9-3-2010

Listener reliability and agreement of a brief intelligibility rating task

Cai Ewing-Buck

Cai Ewing-Buck
Candidate

Department of Speech and Hearing Sciences
Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Amy T. Neel, Ph.D.
Chairperson

Philip Dale, Ph.D.

Phyllis Palmer, Ph.D.

LISTENER RELIABILITY AND AGREEMENT OF A BRIEF INTELLIGIBILITY RATING TASK

BY

CAI EWING-BUCK

B.A., SPEECH AND HEARING SCIENCES.
UNIVERSITY OF NEW MEXICO, 2008

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Speech-Language Pathology

The University of New Mexico
Albuquerque, New Mexico

August, 2010

DEDICATION

I would like to thank my husband and my parents for their exceptional support and love during the completion of my thesis. I truly cannot thank you enough. The Robin and the Little Blue Man live on.

ACKNOWLEDGMENTS

I am greatly thankful to my advisor and thesis chair, Amy T. Neel, Ph.D., CCC-SLP, for continuing to encourage me and guide me throughout the project. Her passion, guidance, and professional style will remain with me as I continue my career.

I also thank my committee members, Phillip Dale, Ph.D. and Phyllis Palmer, Ph.D., CCC-SLP, for their valuable recommendations pertaining to this study and assistance in my professional development.

Gratitude is extended to the University of New Mexico Department of Speech and Hearing Sciences for the funding to pursue this research.

LISTENER RELIABILITY AND AGREEMENT OF A BRIEF INTELLIGIBILITY RATING TASK

BY

CAI EWING-BUCK

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Speech-Language Pathology

The University of New Mexico
Albuquerque, New Mexico

August, 2010

LISTENER RELIABILITY AND AGREEMENT OF A BRIEF INTELLIGIBILITY RATING TASK

by

Cai Ewing-Buck
B.A., Speech and Hearing Sciences, University of New Mexico, 2008
M.S., Speech-Language Pathology, University of New Mexico, 2010

ABSTRACT

Purpose

This study addresses the development of the I-RAVN Test of Speech Intelligibility, an assessment instrument designed to identify which speech components most affect speech intelligibility. The I-RAVN consists of ratings of overall intelligibility, and ratings of four speech components: rate/rhythm/prosody, articulation, voice quality/breath support, and nasality using a rating scale technique adapted from the CAPE-V instrument for voice. This study seeks to establish that listeners can reliably rate overall intelligibility and the four speech components in speakers with dysarthria.

Methods

Twenty-two graduate students listened to recordings from 24 talkers (7 normal, 6 with Parkinson Disease, 11 with oculopharyngeal muscular dystrophy) producing 3 sentences. The listeners rated each talker using the I-RAVN tool, which uses a visual analog scale (100 mm lines) to evaluate the following speech dimensions: overall impression of intelligibility; rate/rhythm/prosody; articulatory precision; voice quality/breath support; and nasality.

Results

To assess intra-rater reliability and agreement, listeners rated sentences from five of the speakers chosen at random a second time, and Pearson product-moment correlations, t-tests, and percent close agreement calculations were performed for all pairs of 22 listeners. Pearson t-tests showed that there were no significant differences between the first and second ratings of the repeated talkers, though percent close agreement calculations demonstrated that nasality, intelligibility, and articulation were more likely to be rated consistently than rate and voice. Overall, intra-rater reliability was high for intelligibility, articulation and voice, and lower for rate and nasality. To assess inter-rater reliability and agreement, Pearson product-moment correlations, factor analysis, intraclass correlation coefficients (ICCs), rater bias one way analyses of variance (ANOVAs), and percent close agreement calculations were performed. The Pearson correlations demonstrated that more than 85% of the ratings were consistent for intelligibility and articulation, and less than 50% for rate. The ICCs showed that listeners had high consistency when rating intelligibility, moderate consistency when rating articulation, voice, and nasality, and lower consistency when rating rate. Inter-rater reliability and agreement across measures were high for intelligibility, somewhat lower for articulation, voice, and nasality, and consistently lower for rate. Overall, good reliability and agreement were noted for intelligibility and articulation, with moderate values for voice quality and nasality. Lower levels of reliability and agreement were obtained for the rate/rhythm/prosody scale on both intra- and inter-rater tests.

Conclusions

Preliminary results indicate adequate inter- and intra-rater reliability and agreement for the I-RAVN Test of Speech Intelligibility for dysarthric speech. Further research will determine if the I-RAVN can be used as an explanatory, streamlined assessment technique to determine treatment targets for individuals with speech intelligibility deficits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

In this study, listeners used a clinically-motivated explanatory tool called the Brief Intelligibility Rating Task (I-RAVN) to rate sentence intelligibility along several speech dimensions: overall intelligibility, rate/rhythm/prosody, articulation, voice quality/breath support, and nasality. The purpose of this study was to determine the intra- and inter-rater reliability and agreement of the explanatory tool for normal speakers and speakers with dysarthria, a group of speech disorders related to neurogenic disorders. Despite limitations (i.e., limited number of talkers per group and mainly mild diagnoses of dysarthria), these measurements provided initial information about the reliability of the I-RAVN explanatory tool.

Chapter 2

Review of Related Literature

Speech intelligibility can be defined as "the degree to which the speaker's intended message is recovered by the listener" (Kent, Weismer, Kent, & Rosenbek, 1989). Decreased intelligibility is a main deficit of dysarthria, a group of disorders that is characterized by difficulty controlling the muscles involved in speech.

Measuring a speaker's intelligibility allows clinicians to appreciate the functional impact of the speaker's communication disorder (De Bodt, Huici, Van de Heyning, 2002). Several methods have been used to measure speech intelligibility. Ratings of overall intelligibility (Most, Weisel, & Lev-Matezky, 1996; Neel, Palmer, Sprouls, & Morrison, 2006; Van Nuffelen, De Bodt, Wuyts, & Van de Heyning, 2009) and calculation of percent of phonemes/words correctly transcribed by listeners (Keintz, Bunton, & Hoit, 2007; Bunton, 2006; Donovan, Kendall, Young, & Rosenbek, 2006; Hustad, 2006; Hustad & Cahill, 2003; Laures & Weismer, 1999) are common. However, these approaches only give an estimate of severity. Tests that are useful for clinicians must provide explanations of speech deficits (Weismer & Martin, 1992) since speakers can have similar overall intelligibility scores but very different perceptual features contributing to their decreased intelligibility (Kent et al., 1989). After a perceptual analysis has been completed and the most deviant areas have been determined, treatment to increase intelligibility can begin, which is the main goal of therapy for many dysarthric speakers (Hustad, 2006).

Several researchers have focused on analyzing articulatory errors. Platt, Andrews, Young, and Quinn (1980), in an attempt to explain speech intelligibility deficits, focused on articulatory errors of speakers with cerebral palsy (CP). Kent, Weismer, Kent, &

Rosenbek (1989) developed a phonetic intelligibility approach using a continuum scaling procedure for talkers with amyotrophic lateral sclerosis (ALS) that they determined was useful in clarifying the most influential components on phonetic intelligibility. However, the focus on articulation leaves out other aspects of speech that are known to affect intelligibility in dysarthric speakers, such as voice quality, hypernasality, and prosody (Chenery, 1998).

Darley, Aronson, and Brown (1969) developed an early explanatory approach to evaluating dysarthria using a set of 38 perceptual features. The 38 features were chosen based on author discussion as well as participant input. The listeners rated the features, ranging from imprecise consonants to excess and equal stress, using a 7-point scale. There is conflicting evidence regarding reliability using this approach. Darley et al. (1969) concluded that reliability was adequate. Bunton et al. (2007), Zeplin and Kent (1996), and Zyski and Weisiger (1987) all determined that the ratings from the Darley et al., (1969) scale did not have adequate reliability. Bunton et al. (2007) found that when average parameter ratings were in the mid-range rather than the extremes, lower reliability was obtained. Zeplin and Kent (1996) found that reliability varied across speech tasks and perceptual features. Zyski and Weisinger (1987) suggested that the reliability ratings from the original study may have shown overinflated numbers due to the presence of a large number of features that were likely to be similarly rated but that did not help differentiate between dysarthria types. Regardless of reliability, this process is time-consuming and may not facilitate treatment planning for clinicians.

In developing a clinically useful scale, a smaller set of perceptual features was selected for this study based on the physiologic approach to dysarthria put forth by

Netsell and Daniel (1979). In the physiologic approach, the contributions of respiration, phonation, articulation, resonance, and prosody to dysarthria are considered. Impairments of respiration are frequently observed in dysarthric speech: altered lung volumes, shorter breath groups, abnormal chest wall movements, and accessory muscle use are seen in some flaccid dysarthrias affecting spinal nerves (e.g., ALS); reduced lung volumes and chest wall movements, reduced breath groups, and reduced intraoral pressures are found in hypokinetic dysarthrias (e.g., Parkinson Disease); and hyperkinetic dysarthrias (e.g., Huntington Disease) are associated with interruptions in breath support (Duffy, 2005). Impairments of phonation are seen in various dysarthrias as well: flaccid dysarthrias affecting the Vagus Nerve (CN X) can result in dysphonia; Parkinson Disease (PD), a hypokinetic dysarthria, is associated with deficits in intensity, monopitch and monoloudness; and spastic dysarthrias, such as primary lateral sclerosis (PLS), are associated with a strained-strangled voice quality (Duffy, 2005). Articulation impairments are commonly found in speakers with dysarthria: irregular articulatory breakdowns, distorted vowels, and prolonged phonemes are found with ataxic dysarthrias (e.g., multiple sclerosis); repeated phonemes and morphemes are found in speakers with hypokinetic dysarthrias; and imprecise consonants are noted in flaccid, spastic and hypokinetic dysarthrias (Duffy, 2005). Resonance impairments are also observed with many dysarthric speakers: hypokinetic dysarthria and some flaccid dysarthrias associated with hypernasality; (Duffy, 2005). Rate, rhythm, and prosody are also affected in dysarthric speech: reduced speech rates, reduced movement rates, and altered stress patterns, are found in speakers with spastic dysarthria; and reduced rate, inconsistency of rate and prosody, and inconsistency of pitch characterize ataxic dysarthric speech (Duffy,

2005). It is crucial to consider each of these speech dimensions with regard to how they might affect intelligibility of dysarthric speech.

In the current study, we adapted an assessment approach from the field of voice disorders, the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). The CAPE-V is an explanatory tool used to show what aspect(s) of voice would benefit from therapy using visual analog scales (100mm lines) to rate several dimensions of disordered voices (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009). For the I-RAVN, our assessment approach used visual analog scales in the form of 100mm lines to separately rate five perceptual dimensions: overall intelligibility, rate/rhythm/prosody, articulation, voice quality/breath support, and nasality.

This study focuses on determining the intra- and inter-rater reliability and agreement for the I-RAVN instrument. Reliability is often described as the consistency of a measurement (Schiavetti & Metz, 2006; Uebersax, 2010), determining whether or not listeners consistently assign the same meaning to the various scale values (Chenery, 1998). Agreement is another measure of listener consistency, determining if the listeners use similar values to rate the talkers (Chenery, 1998). As in the field of voice disorders, there is no agreed upon method for determining reliability (Kreiman, Kempster, Erman, & Berks, 1993), so we used several techniques to measure reliability and agreement in the current study.

For intra-rater reliability, or the consistency of each listener's ratings, we compared each listeners' first ratings to their second ratings of five talkers. We calculated Pearson correlations, t-tests, and percent close agreement to determine whether the listeners rated the talkers in a similar fashion both times. Pearson correlations were

calculated to determine the relationship between the listeners' first and second ratings. T-test calculations were performed to determine if there were significant differences between the listeners' first and second ratings. Percent close agreement showed how often a listener's first and second ratings fell within 10 scale values of each other.

For inter-rater reliability, or the consistency of ratings between listeners, we compared the ratings of each listener to the ratings of all other listeners. We calculated Pearson correlations, ICCs, factor analyses, percent close agreement, and rater bias to determine if the listeners rated the talkers in a similar fashion to the other listeners. Rater-to-rater Pearson correlations were performed to determine how well one listener's ratings agree with every other listener's ratings. Rater-to-group Pearson correlations were completed to determine the likelihood of one listener's ratings agreeing with the group mean. ICCs were obtained in order to establish the average agreement between listeners as a view of the overall unity of the group. Factor analysis was completed to determine the amount of variability between ratings that could be accounted for by forcing all of the ratings to act in a similar fashion (i.e., to determine if the variability seen in the ratings could be accounted for by one potentially unobserved variable). Percent close agreement calculations were performed in order to determine if the listeners used similar ratings as each other (i.e., fell within 10 scale values of one another). Rater bias calculations were completed using ANOVAs and Tukey HSD post-hoc tests to determine if specific listeners behaved significantly differently than the others. Once reliability has been established, further calculations and research can be completed to determine how each perceptual feature relates to overall intelligibility for speakers with similar disorders.

Chapter 3

Methodology

Participants

This study was reviewed and approved by the Human Subjects Committee of the Institutional Review Board at the University of New Mexico. Twenty-two graduate students from the Department of Speech and Hearing Sciences at the University of New Mexico with normal hearing and no history of speech or language problems served as volunteer listeners in the study. The decision to use graduate students was based on results from two studies completed by Bunton et al. (2007) and Van der Graff et al. (2009). The studies concluded that there is no significant difference in perceptual judgments between experienced judges (i.e. clinicians with five years of experience with dysarthric speakers) and inexperienced judges (i.e. graduate students with limited experience with dysarthric speech (Bunton et al., 2007; Van der Graff et al., 2009). Listeners were paid for their participation. Table 1 shows a description of the listeners.

Table 1. Description of listeners.

Participant	Age	Gender
L1	23	F
L2	55	F
L3	31	F
L4	35	F
L5	34	F
L6	29	F
L7	37	M
L8	35	F
L9	24	F
L10	30	F
L11	25	F

Table 1 (cont.)

L12	46	F
L13	23	F
L14	24	F
L15	39	F
L16	29	F
L17	35	F
L18	30	F
L19	35	F
L20	38	F
L21	24	F
L22	38	F

Speech samples

The current study used speech samples from previous studies (Neel, 2009; Neel, Palmer, Sprouls, & Morrison, 2006). Recordings of 7 normal speakers, 11 speakers with OPMD, and 6 speakers with PD were used. Each of the speakers read three sentences derived from Weismer & Laures (2002): 1) "Bob fell down and hurt his right leg"; 2) "Guide them to where trees and plants grow"; and 3) "Dues can be paid each night this week". The talkers were recorded in a quiet room with a Shure SM 10-A head-mounted microphone positioned about 4 to 5 cm from the corner of the mouth connected to a Marantz PMD670 digital recorder for the normal and OPMD talkers, and an HHb Portadisk Pro MDP500 minidisk recorder for the PD talkers. The sentences were read in the habitual speech mode, with the talkers being instructed to produce the sentences in their everyday voice without extra effort or volume. Table 2 shows a description of the talkers.

Table 2. Description of talkers.

Participant	Age	Gender	Years Since Dx or Onset
C2	67	F	n/a
C3	61	F	n/a
C4	76	M	n/a
C5	56	M	n/a
C6	52	M	n/a
C7	61	F	n/a
C8	58	M	n/a
O1	63	F	3
O2	62	F	4
O3	61	M	10
O5	57	F	0.3
O6	66	F	10
O7	67	F	0.5
O8	59	M	4
O9	67	F	2.5
O10	57	F	0.5
O11	73	F	10
O12	50	M	1
P1	72	M	7
P2	73	F	23
P3	76	M	5
P4	76	M	2
P5	86	M	8
P6	54	M	5

Procedure

The listeners were provided with written instructions to rate each talker on overall intelligibility, rate/rhythm/prosody, articulation, voice quality/breath support, and nasality. Descriptions of each perceptual feature were provided within the written instructions. The listeners were instructed to make a small vertical mark along the gray horizontal line (the visual analog scale in the form of a 100 mm line), near "NO" for normal at 0mm if the aspect of speech was normal, near "MI" for mildly deviant at 33mm

if the aspect of speech was mildly abnormal, near "MO" for moderately deviant at 67mm if the aspect of speech was moderately abnormal, and "SE" for severely deviant at 100mm if the speech was severely abnormal. The sentences were presented using Alvin experiment-control software (Gayvert & Hillenbrand, 2003) on a Dell laptop computer with a Creative Extigy external sound card. The speech samples from the 24 talkers were presented in random order, with each talker appearing in a separate block. Five of the talkers (2 normal talkers, 2 talkers with OPMD, and 1 talker with PD), who were chosen in a quasirandom fashion to represent the three types of talkers in the study, were repeated randomly throughout the other 24 speech samples. Listeners were seated in a quiet room and heard the stimuli at a comfortable level through Sennheiser HD 580 headphones. The listeners were allowed to play each speech sample up to 10 times. The listening task lasted about 1 hour. After each listener had completed the listening task, we used a ruler to determine the distance (in mm) on the visual analog scale, and those measurements were transposed as ratings with a range of 1 to 100. The written instructions and visual analog scale can be found in Appendices A and B.

Analysis Techniques

In the current study, there were two reasons for measuring intra- and inter-rater reliability. The first reason is that obtaining intra- and inter-rater reliability calculations can serve to estimate the validity of the I-RAVN rating scale, since there is no gold standard for auditory-perceptual ratings. This reasoning is based on the assumption that if two ratings do not agree, then at least one of them must be incorrect (Uebersax, 2010). The second reason for obtaining intra- and inter-rater reliability calculations is to determine the consistency of listeners' ratings (Uebersax, 2010). An estimate of the precision of a measurement can be obtained through calculations of its stability and

consistency (Schiavetti & Metz, 2006). Various calculations, including Pearson correlations and percent close agreement values, were performed in order to determine if the listeners ranked the talkers in the same order and if they assigned similar values to the stimuli produced by each talker.

In this study, intra-rater reliability calculations were performed based on ratings of five talkers who were chosen in a quasirandom fashion in order to represent the three types of the talkers (control, OPMD, and PD) in the study. These calculations provided substantial information regarding the validity and consistency of listeners' ratings, as the listeners were not told that any speakers were repeated and were thus blindly performing the second ratings for each of the five repeated talkers.

Pearson correlations were performed between each listener's first and second sets of ratings for each of the five talkers to determine if the two sets were similar. A Pearson value of 0.00 shows that the variables are not related, and a value of 1.00 shows that the variables are perfectly related; Pearson correlations of 0.60 and higher can be considered adequate in the early stages of research (Shiavetti & Metz, 2006).

T-tests were performed in order to determine if there were significant differences between the first and second ratings of each of the five repeated talkers. This was done by measuring the difference between group means of the listeners' first and second scale value rating differences.

The visual analog scale used in this study ranges from 0 to 100, so the likelihood of obtaining exact agreement between two ratings, even by the same listener, was very low. Thus, in order to determine if a listener's ratings were similar for the first and second listens of the five repeated talkers, close instead of exact agreement values were

calculated. Each of the 22 listeners' first and second ratings of the five repeated talkers was determined to be within 10 scale values of the other (close agreement) or beyond 10 scale values of the other (not close agreement). By chance, agreement within 10 scale values would be expected on 28% of rating occasions (Kreiman et al., 1993). In this study, a percentage of close agreement greater than 70% was considered to be high.

Inter-rater reliability was measured using six different methods: rater-to-rater Pearson correlations, rater-to-group Pearson correlations, factor analysis, ICCs, rater bias, and percent close agreement.

Rater-to-rater correlations were calculated by comparing each of the 22 listeners' ratings with each of the other listeners' ratings using Pearson product-moment correlations. This was competed to determine how well one listener's rank order of talkers agreed with another's. In early stages of research, such as this study, Pearson correlations can be considered adequate when they are above .60 (Schiavetti & Metz, 2006).

Rater-to-group correlations were calculated by comparing each of the 22 listeners' ratings with the group mean. As with the rater-to-rater correlations, the Pearson correlations discussed here were considered to be high when they were greater than .60 (Schiavetti & Metz, 2006).

Factor analysis was used as another way to construct a norm for each talker, forcing one factor in order to account for some error, and determining what percent of the variance this one factor accounted for. The calculation was performed using principal axis factoring. One latent factor was extracted for each perceptual feature, meaning that the ratings for each perceptual feature were compared to the mean when error was

accounted for. In this study, percent of variance values greater than .70 were considered to be high. We have not seen this technique used in the speech literature, though it was recommended by Uebersax (2010) in order to assess interrater reliability.

ICCs were calculated in SPSS Version 15.0 (SPSS, Inc., Chicago, IL) as another way to assess interrater reliability. ICCs calculate the ratio of variance associated with the rated perceptual features over the sum of the variance plus the error variance (Sheard, Adams, & Davis, 1991). ICCs are the most generalizable measure of interrater reliability (Sheard et al., 1991). The 24 ratings produced by each of the 22 listeners for each of the 5 perceptual features (i.e., each of the 22 listeners rated each of the 24 talkers on each perceptual feature, so 24 ratings of intelligibility for each listener) were submitted to a two-way mixed effects ANOVA to determine consistency of ratings among listeners. Typically, ICC coefficients above.70 are considered to represent good levels of reliability (Sheard et al., 1991), though it has also been proposed that coefficients at this level are inappropriately high and that coefficients as low as .50 or .60 may be adequate (Mitchell, 1979).

Rater bias is a measure of a listener's ratings across all talkers compared to those of all other listeners. This is a way of determining if certain raters performed substantially differently than other raters. The rater bias values were calculated by using a two-way ANOVA with Tukey HSD post-hoc tests. There were 231 pairwise comparisons performed for each perceptual feature. As such, the percentage of pairs that had significant differences could be calculated. In this study, values lower than 10% were considered to represent low levels of bias.

The frequency of close agreement ratings between listeners for all talkers was calculated in order to further examine inter-rater reliability. It would be unlikely to obtain exact agreement for a scale ranging from 0 to 100, such as is used in the current study. In this study, close agreement was defined as a difference of equal to or less than 10. Calculations were performed by comparing each listener's scale value ratings of a perceptual feature for a single talker to every other listener's scale value ratings of that feature in a pairwise manner. This calculation was performed for all 24 talkers. For example, Listener 1's intelligibility rating of Talker PD4 was determined to be within or beyond scale values of the intelligibility ratings for Talker PD4 from Listeners 2 through 22. These values were then collapsed across all talkers and all listeners for each perceptual feature. In this study, percentages of close agreement above 70% were considered to be high.

Chapter 4

Results

Reliability of Line Measurements

The first investigator measured each listener's markings on the visual analog scales with a ruler to the nearest mm. To measure intra-judge reliability, the first investigator measured 10% of the markings a second time. To measure inter-judge reliability, the second investigator measured 10% of the markings. Pearson correlation calculations were performed to determine the consistency of intra-judge and inter-judge measurements. The Pearson correlation for intra-judge reliability was .99 ($p<.01$) and for inter-judge reliability was .96 ($p<.01$). From these values, it can be seen that the visual analog scale measurements were reliable.

Intra-rater Reliability

Pearson Correlations.

In the current study, Pearson correlations (Table 3) were calculated to compare the first listen ratings to the second listen ratings in order to determine the consistency within listeners. The Pearson correlations (see Table 3) in the current study ranged from high (for the intelligibility and articulation scales) to low (for nasality) across the perceptual features. The mean for intelligibility was .742 ($p<.01$; range = .552-.911). The mean for rate was .652 ($p<.01$ except Talker O8 with $p=.031$; range = .460-.762). The mean for articulation was .587 ($p<.01$ except Talker O8 with $p=.181$; range = .296-.762). The mean for voice was .703 ($p<.01$ except Talker O8 with $p=.291$; range = .236-.943). The mean for nasality was .479 ($p<.01$ only for talkers C2 and P5; range = .138-.875).

Consistently low Pearson correlation values were found for Talker O8, indicating poor agreement within listeners. Values for the other talkers were low to high.

Table 3. Intra-rater Pearson correlations

Pearson Correlations	Intelligibility	Rate	Articulation	Voice	Nasality
Talker C2	.911 (p=.000)	.762 (p=.000)	.762 (p=.000)	.844 (p=.000)	.686 (p=.000)
Talker C5	.718 (p=.000)	.741 (p=.000)	.697 (p=.000)	.664 (p=.001)	.260 (p=.243)
Talker O3	.708 (p=.000)	.733 (p=.000)	.709 (p=.000)	.943 (p=.000)	.433 (p=.044)
Talker O8	.552 (p=.008)	.460 (p=.031)	.296 (p=.181)	.236 (p=.291)	.138 (p=.539)
Talker P5	.821 (p=.000)	.567 (p=.006)	.470 (p=.027)	.831 (p=.000)	.875 (p=.000)
Mean across all talkers	.742	.652	.586	.703	.479
Range (min - max)	.552 to .911	.460 to .762	.296 to .709	.236 to .943	.138 to .875

T-Tests.

T-tests (Table 4) were calculated to determine if there were significant differences between the first and second ratings of the repeated talkers. No significant differences ($p<.01$) were found. Thus, it can be inferred that the listeners were consistent between their first and second ratings.

Table 4. Intra-rater t-tests

T-Test	Intelligibility	Rate	Articulation	Voice	Nasality
Significance ≤0.01	0 pairs	0 pairs	0 pairs	0 pairs	0 pairs

Percent Close Agreement.

Percent close agreement (within 10 out of 100 scale values) was calculated to determine if the listeners used similar values for their first and second ratings of the

repeated talkers (see Table 5). Close agreement was found for 75.45% of intelligibility ratings, 55.45% of the rate ratings, 74.55% of the articulation ratings, 67.27% of the voice ratings, and 86.36% of the nasality ratings. From these values, it can be seen that intra-rater reliability varied by talker, and that nasality, intelligibility, and articulation were more likely to be rated consistently than rate and voice.

Table 5. Intra-rater percent close agreement

Percent Close Agreement	Intelligibility	Rate	Articulation	Voice	Nasality	Overall
Talker C2	86.36	77.27	90.91	68.18	100	84.85
Talker C5	59.09	45.45	72.73	68.18	90.91	67.42
Talker O3	77.27	63.64	86.36	77.27	75.00	75.00
Talker O8	81.82	36.36	86.36	59.09	81.82	70.45
Talker P5	72.73	54.55	59.09	54.55	81.82	72.73
Mean % close agreement across all 5 Talkers	75.45	55.45	74.55	67.27	86.36	72.73

In this study, gross disagreement (Table 6) was defined as greater than 30 scale values of difference (Bunton et al., 2007). Gross disagreement was found for 3.64% of the intelligibility ratings, 8.18% of the rate ratings, 9.09% of the articulation ratings, 6.36% of the voice ratings, and 3.64% of the nasality ratings. Overall, 6.18% of the ratings grossly disagreed. As stated by Bunton et al., (2007), gross disagreements do not have any clinical use. The percentages of gross disagreement found in this study are low, and thus nearly all of the ratings in this study are clinically useful.

Table 6. Intra-rater percent gross disagreement

Percent Gross Disagreement	Intelligibility	Rate	Articulation	Voice	Nasality
Across all 5 Talkers	3.64	8.18	9.09	6.36	3.64

Inter-rater Reliability

Pearson correlations.

Rater-to-rater correlation calculations (Table 7) were completed to determine the relationship between listeners' ratings. The mean rater-to-rater correlation for intelligibility was r = 0.739 (range = .265 to .972); for rate the mean was .460 (range = .006 to .827); for articulation, the mean was .736 (range = .428 to .987); for voice, the mean was .630 (range = -.220 to .932); and for nasality, the mean was .578 (range = -.215 to .937). For intelligibility and articulation, more than 90% of listener pairs were significantly correlated with one another. For voice and nasality, more than half of the listener pairs were significantly correlated. For rate, however, only 42% of listener pairs were significantly correlated.

Table 7. Rater-to-rater Pearson correlations

Mean Correlation	Intelligibility	Rate	Articulation	Voice	Nasality
Across all pairs	0.739	0.460	0.736	0.630	0.578
Range (min - max)	.265 to .972	-.006 to .827	.428 to .987	-.220 to .932	-.215 to .937
Significance p<.01	220 of 231 pairs (95.24%)	96 of 231 pairs (41.56%)	223 of 231 pairs (96.54%)	185 of 231 pairs (80.09%)	157 of 231 pairs (67.97%)

Rater-to-group correlations (Table 8) were completed to determine the relationship between listeners' ratings and the group average. The mean rater-to-group correlation for intelligibility was .889 (range = .590 to .962); for rate, the mean was .313 (range = -.062 to .577); for articulation, the mean was .864 (range = .743 to .961); for voice, the mean was .803 (range = .497 to .960); and for nasality, the mean was .771 (range = .024 to .881). For intelligibility, articulation, voice, and nasality, high

correlations between listeners and the group mean were obtained. But for rate, few of the listeners' ratings significantly correlated with the group mean.

Table 8. Rater-to-group Pearson correlations

Mean Correlation	Intelligibility	Rate	Articulation	Voice	Nasality
Across all listeners	0.889	0.313	0.864	0.803	0.771
Range (min - max)	.590 to .962	-.062 to .577	.743 to .961	.497 to .960	.024 to .881
Significance <.01	22 of 22 pairs (100%)	2 of 22 pairs (9.09%)	22 of 22 pairs (100%)	21 of 22 pairs (95.45%)	21 of 22 pairs (95.45%)

Factor analysis.

Factor analyses (Table 9) were completed to determine how much variance was accounted for by forcing the ratings to have one latent factor. High percentages of variance were accounted for with both intelligibility and articulation. The significance values for all of the perceptual features were low, however, ranging from .082 for intelligibility and .116 for articulation to .218 for nasality. Loadings for each perceptual feature on the latent factor were largely equivalent to the rater-to-group Pearson correlation values with the exception of better performance for rate in the factor analysis loadings. Rate still had the worst performance in the factor analysis of all the five variables.

Table 9. Inter-rater factor analysis

Factor Analysis	Intelligibility	Rate	Articulation	Voice	Nasality
Percent of variance	74.59	50.01	75.24	65.97	63.98
Mean loading on latent factor for all listeners	0.86	0.696	0.804	0.804	0.771
Range (min - max)	.584 to .968	.360 to .887	.494 to .966	.494 to .966	.153 to .984

Intraclass coefficients.

ICC coefficients (Table 10) were calculated to determine the consistency of the entire group of listeners by means of calculating the average agreement between listeners (Kreiman et al., 1993). The ICC coefficient for intelligibility was 0.723, for rate 0.445, for articulation 0.581, for voice 0.586, and for nasality 0.581. All values were significant at the 0.05 level. The group of listeners had good reliability when rating intelligibility (Kreiman et al., 1993). The group of listeners had moderate reliability when rating articulation, voice, and nasality (Kreiman et al., 1993).

Table 10. Inter-rater intraclass correlation coefficients

ICC (2, 1)	Intelligibility	Rate	Articulation	Voice	Nasality
Across all listeners ($p<.01$)	0.723	0.441	0.696	0.586	0.581

Rater bias.

Rater bias calculations (Table 11), which were performed to determine if specific listeners performed significantly differently from others, were completed using one-way analyses of variance (ANOVAs) and Tukey HSD post-hoc tests. The ANOVA f-tests

showed that there were significant differences ($p<.01$) found between listener ratings for each of the five perceptual features. The Tukey HSD post-hoc tests were used to show which listeners differed significantly from others: 6.93% of the pairwise comparisons were significantly different for intelligibility, 13.86% for rate, 1.30% for articulation, 9.96% for voice, and 0.00% for nasality. There were specific listeners who differed significantly from others across the five perceptual features: 23.81% (25 out of 105 chances for agreement) of L2's ratings, 14.29% (15 out of 105) of L8's ratings, 20.00% (21 out of 105) of L15's ratings, 12.38% (13 out of 105) of L20's ratings, and 10.48% (11 out of 105) of L21's ratings were significantly different from the other listeners'. These five listeners accounted for 59.03% of the total variance. Overall, intelligibility, articulation, voice, and nasality had low levels (<10%) of rater bias.

Table 11. Inter-rater bias

Rater Bias	Intelligibility	Rate	Articulation	Voice	Nasality
ANOVA					
F-test ($p<.01$)	$F(1,21) = 4.09$	$F(1,21) = 5.70$	$F(1,21) = 2.19$	$F(1, 21) = 4.97$	$F(1, 21) = 1.82$ ($p=.014$)
Tukey HSD					
Percent of total chances with significant differences ($p<.05$)	6.93% (16 of 231 pairs)	13.86% (32 of 231 pairs)	1.30% (3 of 231 pairs)	9.96% (23 of 231 pairs)	0.00% (0 of 231 pairs)
Listeners with >4 pairwise differences	L8 (8 pairs), L15 (6 pairs)	L2 (9 pairs), L15 (12 pairs), L20 (7 pairs)		L2 (13 pairs)	

Percent close agreement.

Percent close agreement (Table 12) between listeners was calculated in order to determine if the listeners rated the talkers at similar levels. For intelligibility, 53.97% of the listeners' ratings were within 10 scale values of each other; for rate, 42.45% were in close agreement; for articulation, 61.03% were in close agreement; for voice, 50.67% were in close agreement; and for nasality, 70.25% were in close agreement. The levels of agreement seen here are much higher than the 28% chance level of agreement expected for a 10-point scale (i.e. close agreement in the current study was defined as within 10 scale values, so our 100 point scale was adjusted to be a 10-point scale for this calculation) (Kreiman et al., 1993). High levels of close agreement were obtained with intelligibility, articulation, and nasality. Rate, again, had lower levels of agreement than the other four variables.

Table 12. Inter-rater percent close agreement

Percent Close Agreement	Intelligibility	Rate	Articulation	Voice	Nasality
Across All Listeners	53.97	42.45	61.03	50.67	70.25
Mean Difference	5.91	4.65	6.68	5.55	7.69
Standard Deviation	4.95	4.45	5.54	5.04	6.02

Gross disagreement (Table 13), defined as greater than 30 scale values of difference (Bunton et al., 2007), was found for 16.68% of the intelligibility ratings, 26.36% of the rate ratings, 15.36% of the articulation ratings, 22.93% of the voice ratings, and 14.76% of the nasality ratings. Gross disagreement was found for 19.22% of the ratings for all perceptual features, similar to levels reported by Bunton et al. (2007).

Table 13. Inter-rater percent gross disagreement

Percent Gross Disagreement	Intelligibility	Rate	Articulation	Voice	Nasality
Across all talkers	16.68	26.36	15.36	22.93	14.76

Summary

Analysis of intra-rater reliability revealed high consistency for intelligibility, articulation, and nasality, and lower consistency for rate and voice rating scales Intelligibility, articulation, and nasality were all found to have high percentages of close agreement ratings and no significant differences between the first and second ratings. Voice was found to have a moderate percentage of close agreement ratings and no significant difference between the first and second ratings. Rate was found to have a low percentage of close agreement ratings and one talker (Talker O8) who received significantly different first and second ratings. Analysis of intra-rater reliability also revealed that one talker (Talker O8) was found to have poor consistency of ratings.

Analysis of inter-rater reliability revealed lower levels of consistency than were found for intra-rater reliability. Intelligibility was found to have a moderate level of close agreement between first and second ratings, a high level of listener-to-listener correlation, a high level of listener-to-group correlation, a high level of comparable variance between listeners, a moderate level of variance accounted for with one latent factor, and a low level of rater bias. Rate was found to have a low level of close agreement between first and second ratings, a low level of listener-to-listener correlation, a low level of listener-to-group correlation, a low level of comparable variance between listeners, a moderate level of variance accounted for with one latent factor, and a moderate level of rater bias.

Articulation was found to have a moderate level of close agreement between first and second ratings, a high level of listener-to-listener correlation, a high level of listener-to-group correlation, a moderate level of comparable variance between listeners, a high level of variance accounted for with one latent factor, and a low level of rater bias. Voice was found to have a low level of close agreement between first and second ratings, a high level of listener-to-listener correlation, a high level of listener-to-group correlation, a moderate level of comparable variance between listeners, a moderate level of variance accounted for with one latent factor, and a low level of rater bias. Nasality was found to have a high level of close agreement between first and second ratings, a low level of listener-to-listener correlation, a high level of listener-to-group correlation, moderate level of comparable variance between listeners, a moderate level of variance accounted for with one latent factor, and a low level of rater bias.

Overall, good reliability and agreement were noted for intelligibility and articulation with moderate values for voice quality and nasality. Relatively poor reliability and agreement were obtained for the rate/rhythm/prosody scale on both intra-rater and inter-rater tests.

Chapter 5

Discussion

Summary

The purpose of the study was to determine intra- and inter-rater reliability and agreement for the I-RAVN explanatory tool. Overall, intra-rater reliability and agreement was high for intelligibility, fairly high for articulation and voice, and somewhat lower for rate and nasality. Inter-rater reliability and agreement was high for intelligibility and articulation as well. Nasality and voice had moderate inter-rater reliability and agreement, and rate was less reliable and had less agreement.

Results Compared with Previous Studies

These findings are in general agreement with findings from previous studies. Reasonable levels of agreement were found by Bunton et al. (2007) for the 38-feature analysis, and reasonable levels of reliability and agreement were found in this study. Though Bunton et al. (2007) determined that there were no significant differences in agreement between the perceptual features, this study found that intelligibility, articulation, voice, and nasality were rated more consistently than rate. This finding is in accordance with findings by Kreiman et al. (1993), Sheard et al. (1991), and Zeplin and Kent (1996), in which agreement and reliability levels differed across features.

Lower Reliability and Agreement for Talker O8

Analysis revealed that intra-rater reliability and agreement levels were low for one talker in particular, Talker O8. Talker O8 had mild to moderate dysarthria. It is common for listeners to have higher levels of reliability and agreement when rating normal and severe attributes, and lower levels when rating mild to moderate attributes. Though the majority of talkers in the study fell in the mild to moderate range, it is

possible that this trend was only exhibited with Talker O8, since ratings from only five talkers were examined for intra-rater reliability. The within listener ratings disagreed especially for rate and voice for Talker O8. Although his dysarthria was not judged to be particularly severe (mean intelligibility rating = 10.125, where 0 = normal), his slow rate and occasional pausing may have caused some of the rate rating variability. His tendency to have a rising intonation contour rather than a typical falling intonation contour for his sentence productions may have also lead to poor reliability of rate and voice quality.

Lower Reliability and Agreement Levels for Rate/Rhythm/Prosody

Lower levels of reliability and agreement were found for rate/rhythm/prosody both within and between listeners, though it is interesting to note that rate was rated more consistently within than between listeners. There are several potential explanations for the poor reliability and agreement of ratings for the rate/rhythm/prosody category. It is possible that the rate category was too large, with too many elements (rate, pauses, stress, intonation) to combine. Listeners are not always good at separating some perceptual features into their components (Kreiman et al., 1993). It is also possible that the descriptions provided to the listeners for this category need to be modified to reduce variability of interpretation. Another possibility is that rate and other aspects of speech may interact with each other, making it difficult to separate the perceptual features, such as the interaction between rate and nasality (Dwyer, Robb, O'Beirne, & Gilbert, 2009).

Potential Limitations of the Study

Some potential limitations of the study should be addressed. As for the talkers, there were few severe cases of dysarthria, leading to a relatively small range of severity. When there are large numbers of normal parameters being rated, reliability and agreement levels can be overinflated (Sheard et al., 1991). With only PD and OPMD

talkers, there was also a small range of etiologies represented in this study. Further research should involve a wider variety of severity levels and etiologies. As for the listeners, there was a large number of them, but their experience with disordered speech was limited (i.e., graduate students with variable amounts of exposure to dysarthric speech and assessment methods for dysarthric speech). There were also some limitations of the listening task itself. We provided instructions but no training for the listeners prior to beginning the listening task, though structured training has been recommended for higher levels of intra- and inter-rater consistency (Chenery, 1998) The use of training might limit the generalizability of the study findings though (Sheard et al., 1991). In future research, conversational samples rather than short sentences read aloud should be used to evaluate a more clinically valid representation of connected speech (Weismer, Jeng, Laures, Kent, & Kent, 2001). This would also minimize the flattening effect of familiar material (i.e., the same sentences for each talker) on ratings (Sheard et al., 1991). The use of reference talkers (i.e., speech samples with moderate severity for the perceptual features being rated) has also been recommended (Kreiman et al., 1993; Chan & Yiu, 2002), though reference talkers were not utilized in the current study. Further research should be completed to evaluate the effect of reference talkers on the reliability of the I-RAVN. Another limitation of the study is that the talkers were presented in the same order to each of the listeners, opening up the possibilities of sequencing and order effects (Schiavetti & Metz, 2006).

Some disadvantages of perceptual analysis in general have been identified in the motor speech literature. Perceptual analysis can be influenced by listener experience and skill, as well as environmental effects on the talker, leading to difficulty with

standardization of this type of analysis (Chenery, 1998). As previously mentioned, certain aspects of speech may influence the perception of other aspects (Sheard et al., 1991), and a number of deficits can result in similar perceptual differences, making it difficulty in some cases to determine the pathophysiology (Chenery, 1998). For this reason, the I-RAVN, like other auditory-perceptual rating tools, should be combined with other assessment tools in order to obtain a complete view of the speakers' strengths and weaknesses (Oates, 2009). It has been suggested that having the listeners make multiple ratings for each stimulus and averaging those ratings might lead to better consistency both within and between listeners (Shrivastav, Sapienza, & Nandur, 2005). This process should be addressed in future research on the I-RAVN explanatory tool.

Strengths of the Study

The study also has numerous strengths, including the fact that there was a large number of listeners, and that good reliability and agreement levels were obtained for intelligibility and articulation even without listener training or reference talkers. The results of this experiment demonstrate that the I-RAVN tool can be utilized by even unfamiliar listeners, which is representative of many communication partners throughout daily life (e.g., cashiers and bank tellers) (Hustad & Cahill, 2003). The experiment was conducted in a manner conducive to quality measurement (Schiavetti & Metz, 2006), with a consistent and minimally distracting testing environment, consistent equipment that had been calibrated, and consistent written instructions to the listeners. The listening task itself is an easy procedure that takes minimal time to complete, unlike some other perceptual analysis procedures (e.g., Darley et al., 1969; Bunton et al., 2007). The I-RAVN tool would be easy to use clinically, and would be inexpensive and readily available. Identification of perceptual features is commonly used as the first tool of

evaluation (Chenery, 1998), especially since perceptual analyses in general are more meaningful to clients, families, caregivers, and other professionals than some other types of analysis, such as acoustic analysis (Oates, 2009). For this reason, it is important to establish a quick, easy, and reliable perceptual analysis tool. The I-RAVN tool is similar to other perceptual evaluation techniques in that it would require little extra training, since clinicians' training for evaluation of dysarthria involves substantial training in the identification of perceptual features (Chenery, 1998). The I-RAVN tool could also be used to monitor change during therapy, since perceptual evaluation tools in general are sensitive to subtle changes in performance (Chenery, 1998). Another advantage to the I-RAVN tool is that because it is based on the physiologic approach to motor speech disorders (assessing the individual motor subsystems) (Netsell & Daniel, 1979), it is more useful than rating overall intelligibility; this can be seen by examining individual talker profiles. For example, Talker O5 had a mean intelligibility rating of 21.7, a mean rate/rhythm/prosody rating of 45.2, a mean articulation rating of 12.8, a mean voice quality/breath support rating of 40.8, and a mean nasality rating of 5.8. A variable profile was found for many other talkers as well. Talker P4 received a mean intelligibility rating of 67.6, a mean rate rating of 43.0, a mean articulation rating of 71.6, a mean voice rating of 48.3, and a mean nasality rating of 56.9. Since Talker O5's ratings were most deviant from normal for rate and voice, it would be expected that she would receive greatest gains initially by beginning therapy in those areas. Based on Talker P4's profile, though, it would be expected that he would most benefit initially from articulation therapy. The profiles for Talkers O5 and P4 can be seen in Figures 1 and 2. Further research should be completed to confirm that certain I-RAVN perceptual features are more highly correlated

with specific dysarthrias (De Bodt et al., 2002). As with all perceptual analyses, the I-RAVN should be used to identify further assessments to provide information about specific goals for therapy (Kent et al., 1989), including acoustic analysis (Weismer et al., 2001).

Figure 1. Profile for Talker O5.

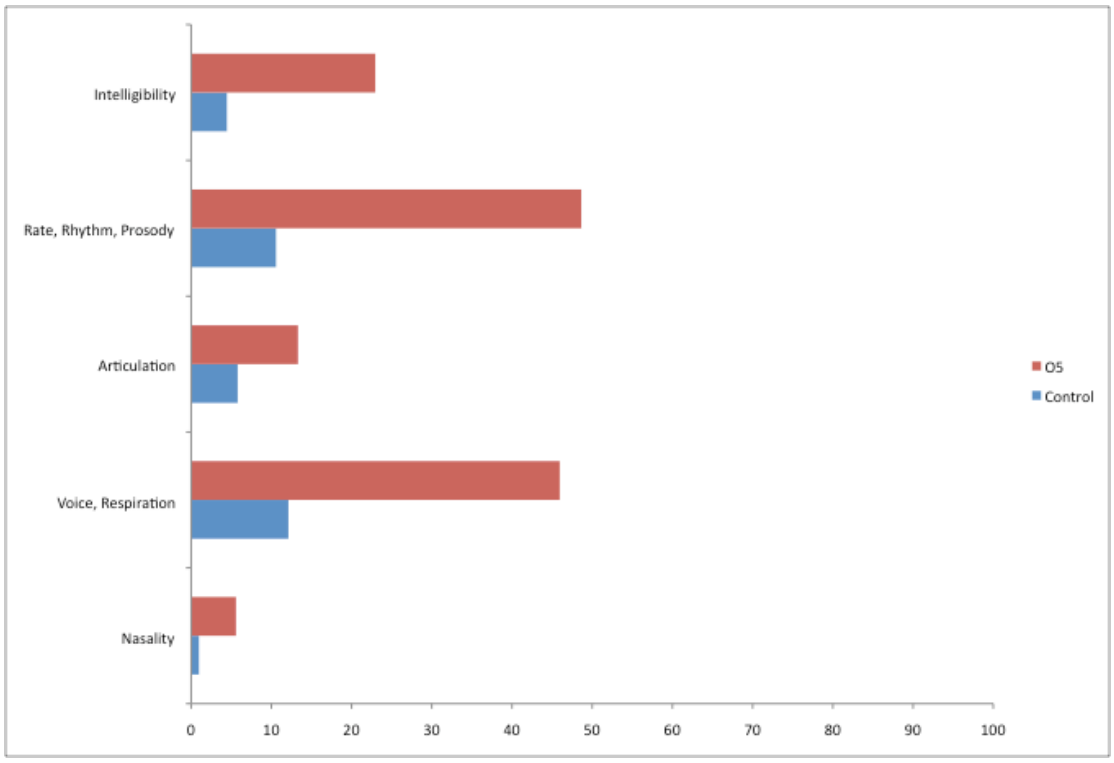
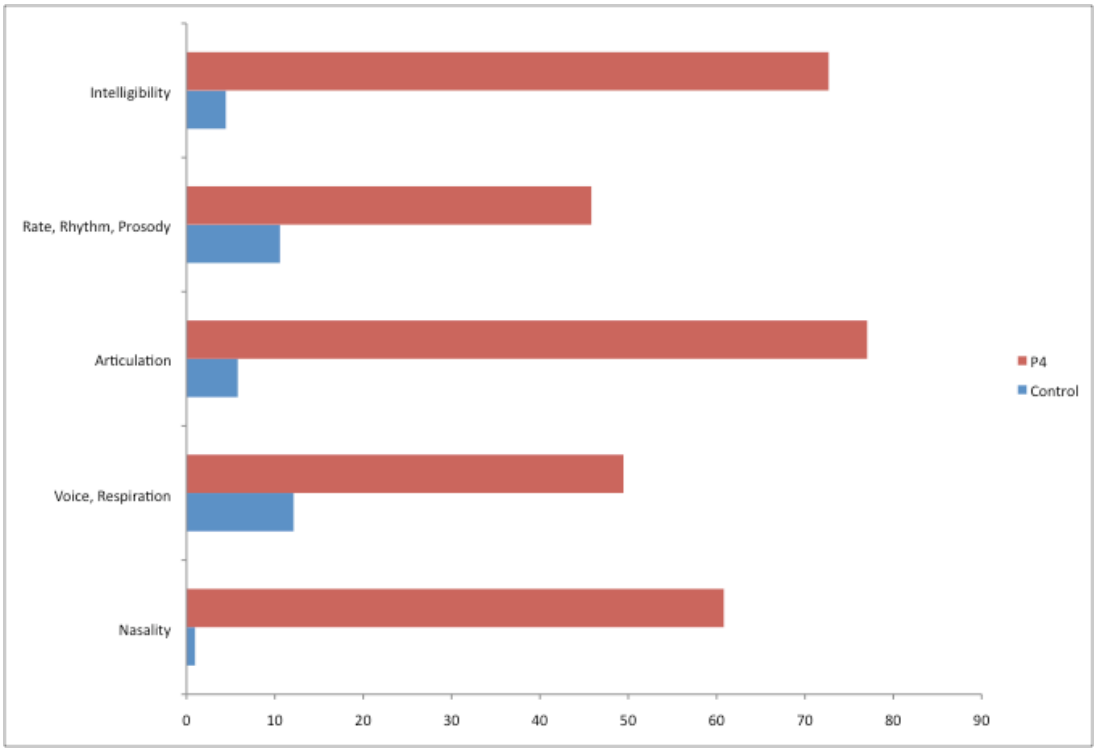


Figure 2. Profile for Talker P4.



It has been suggested that reliability and agreement be measured in various ways to ensure a more complete view of rater variance (Sheard et al., 1991). We found similar results across a number of statistical methods (Pearson correlations, factor analyses, ICCs, rater bias ANOVAs, and percent close agreements) to determine the intra- and inter-rater reliability and agreement for the I-RAVN explanatory tool. Reasonable levels of reliability and agreement were found for the intelligibility, articulation, voice, and nasality perceptual features. Rate should be studied further to determine possible reasons for its lower levels of reliability and agreement.

LIST OF APPENDICES

Appendix A

Listener Instructions

IRAVN Rating Scale Instructions

You'll be listening to a set of sentences produced by talkers with normal speech and with dysarthria. For each talker, you will first rate the ***overall intelligibility*** of their speech. Intelligibility refers to how easy or hard it is to understand their speech. You'll be hearing the same sentences over and over again, so you may want to imagine how easy or difficult it would be to understand the talker if you were to have a conversation with them.

Next you'll rate several aspects of speech that may affect speech intelligibility.

- First you'll rate the ***rate, rhythm, and prosody*** of their speech. You'll need to decide if their speech rate is normal, too fast, or too slow. You can comment on the number and location of pauses in their sentences. You may address the appropriateness of syllable stress on stressed and unstressed syllables in the sentences. Finally, you can comment on the adequacy of intonation – the use of pitch over the course of the sentence.
- Second, you'll rate the talker's ***articulatory precision*** – how accurately the speech sounds were pronounced. You can make comments about the production of consonants and vowels.
- Third, you'll rate the talker's ***voice quality and breath support***. Voice quality includes the appropriateness of pitch to the talker's age and gender, the loudness of their voice, and whether their voice is characterized by roughness, hoarseness, breathiness, harshness, strain, weakness, or tremor. ***Breath support*** refers to the sufficiency of air supply and ability to control expiratory flow of air for speech.
- Finally, you'll rate the talker's ***nasality*** or resonance balance. You can comment on whether the talker seems hypernasal (air being resonated in nasal cavity for oral sounds) or hyponasal (air is prevented from being resonated in the nasal cavity for nasal sounds), or if audible nasal air emission is present.

To rate each dimension, you'll make a small vertical mark along the gray horizontal line. If the intelligibility or aspect of speech is normal, you'll place a mark near the left end of the line marked "NO" for normal. If the intelligibility or aspect of speech is abnormal, you can choose to mark the line anywhere normal to "SE" for severely deviant. You can mark the line between categories. For example, if you believe that the dimension is mildly to moderately impaired, you can mark the line between "MI" and "MO." If the talker is consistent in their behavior, circle the "C" to the right of the rating line. If they are inconsistent in that dimension, circle the "I."

You can write any comments about each of the dimensions in the spaces provided below the rating line. For example, if a talker seems to be having difficulty producing a clear /s/ sound, you could write "distorted /s/" in the "Consonants" box of the "Articulatory Precision" section. You may listen to the set of sentences a second time if you need to.

Legend	
C = Consistent	**I** = Inconsistent
NO	Normal
MI	Mildly deviant
MO	Moderately deviant
SE	Severely deviant

C I

NO MI MO SE

Descriptive Terms for Impairments of Speech

Rate, Rhythm and Prosody

- Rate – speed of movement of the articulators, can be measured in syllables or phonemes per second
  - Too fast
  - Too slow
  - Inappropriate variability (rushes of speech, uneven rate)
- Pauses – intervals of silence between syllables or words
  - Too few pauses occur
  - Too many pauses occur
  - Pauses are too long
  - Pauses occur in inappropriate locations (within word or syntactic units)
- Stress – use of emphasis to convey word or sentence meaning
  - Excess stress on syllables that are usually unstressed
  - Equal stress on all syllables
- Intonation – use of pitch across an utterance to convey meaning
  - Monopitch
  - Excessive or unpredictable pitch variation

Articulatory Precision

- Consonants – accurate production of consonant phonemes
  - Imprecise consonants could include distortions, substitutions, omissions, additions
- Vowels
  - Distorted vowels could include vowel distortions, substitutions, vowel length errors
- Irregular articulatory breakdown

Voice Quality and Breath Support for Speech

- Pitch – f_0 is appropriate for age and gender
  - Too low
  - Too high
  - Too variable
- Loudness – volume of voice is adequate
  - Too loud
  - Too soft
  - Too variable
- Quality – timbre of voice, characteristics other than voice and loudness
  - Harsh, hoarse, rough, or raspy
  - Too breathy
  - Strained-strangled – effortful squeezing of voice through glottis
  - Weak – voice is not rich and resonant
  - Tremor – rhythmic unsteadiness in pitch or volume
  - Wetness – gurgling quality
- Breath support – sufficient supply of air and control of expired air for speech
  - Speaker runs out of air before end of breath group in phrase

Nasality

- Hypernasal – inappropriate nasal resonance for oral phonemes
- Audible nasal air emission
- Hyponasal – lack of nasal resonance for nasal consonants because of obstructed nasal tract

Appendix B

I-RAVN Rating Form

I-RAVN Rating Scale for Speech Intelligibility

Listener: _____ **Date:** _____ **Talker:** _____

I **Overall impression of speech intelligibility**

NO MI MO SE C I ____/100

R **Rate, rhythm and prosody**

NO MI MO SE C I ____/100

Rate of speech	
Pauses	
Stress	
Intonation	

A **Articulatory precision**

NO MI MO SE C I ____/100

Consonants	
Vowels	

V **Voice Quality and Breath Support**

NO MI MO SE C I ____/100

Pitch	
Loudness	
Quality	
Breath support	

N **Nasality**

NO MI MO SE C I ____/100

Resonance balance	

I2 **Overall impression of speech intelligibility**

NO MI MO SE C I ____/100

References

Bunton, K. (2006). Fundamental frequency as a perceptual cue for vowel identification in speakers with Parkinson's Disease. *Folia Phoniatrica et Logopaedica, 58*, 323-339.

Bunton, K., Kent, R.D., Duffy, J.R., Rosenbek, J.C., & Kent, J.F. (2007). Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech- Language Pathology, 50,* 1481-1495.

Chan, K.M.K., & Yiu, E.M-L. (2002). The effects of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45*(1), 111-126.

Chenery, H.J. (1998). Perceptual analysis of dysarthric speech. In B.E. Murdoch (Ed.), *Dysarthria: A physiological approach to assessment and treatment* (36-67). Cheltenham, United Kingdom: Stanley Thomas Ltd.

Darley, F.L., Aronson, A.E., & Brown, J.R. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12*, 246-269.

De Bodt, M.S., Huici, M.E.H-D., Van de Heyning, P.H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders, 35,* 283-292.

Donovan, N.J., Kendall, D.L., Young, M.E., & Rosenbek, J.C. (2008). The communicative effectiveness survey: preliminary evidence of construct validity. *American Journal of Speech-Language Pathology, 17,* 335-347.

Duffy, J.R. (2005). *Motor speech disorders: Substrates, differential diagnosis, and management* (2nd ed.). St. Louis, Missouri: Elsevier Mosby.

Dwyer, C.H., Robb, M.P., O'Beirne, G.A., & Gilbert, H.R. (2009). The influence of speaking rate on nasality in the speech of hearing-impaired individuals. *Journal of Speech, Language, and Hearing Research, 52*(5), 1321-1333.

Gayvert, R., & Hillenbrand, J. (2003). Open source software for speech perception research. *The Journal of the Acoustical Society of America, 113*, 2260.

Hustad, K.C. (2006). Closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners *American Journal of Speech-Language Pathology, 15,* 268-277.

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.

Keintz, C.K., Bunton, K., & Hoit, J.D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 16*, 222-234.

Kempster, G.B., Gerratt, B.R., Abbott, K.V., Barkmeier-Kraemer, J., & Hillman, R.E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18,* 124-132.

Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*(4), 482-499.

Kreiman, J., Kempster, G.B., Erman, A., & Berks, 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.

Laures, J.S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42,* 1148-1156.

Mitchell, S.K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86*(2), 376-390.

Moser, H.M., Dreher, J.J., & Adler, S. (1955). Comparison of hyponasality, hypernasality, and normal voice quality on the intelligibility of two-digit numbers. *The Journal of The Acoustical Society of America, 27*(5), 872-874.

Most, T., Weisel, A., & Lev-Matezky, A., (1996). Speech intelligibility and the evaluation of personal qualities by experienced and inexperienced listeners. *Volta Review, 98*(4), 181-190.

Neel, A.T. (2009). Intelligibility of loud, amplified, and habitual speech in Parkinson Disease. *Journal of Speech-Language Hearing Research, 52,* 1021-1033.

Neel, A.T., Palmer, P.M., Sprouls, G., and Morrison, L. (2006). Tongue strength and speech intelligibility in oculopharyngeal muscular dystrophy. *Journal of Medical Speech-Language Pathology, 14,* 273-277.

Netsell, R., & Daniel, B. (1979). Dysarthria in adults: Physiologic approach to rehabilitation. *Archives of Physical Medicine & Rehabilitation, 60*, 502-508.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica, 61*, 49-56.

Platt, L.J., Andrews, G., Young, M., & Quinn, P.T., (1980). Dysarthria of adult cerebral palsy: I. Intelligibility and articulatory impairment. *Journal of Speech and Hearing Research, 23*(1), 28-40.

Schiavetti, N., & Metz, D.E. (2006). *Evaluation research in communication disorders* (5th ed.). Boston, MA: Pearson Education, Inc.

Sheard, C., Adams, R.D., Davis, P.J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech and Hearing Research, 34*, 285-293.

Shrivastav, R., Sapienza, C.M., & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research, 48*, 323-335.

Uebersax, J. (2010). *Statistical methods for rater and diagnostic agreement*. Retrieved from http://www.john_eubersax.com/stat/agree.htm

Van der Graaff, M.V., Kuiper, T., Zwinderman, A., Warrenburg, B.V., Poels, P., Offeringa, A., … De Visser, M.(2009). Clinical identification of dysarthria types among neurologists, residents in neurology and speech therapists. *European Neurology, 61,* 295-300.

Van Nuffelen, G., De Bodt, M., Wuyts, F., & Van de Heyning, P. (2009). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatrica et Logopaedica, 61*, 69-75.

Weismer, G., Jeng, J-Y., Laures, J.S., Kent, R.D., & Kent, J.F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*, 1-18.

Weismer, G., & Laures, J.S. (2002). Direct magnitude estimates of speech intelligibility in dysarthria: Effects of a chosen standard. *Journal of Speech and Hearing Research, 45*, 421-433.

Weismer, G., & Martin, R.E. (1992). Acoustic and perceptual approaches to the study of intelligibility. In R.D. Kent (Ed.). *Intelligibility in speech disorders: Theory, measurement and management* (68-118). Amsterdam: John Benjamin.

Zeplin, J., & Kent, R.D. (1996). Reliability of auditory-perceptual scaling of dysarthria. In D.A. Robin, K.M. Yorkston, & D.R. Beukelman (Eds.), *Disorders of Motor Speech: Assessment, Treatment, and Clinical Characterization* (145-154). Baltimore: Paul H. Brookes Publishing Co.

Zyski, B.J., & Weisiger, B.E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders, 20*(5), 367-378.