| 39 | 1 | 1 | 1 | 1 | 3 | 1 | 9 | 5 | 5 | 9 | 1 | 5 | 9 | 9 | 9 | 9 | 2 | 9 | 9 | 5 |
| 42 | 7 | 9 | 8 | 8 | 7 | 9 | 9 | 9 | 3 | 4 | 3 | 2 | 4 | 3 | 4 | 3 | 4 | 2 | 2 | 4 |
| 45 | 6 | 8 | 7 | 9 | 5 | 9 | 9 | 9 | 3 | 3 | 3 | 4 | 5 | 5 | 6 | 6 | 3 | 2 | 4 | 5 |
| 48 | 7 | 6 | 6 | 8 | 8 | 6 | 8 | 8 | 6 | 7 | 6 | 6 | 7 | 6 | 7 | 6 | 8 | 7 | 6 | 7 |
| 51 | 4 | 8 | 8 | 9 | 6 | 5 | 8 | 5 | 6 | 6 | 7 | 6 | 7 | 5 | 9 | 6 | 3 | 7 | 7 | 7 |
| 54 | 9 | 9 | 6 | 9 | 7 | 9 | 8 | 7 | 8 | 6 | 7 | 5 | 9 | 7 | 4 | 7 | 6 | 7 | 6 | 7 |
| 57 | 8 | 8 | 7 | 9 | 9 | 9 | 7 | 8 | 7 | 9 | 9 | 8 | 9 | 8 | 1 | 6 | 7 | 7 | 9 | 7 |

| Speech Sample | 14.6 | | | |
|---|---|---|---|---|
| Judge | Occ1 | Occ2 | Occ3 | Occ4 |
| 3 | 5 | 5 | 2 | 3 |
| 6 | 5 | 5 | 4 | 4 |
| 10 | 3 | 5 | 2 | 3 |
| 14 | 5 | 7 | 6 | 5 |
| 17 | 3 | 1 | 1 | 4 |
| 20 | 2 | 5 | 3 | 4 |
| 23 | 6 | 5 | 4 | 5 |
| 26 | 5 | 7 | 6 | 4 |
| 29 | 5 | 6 | 6 | 7 |
| 33 | 4 | 5 | 5 | 6 |
| 36 | 5 | 7 | 7 | 6 |
| 39 | 6 | 9 | 9 | 9 |
| 42 | 4 | 3 | 3 | 2 |
| 45 | 3 | 4 | 4 | 6 |
| 48 | 7 | 6 | 6 | 6 |
| 51 | 6 | 8 | 8 | 7 |
| 54 | 8 | 7 | 6 | 4 |
| 57 | 8 | 9 | 7 | 7 |

APPENDIX AD

STUDY TWO: INTRARATER AGREEMENT OF SPEECH NATURALNESS - THE PERCENTAGE OF RERATINGS WITHIN +/- 1 OF ORIGINAL VALUE.

| Judge | Group | % occasion 1 versus occasion 2 | % occasion 3 versus occasion 4 |
|---|---|---|---|
| 1 | 1 | 83.33 | 77.78 |
| 2 | 2 | 83.33 | 77.78 |
| 3 | 3 | 77.78 | 94.44 |
| 4 | 1 | 75.00 | 91.67 |
| 5 | 2 | 86.11 | 77.78 |
| 6 | 3 | 83.33 | 88.89 |
| 7 | 1 | 80.56 | 94.44 |
| 8 | 2 | 80.56 | 94.44 |
| 10 | 3 | 86.11 | 75.00 |
| 11 | 1 | 75.00 | 69.44 |
| 12 | 2 | 75.00 | 86.11 |
| 14 | 3 | 86.11 | 88.89 |
| 15 | 1 | 75.00 | 91.67 |
| 16 | 2 | 88.89 | 75.00 |
| 17 | 3 | 63.89 | 77.78 |
| 18 | 1 | 75.00 | 83.33 |
| 19 | 2 | 63.89 | 72.22 |
| 20 | 3 | 80.56 | 80.56 |
| 21 | 1 | 88.89 | 91.67 |
| 22 | 2 | 69.44 | 69.44 |
| 23 | 3 | 72.22 | 86.11 |
| 24 | 1 | 72.22 | 80.56 |
| 25 | 2 | 77.78 | 91.67 |
| 26 | 3 | 83.33 | 83.33 |
| 27 | 1 | 61.11 | 50.00 |
| 28 | 2 | 83.33 | 97.22 |
| 29 | 3 | 75.00 | 97.22 |
| 30 | 1 | 86.11 | 77.78 |
| 32 | 2 | 86.11 | 80.56 |
| 33 | 3 | 72.22 | 91.67 |
| 34 | 1 | 83.33 | 91.67 |
| 35 | 2 | 83.33 | 80.56 |

| | | | |
|---|---|---|---|
| 36 | 3 | 72.22 | 72.22 |
| 37 | 1 | 75.00 | 77.78 |
| 38 | 2 | 80.56 | 72.22 |
| 39 | 3 | 72.22 | 83.33 |
| 40 | 1 | 86.11 | 97.22 |
| 41 | 2 | 47.22 | 91.67 |
| 42 | 3 | 83.33 | 80.56 |
| 43 | 1 | 80.56 | 94.44 |
| 44 | 2 | 80.56 | 75.00 |
| 45 | 3 | 86.11 | 83.33 |
| 46 | 1 | 80.56 | 86.11 |
| 47 | 2 | 77.78 | 77.78 |
| 48 | 3 | 69.44 | 91.67 |
| 49 | 1 | 77.78 | 88.89 |
| 50 | 2 | 69.44 | 86.11 |
| 51 | 3 | 77.78 | 77.78 |
| 52 | 1 | 77.78 | 94.44 |
| 53 | 2 | 83.33 | 91.67 |
| 54 | 3 | 63.89 | 72.22 |
| 55 | 1 | 66.67 | 91.67 |
| 56 | 2 | 88.89 | 72.22 |
| 57 | 3 | 69.44 | 52.78 |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group

APPENDIX AE

STUDY TWO: AVERAGE INTRARATER AGREEMENT OF SPEECH NATURALNESS (% OF RERATINGS WITHIN +/- 1 SCALE VALUE OF THE ORIGINAL RATING) FOR ALL 54 RATERS BY STUTTERING SEVERITY LEVEL.

| Judge | Group | Nor x1 | Mil x1 | Mod x 1 | Sev x1 | Nor x2 | Mil x2 | Mod x2 | Sev x2 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 77.78 | 66.67 | 88.89 | 100 | 100 | 44.44 | 66.67 | 100 |
| 2 | 2 | 100 | 100 | 44.44 | 88.89 | 100 | 55.56 | 66.67 | 88.89 |
| 3 | 3 | 66.67 | 66.67 | 77.78 | 100 | 100 | 77.78 | 100 | 100 |
| 4 | 1 | 77.78 | 44.44 | 77.78 | 100 | 100 | 77.78 | 100 | 88.89 |
| 5 | 2 | 88.89 | 55.56 | 100 | 100 | 88.89 | 66.67 | 55.56 | 100 |
| 6 | 3 | 100 | 55.56 | 88.89 | 88.89 | 100 | 77.78 | 77.78 | 100 |
| 7 | 1 | 100 | 66.67 | 77.78 | 77.78 | 77.78 | 100 | 100 | 100 |
| 8 | 2 | 100 | 88.89 | 55.56 | 77.78 | 100 | 100 | 77.78 | 100 |
| 10 | 3 | 100 | 100 | 44.44 | 100 | 100 | 77.78 | 44.44 | 77.78 |
| 11 | 1 | 100 | 77.78 | 33.33 | 88.89 | 77.78 | 66.67 | 55.56 | 77.78 |
| 12 | 2 | 77.78 | 88.89 | 55.56 | 77.78 | 77.78 | 77.78 | 100 | 88.89 |
| 14 | 3 | 88.89 | 77.78 | 77.78 | 100 | 100 | 88.89 | 77.78 | 88.89 |
| 15 | 1 | 55.56 | 66.67 | 88.89 | 88.89 | 88.89 | 100 | 88.89 | 88.89 |
| 16 | 2 | 88.89 | 77.78 | 88.89 | 100 | 77.78 | 55.56 | 66.67 | 100 |
| 17 | 3 | 55.56 | 33.33 | 77.78 | 88.89 | 77.78 | 66.67 | 77.78 | 88.89 |
| 18 | 1 | 88.89 | 44.44 | 77.78 | 88.89 | 66.67 | 88.89 | 88.89 | 88.89 |
| 19 | 2 | 100 | 55.56 | 22.22 | 77.78 | 100 | 55.56 | 44.44 | 88.89 |
| 20 | 3 | 88.89 | 88.89 | 77.78 | 66.67 | 88.89 | 88.89 | 77.78 | 66.67 |
| 21 | 1 | 100 | 66.67 | 100 | 88.89 | 100 | 88.89 | 88.89 | 88.89 |
| 22 | 2 | 66.67 | 55.56 | 55.56 | 100 | 66.67 | 44.44 | 88.89 | 77.78 |
| 23 | 3 | 100 | 77.78 | 44.44 | 66.67 | 100 | 100 | 55.56 | 88.89 |
| 24 | 1 | 100 | 77.78 | 33.33 | 77.78 | 100 | 55.56 | 100 | 66.67 |
| 25 | 2 | 66.67 | 55.56 | 88.89 | 100 | 100 | 88.89 | 77.78 | 100 |
| 26 | 3 | 88.89 | 77.78 | 66.67 | 100 | 100 | 77.78 | 77.78 | 77.78 |
| 27 | 1 | 66.67 | 55.56 | 44.44 | 77.78 | 11.1 | 11.1 | 88.89 | 88.89 |
| 28 | 2 | 100 | 77.78 | 66.67 | 88.89 | 100 | 100 | 88.89 | 100 |
| 29 | 3 | 88.89 | 77.78 | 66.67 | 66.67 | 100 | 100 | 88.89 | 100 |
| 30 | 1 | 88.89 | 88.89 | 77.78 | 88.89 | 77.78 | 66.67 | 88.89 | 77.78 |
| 32 | 2 | 88.89 | 66.67 | 100 | 88.89 | 100 | 66.67 | 77.78 | 77.78 |
| 33 | 3 | 100 | 55.56 | 33.33 | 100 | 100 | 88.89 | 88.89 | 88.89 |
| 34 | 1 | 100 | 66.67 | 88.89 | 77.78 | 100 | 77.78 | 100 | 88.89 |
| 35 | 2 | 88.89 | 77.78 | 66.67 | 100 | 88.89 | 77.78 | 77.78 | 77.78 |
| 36 | 3 | 88.89 | 44.44 | 66.67 | 88.89 | 44.44 | 66.67 | 88.89 | 88.89 |
| 37 | 1 | 100 | 88.89 | 33.33 | 77.78 | 100 | 88.89 | 77.78 | 44.44 |
| 38 | 2 | 88.89 | 88.89 | 77.78 | 66.67 | 88.89 | 55.56 | 66.67 | 77.78 |
| 39 | 3 | 100 | 44.44 | 55.56 | 88.89 | 100 | 77.78 | 66.67 | 88.89 |
| 40 | 1 | 100 | 100 | 55.56 | 88.89 | 100 | 88.89 | 100 | 100 |
| 41 | 2 | 88.89 | 44.44 | 33.33 | 22.22 | 100 | 88.89 | 88.89 | 88.89 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 42 | 3 | 100 | 100 | 77.78 | 55.56 | 100 | 88.89 | 44.44 | 88.89 |
| 43 | 1 | 100 | 77.78 | 55.56 | 88.89 | 100 | 77.78 | 100 | 100 |
| 44 | 2 | 88.89 | 66.67 | 66.67 | 100 | 44.44 | 77.78 | 77.78 | 100 |
| 45 | 3 | 100 | 100 | 77.78 | 66.67 | 100 | 88.89 | 77.78 | 66.67 |
| 46 | 1 | 88.89 | 100 | 33.33 | 100 | 100 | 88.89 | 77.78 | 77.78 |
| 47 | 2 | 77.78 | 66.67 | 66.67 | 100 | 88.89 | 44.44 | 77.78 | 100 |
| 48 | 3 | 77.78 | 44.44 | 66.67 | 88.89 | 88.89 | 100 | 88.89 | 88.89 |
| 49 | 1 | 77.78 | 66.67 | 77.78 | 88.89 | 100 | 100 | 55.56 | 100 |
| 50 | 2 | 88.89 | 66.67 | 55.56 | 66.67 | 100 | 77.78 | 77.78 | 88.89 |
| 51 | 3 | 100 | 66.67 | 66.67 | 77.78 | 88.89 | 66.67 | 77.78 | 77.78 |
| 52 | 1 | 88.89 | 66.67 | 55.56 | 100 | 100 | 100 | 88.89 | 88.89 |
| 53 | 2 | 66.67 | 66.67 | 100 | 100 | 100 | 88.89 | 88.89 | 88.89 |
| 54 | 3 | 77.78 | 44.44 | 66.67 | 66.67 | 88.89 | 66.67 | 44.44 | 88.89 |
| 55 | 1 | 55.56 | 77.78 | 33.33 | 100 | 100 | 88.89 | 77.78 | 100 |
| 56 | 2 | 100 | 77.78 | 77.78 | 100 | 100 | 55.56 | 55.56 | 77.78 |
| 57 | 3 | 55.56 | 55.56 | 77.78 | 88.89 | 55.56 | 33.33 | 66.67 | 55.56 |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group

APPENDIX AF

STUDY TWO: INTERRATER AGREEMENT OF SPEECH NATURALNESS - THE PERCENTAGE OF RATINGS FOR EACH JUDGE WITHIN +/- 1 OF EVERY OTHER JUDGE IN HER GROUP ON EACH OCCASION.

| Judge | Group | Occasion 1 | Occasion 2 | Occasion 3 | Occasion 4 |
|---|---|---|---|---|---|
| 1 | 1 | 71.08 | 73.69 | 71.57 | 69.61 |
| 2 | 2 | 54.58 | 59.15 | 66.18 | 63.56 |
| 3 | 3 | 53.59 | 55.23 | 57.19 | 56.21 |
| 4 | 1 | 72.71 | 62.42 | 73.04 | 79.90 |
| 5 | 2 | 63.07 | 66.99 | 71.90 | 67.81 |
| 6 | 3 | 67.81 | 67.48 | 67.16 | 69.44 |
| 7 | 1 | 66.01 | 71.57 | 67.16 | 72.88 |
| 8 | 2 | 64.54 | 68.63 | 69.44 | 69.77 |
| 10 | 3 | 62.75 | 58.33 | 63.24 | 56.54 |
| 11 | 1 | 63.24 | 64.71 | 69.61 | 62.75 |
| 12 | 2 | 58.33 | 68.63 | 61.60 | 72.22 |
| 14 | 3 | 54.74 | 54.74 | 55.39 | 59.15 |
| 15 | 1 | 64.71 | 70.42 | 72.71 | 76.14 |
| 16 | 2 | 69.28 | 66.67 | 64.22 | 68.14 |
| 17 | 3 | 60.95 | 51.96 | 62.42 | 65.20 |
| 18 | 1 | 57.35 | 66.99 | 74.51 | 72.22 |
| 19 | 2 | 54.08 | 73.20 | 61.44 | 53.76 |
| 20 | 3 | 55.07 | 60.29 | 48.20 | 55.39 |
| 21 | 1 | 71.73 | 71.24 | 76.47 | 71.41 |
| 22 | 2 | 52.29 | 63.40 | 63.24 | 64.22 |
| 23 | 3 | 59.15 | 59.48 | 57.35 | 65.52 |
| 24 | 1 | 62.58 | 57.52 | 62.25 | 70.10 |
| 25 | 2 | 52.94 | 51.63 | 58.50 | 58.33 |
| 26 | 3 | 62.25 | 55.88 | 56.86 | 63.73 |
| 27 | 1 | 67.65 | 69.12 | 57.68 | 67.65 |
| 28 | 2 | 66.50 | 64.22 | 46.41 | 58.17 |
| 29 | 3 | 58.50 | 59.64 | 59.31 | 62.75 |
| 30 | 1 | 69.93 | 70.26 | 68.63 | 68.95 |
| 32 | 2 | 53.76 | 57.68 | 69.77 | 67.97 |
| 33 | 3 | 58.99 | 62.75 | 62.09 | 65.20 |
| 34 | 1 | 64.22 | 67.97 | 76.63 | 69.12 |
| 35 | 2 | 52.94 | 52.94 | 48.37 | 45.26 |
| 36 | 3 | 60.78 | 53.27 | 49.84 | 69.93 |
| 37 | 1 | 56.70 | 59.64 | 69.93 | 49.84 |
| 38 | 2 | 56.21 | 59.80 | 63.73 | 60.95 |
| 39 | 3 | 56.21 | 48.04 | 51.63 | 58.01 |
| 40 | 1 | 57.35 | 65.20 | 73.37 | 72.39 |
| 41 | 2 | 56.21 | 71.24 | 67.81 | 65.20 |

| | | | | | |
|---|---|---|---|---|---|
| 42 | 3 | 55.39 | 56.05 | 58.82 | 62.25 |
| 43 | 1 | 57.35 | 66.18 | 73.86 | 69.61 |
| 44 | 2 | 64.71 | 66.83 | 58.33 | 69.77 |
| 45 | 3 | 56.21 | 58.33 | 58.17 | 64.71 |
| 46 | 1 | 60.78 | 55.72 | 69.28 | 66.50 |
| 47 | 2 | 59.64 | 70.42 | 68.95 | 69.61 |
| 48 | 3 | 62.75 | 63.89 | 66.67 | 66.18 |
| 49 | 1 | 70.26 | 69.77 | 69.12 | 59.15 |
| 50 | 2 | 62.91 | 67.32 | 68.30 | 68.30 |
| 51 | 3 | 63.24 | 64.71 | 59.97 | 65.85 |
| 52 | 1 | 69.28 | 56.54 | 72.06 | 63.24 |
| 53 | 2 | 60.46 | 66.01 | 72.39 | 69.93 |
| 54 | 3 | 49.51 | 55.23 | 54.41 | 63.07 |
| 55 | 1 | 61.44 | 58.50 | 74.02 | 64.71 |
| 56 | 2 | 64.22 | 70.26 | 58.66 | 64.87 |
| 57 | 3 | 48.20 | 48.37 | 35.13 | 47.22 |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group

APPENDIX AG

STUDY TWO: INTERRATER AGREEMENT FOR 54 INDIVIDUAL JUDGES FOR ALL FOUR RATING OCCASIONS BY SPEAKER GROUP

| Judge | Group | Nor1 | Nor2 | Nor3 | Nor4 | Mil1 | Mil2 | Mil3 | Mil4 | Mod1 | Mod2 | Mod3 | Mod4 | Sev1 | Sev2 | Sev3 | Sev4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 77.12 | 84.31 | 74.51 | 70.59 | 63.40 | 61.44 | 62.75 | 65.36 | 62.09 | 67.97 | 67.97 | 66.01 | 81.70 | 81.05 | 81.05 | 76.47 |
| 2 | 2 | 69.28 | 73.86 | 75.82 | 83.66 | 32.03 | 27.45 | 46.41 | 43.79 | 43.79 | 56.86 | 69.94 | 56.86 | 73.20 | 78.43 | 72.55 | 69.94 |
| 3 | 3 | 55.56 | 47.06 | 57.52 | 67.32 | 40.52 | 44.44 | 54.25 | 43.79 | 52.29 | 54.25 | 45.10 | 37.91 | 66.01 | 75.16 | 71.9 | 75.82 |
| 4 | 1 | 81.05 | 56.86 | 65.36 | 75.16 | 54.25 | 41.83 | 57.52 | 77.12 | 67.32 | 66.01 | 84.31 | 78.43 | 88.24 | 84.97 | 84.97 | 88.89 |
| 5 | 2 | 71.90 | 82.35 | 88.89 | 90.85 | 43.79 | 55.56 | 53.60 | 42.48 | 58.17 | 48.37 | 67.97 | 61.44 | 78.43 | 81.70 | 77.12 | 76.47 |
| 6 | 3 | 78.43 | 75.16 | 82.35 | 84.31 | 58.17 | 54.90 | 62.75 | 59.48 | 63.40 | 59.48 | 54.25 | 52.29 | 71.24 | 80.39 | 69.28 | 81.70 |
| 7 | 1 | 71.24 | 82.35 | 58.17 | 67.97 | 56.21 | 44.44 | 54.90 | 71.90 | 58.17 | 74.51 | 79.09 | 78.43 | 78.43 | 84.97 | 76.47 | 73.20 |
| 8 | 2 | 76.47 | 79.08 | 82.35 | 87.58 | 58.17 | 63.40 | 53.60 | 62.09 | 53.59 | 52.29 | 56.86 | 51.63 | 69.93 | 79.74 | 84.97 | 77.78 |
| 10 | 3 | 79.74 | 73.86 | 81.05 | 82.35 | 58.17 | 56.86 | 57.52 | 46.41 | 43.79 | 28.11 | 52.29 | 30.07 | 69.28 | 74.51 | 62.09 | 67.32 |
| 11 | 1 | 73.20 | 81.70 | 69.94 | 70.59 | 51.63 | 54.90 | 69.94 | 66.01 | 54.90 | 49.02 | 62.75 | 44.44 | 73.2 | 73.20 | 75.82 | 69.94 |
| 12 | 2 | 75.16 | 78.43 | 82.35 | 87.58 | 49.67 | 59.48 | 47.71 | 60.13 | 41.83 | 59.48 | 38.56 | 54.90 | 66.67 | 77.12 | 77.78 | 86.28 |
| 14 | 3 | 44.44 | 40.52 | 53.60 | 54.90 | 41.18 | 39.22 | 44.44 | 42.48 | 59.48 | 52.94 | 49.67 | 67.32 | 73.86 | 86.28 | 73.86 | 71.90 |
| 15 | 1 | 66.01 | 83.01 | 69.94 | 75.16 | 46.41 | 51.63 | 75.82 | 81.70 | 62.09 | 67.32 | 77.12 | 78.43 | 84.31 | 79.74 | 67.97 | 69.28 |
| 16 | 2 | 82.35 | 73.20 | 77.78 | 86.28 | 59.48 | 51.63 | 37.91 | 39.87 | 56.21 | 60.13 | 62.09 | 66.01 | 79.08 | 81.70 | 79.08 | 80.39 |
| 17 | 3 | 63.40 | 50.33 | 81.05 | 71.24 | 50.33 | 42.48 | 49.02 | 60.13 | 56.86 | 48.37 | 45.75 | 51.63 | 73.20 | 66.67 | 73.86 | 77.78 |
| 18 | 1 | 60.78 | 83.01 | 76.47 | 75.16 | 44.44 | 52.94 | 61.44 | 71.90 | 48.37 | 50.98 | 84.31 | 75.82 | 75.82 | 81.05 | 75.82 | 66.01 |
| 19 | 2 | 72.55 | 80.39 | 72.55 | 54.25 | 43.79 | 64.71 | 49.02 | 52.94 | 29.41 | 63.40 | 54.90 | 38.56 | 70.59 | 84.31 | 69.28 | 69.28 |
| 20 | 3 | 61.44 | 52.94 | 83.01 | 75.82 | 57.52 | 50.98 | 43.79 | 53.60 | 41.83 | 56.86 | 35.29 | 43.79 | 59.48 | 80.39 | 30.72 | 48.37 |
| 21 | 1 | 82.35 | 83.66 | 69.28 | 64.05 | 59.48 | 57.52 | 71.90 | 73.86 | 60.13 | 67.97 | 73.01 | 80.39 | 84.97 | 75.82 | 81.70 | 67.32 |
| 22 | 2 | 45.75 | 58.17 | 75.82 | 77.78 | 43.14 | 57.52 | 57.52 | 45.75 | 45.10 | 54.90 | 49.02 | 51.63 | 75.16 | 83.01 | 70.59 | 81.70 |
| 23 | 3 | 75.16 | 70.59 | 81.05 | 82.35 | 49.02 | 49.02 | 48.37 | 55.56 | 50.33 | 54.90 | 45.75 | 60.78 | 62.09 | 63.40 | 54.25 | 63.40 |
| 24 | 1 | 85.62 | 87.58 | 67.32 | 73.86 | 37.91 | 39.22 | 35.95 | 59.48 | 60.13 | 45.10 | 75.82 | 76.47 | 66.67 | 58.17 | 69.93 | 70.59 |
| 25 | 2 | 41.83 | 47.06 | 62.09 | 65.36 | 38.56 | 29.41 | 50.33 | 38.56 | 54.90 | 54.25 | 47.71 | 58.82 | 76.47 | 75.82 | 73.86 | 70.59 |
| 26 | 3 | 80.40 | 65.36 | 78.43 | 85.62 | 41.83 | 51.63 | 50.98 | 52.29 | 54.25 | 47.06 | 47.71 | 50.98 | 72.55 | 59.48 | 50.33 | 66.01 |
| 27 | 1 | 66.01 | 84.97 | 61.44 | 66.67 | 53.59 | 52.29 | 35.95 | 63.40 | 68.63 | 62.75 | 54.25 | 66.67 | 82.35 | 76.47 | 79.08 | 73.86 |
| 28 | 2 | 81.05 | 84.31 | 79.09 | 89.54 | 57.52 | 57.52 | 43.79 | 54.25 | 54.90 | 40.52 | 17.65 | 27.45 | 72.55 | 74.51 | 45.10 | 61.44 |
| 29 | 3 | 77.12 | 58.17 | 83.01 | 85.62 | 32.03 | 37.91 | 44.44 | 37.26 | 53.59 | 57.52 | 38.56 | 51.64 | 71.24 | 84.97 | 71.24 | 76.47 |
| 30 | 1 | 75.16 | 84.31 | 65.36 | 77.78 | 56.86 | 45.75 | 60.13 | 69.94 | 65.36 | 66.67 | 69.94 | 61.44 | 82.35 | 84.31 | 79.08 | 66.67 |
| 32 | 2 | 71.90 | 70.59 | 82.35 | 85.62 | 37.25 | 34.64 | 50.98 | 46.41 | 42.48 | 52.29 | 67.32 | 59.48 | 63.40 | 73.20 | 78.43 | 80.39 |
| 33 | 3 | 71.24 | 70.59 | 81.05 | 82.35 | 55.56 | 50.98 | 52.94 | 54.90 | 37.91 | 50.33 | 48.37 | 50.33 | 71.24 | 79.09 | 66.01 | 73.20 |
| 34 | 1 | 83.66 | 88.89 | 75.82 | 78.43 | 47.71 | 51.63 | 58.17 | 54.25 | 59.48 | 56.86 | 85.62 | 66.67 | 66.01 | 74.51 | 86.93 | 77.12 |
| 35 | 2 | 57.52 | 73.86 | 66.01 | 84.31 | 54.25 | 64.05 | 54.90 | 54.25 | 47.71 | 27.45 | 34.64 | 29.41 | 52.29 | 46.41 | 37.91 | 13.07 |
| 36 | 3 | 62.09 | 56.86 | 38.56 | 71.90 | 53.59 | 37.91 | 44.44 | 60.78 | 60.78 | 48.37 | 45.10 | 65.36 | 66.67 | 69.94 | 71.24 | 81.70 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 37 | 1 | 76.47 | 85.62 | 77.78 | 77.12 | 40.52 | 45.75 | 71.24 | 59.48 | 29.41 | 27.45 | 49.02 | 20.26 | 80.39 | 79.74 | 81.70 | 42.48 |
| 38 | 2 | 70.59 | 64.71 | 79.09 | 85.62 | 58.17 | 52.94 | 42.48 | 43.79 | 54.25 | 54.25 | 60.13 | 57.52 | 41.83 | 67.32 | 73.20 | 56.86 |
| 39 | 3 | 71.24 | 67.97 | 81.05 | 82.35 | 54.25 | 37.91 | 37.26 | 50.98 | 44.44 | 24.84 | 27.45 | 36.60 | 54.90 | 61.44 | 60.78 | 62.09 |
| 40 | 1 | 78.43 | 83.66 | 66.01 | 71.24 | 44.44 | 53.60 | 66.67 | 69.28 | 47.71 | 62.75 | 80.39 | 76.47 | 58.82 | 60.78 | 80.39 | 72.55 |
| 41 | 2 | 79.09 | 81.05 | 80.39 | 79.74 | 51.63 | 56.86 | 54.90 | 50.98 | 52.94 | 59.48 | 58.82 | 56.86 | 41.18 | 87.58 | 77.12 | 73.20 |
| 42 | 3 | 72.55 | 70.59 | 81.05 | 82.35 | 55.56 | 45.75 | 52.94 | 53.60 | 41.18 | 34.64 | 36.60 | 33.33 | 52.29 | 73.20 | 64.71 | 79.74 |
| 43 | 1 | 69.28 | 83.66 | 64.71 | 56.86 | 49.67 | 47.71 | 59.48 | 73.20 | 41.83 | 57.52 | 81.05 | 66.67 | 68.63 | 75.82 | 90.20 | 81.70 |
| 44 | 2 | 68.63 | 77.12 | 44.44 | 84.31 | 55.56 | 54.90 | 57.52 | 59.48 | 58.82 | 60.13 | 66.67 | 63.40 | 75.82 | 75.16 | 64.71 | 71.90 |
| 45 | 3 | 76.47 | 69.93 | 81.05 | 82.35 | 55.56 | 51.63 | 50.98 | 52.94 | 35.95 | 35.95 | 30.07 | 44.44 | 56.86 | 75.82 | 70.59 | 79.09 |
| 46 | 1 | 38.56 | 25.49 | 43.14 | 33.33 | 54.90 | 49.02 | 62.75 | 70.59 | 59.48 | 58.82 | 80.39 | 75.16 | 90.20 | 89.54 | 90.85 | 86.93 |
| 47 | 2 | 69.28 | 81.05 | 83.66 | 83.66 | 47.71 | 59.48 | 55.56 | 53.60 | 51.63 | 63.40 | 62.75 | 65.36 | 69.93 | 77.78 | 73.86 | 75.82 |
| 48 | 3 | 73.20 | 71.90 | 78.43 | 67.97 | 47.06 | 57.52 | 61.44 | 60.13 | 52.29 | 58.17 | 52.94 | 58.17 | 78.43 | 67.97 | 73.86 | 78.43 |
| 49 | 1 | 77.78 | 86.93 | 59.48 | 64.05 | 46.41 | 54.90 | 55.56 | 51.63 | 71.90 | 60.13 | 75.82 | 49.67 | 84.97 | 77.12 | 85.62 | 71.24 |
| 50 | 2 | 81.70 | 77.12 | 84.31 | 84.97 | 45.10 | 47.71 | 47.06 | 57.52 | 51.63 | 61.44 | 65.36 | 58.82 | 73.20 | 83.01 | 76.47 | 71.90 |
| 51 | 3 | 77.12 | 73.20 | 72.55 | 80.39 | 59.48 | 56.21 | 47.71 | 54.25 | 52.29 | 56.21 | 47.06 | 57.52 | 64.05 | 73.20 | 72.55 | 71.24 |
| 52 | 1 | 77.78 | 69.28 | 67.32 | 46.41 | 56.86 | 28.11 | 63.40 | 52.94 | 58.17 | 51.63 | 69.28 | 74.51 | 84.31 | 77.12 | 88.24 | 79.09 |
| 53 | 2 | 67.97 | 77.12 | 87.58 | 88.24 | 40.52 | 44.44 | 61.44 | 60.13 | 54.90 | 62.75 | 64.05 | 64.05 | 78.43 | 79.74 | 76.47 | 67.32 |
| 54 | 3 | 68.63 | 64.05 | 63.40 | 80.39 | 34.64 | 26.14 | 35.95 | 39.87 | 34.64 | 47.06 | 41.83 | 54.90 | 60.13 | 83.66 | 76.47 | 77.12 |
| 55 | 1 | 56.21 | 49.02 | 75.82 | 68.63 | 54.25 | 44.44 | 67.97 | 70.59 | 60.13 | 69.28 | 75.82 | 52.29 | 75.16 | 71.24 | 76.47 | 67.32 |
| 56 | 2 | 62.75 | 75.16 | 64.05 | 60.78 | 52.94 | 54.25 | 50.33 | 59.48 | 58.82 | 68.63 | 50.33 | 62.09 | 82.35 | 83.01 | 69.93 | 77.12 |
| 57 | 3 | 45.75 | 41.18 | 30.72 | 50.33 | 31.37 | 29.41 | 20.92 | 20.92 | 45.75 | 47.71 | 36.60 | 40.52 | 69.93 | 75.16 | 52.29 | 77.12 |

Group 1 = Training Group
Group 2 = Exposure Control Group
Group 3 = Control Group
“Nor” = Normal Speaking Children
“Mil” = Children Who Stutter at a Mild Severity Level
“Mod” Children Who Stutter at a Moderate Severity Level
“Sev” Children Who Stutter at a Severe Severity Level
“1” = Rating Occasion 1
“2” = Rating Occasion 2
“3” = Rating Occasion 3
“4” = Rating Occasion 4