Listener Agreement for Auditory-Perceptual Ratings of Dysarthria

Kate Bunton
Raymond D. Kent
Waisman Center, Madison, WI

Joseph R. Duffy
Mayo Clinic, Rochester, MN

John C. Rosenbek
William S. Middleton Memorial Veterans Hospital, Madison, WI

Jane F. Kent
Waisman Center

Purpose: Darley, Aronson, and Brown (1969a, 1969b) detailed methods and results of auditory-perceptual assessment for speakers with dysarthrias of varying etiology. They reported adequate listener reliability for use of the rating system as a tool for differential diagnosis, but several more recent studies have raised concerns about listener reliability using this approach.
Method: In the present study, the authors examined intrarater and interrater agreement for perceptual ratings of 47 speakers with various dysarthria types by 2 listener groups (inexperienced and experienced). The entire set of perceptual features proposed by Darley et al. was rated based on a 40-s conversational speech sample.
Results: No differences in levels of agreement were found between the listener groups. Agreement was within 1 scale value or better for 67% of the pairwise comparisons. Levels of agreement were lower when the average rating fell in the mid-range of the scale compared with samples that had an average rating near either of the scale endpoints; agreement was above chance level. No significant differences in agreement were found between the perceptual features.
Discussion: The levels of listener agreement that were found indicate that auditory-perceptual ratings show promise during clinical assessment for identifying salient features of dysarthria for speakers with various etiologies.

KEY WORDS: auditory-perceptual ratings, dysarthria, listener agreement

Auditory-perceptual assessment methods are considered the "gold standard" for clinical differential diagnosis, judgment of severity, decisions about management, and the assessment of functional change in dysarthria. The work of Darley, Aronson, and Brown (1969a, 1969b) is the foundation for current clinical methods of auditory-perceptual assessment and classification of the dysarthrias. The descriptions presented in their pair of papers and in their 1975 book, *Motor Speech Disorders,* are central to many contemporary descriptions of dysarthria and have been the basis for many investigations of the underlying acoustic, physiologic, and neuroanatomic bases of dysarthria (Darley, Aronson, & Brown, 1975a). In this article, we refer to their method as the *Mayo Clinic rating system*. This system is also notable in that it is one of the most comprehensive auditory-perceptual systems developed for the clinical assessment of disordered speech. It includes 38 perceptual dimensions relating to respiration, voice, articulation, prosody, and other aspects of speech. Despite the primacy of this auditory-perceptual system in contemporary clinical practice and its influence on acoustic and physiologic studies of speech production, few attempts have been made to establish rater reliability of the rating system for different types of dysarthria. This issue demands further attention if the auditory-perceptual approach to describing and classifying dysarthria is to be a validated clinical and research tool.

Only two published studies have attempted to replicate the work of Darley and colleagues and establish rater reliability using the Mayo Clinic rating system (Zeplin & Kent, 1996; Zyski & Weisiger, 1987). Both studies reported differences from the original publications (Darley, Aronson, & Brown, 1969a, 1969b) on lists of deviant perceptual features for the different dysarthria types and raised questions about interrater reliability. The approach used in these two studies differed from the original studies of Darley, Aronson, and Brown (1969a, 1969b) in that the listeners had no a priori knowledge of the neurologic condition for the speakers. Both studies, however, used recorded speech samples originally collected by Darley and colleagues in their *Audio Seminar Series* (1975b) as stimulus material. In the first study, conducted by Zyski and Weisiger (1987), interrater reliability was calculated for two groups of listeners: experienced clinicians and graduate students. Listeners were asked to rate key perceptual features for each speaker based on presentation of a standard reading passage (from *My Grandfather*; Gray, 1936) and syllable repetitions. Features were rated on a 7-point scale (1 = *does not deviate from normal,* 7 = *severe deviation from normal*). The list of perceptual features presented for the raters included only those that (a) had mean scale values greater than 2.0 in the original Darley, Aronson, and Brown (1969a, 1969b) studies and (b) were present in no more than four of the seven dysarthria types. The 16 perceptual features that met these criteria were selected to maximize differentiation among dysarthria types. In preselecting a smaller number of perceptual features, the authors hoped to control artificially inflated reliability coefficients by eliminating features that were likely to be judged within normal limits. The authors' decision to focus their analysis on those features with the greatest variability likely contributed to lower correlations and the negative conclusion that the Mayo Clinic rating scale was not sufficiently reliable for clinical purposes. In the second study, by Zeplin and Kent (1996), stimuli from two speech tasks—syllable repetition and passage reading—were presented to five judges, all of whom had experience with dysarthria. Results of this study showed that listeners were able to identify key perceptual features of dysarthric speech and had good intrarater reliability; however, significant differences in degree of interrater reliability among perceptual features were reported. The authors concluded that ratings of certain perceptual features may be more reliable than others and that this finding warrants further study. Considering the discordant results from these two studies, it cannot be stated with confidence that the reliability of the Mayo Clinic system has been established. These findings are unsettling because samples used in these studies were the speech materials prepared at the Mayo Clinic for purposes of listener training.

Two additional studies of interrater reliability using the Mayo Clinic rating system should be noted here (Kearns & Simmons, 1988; Sheard, Adams, & Davis, 1991). In contrast to the studies discussed previously, the focus in these two studies was on a single type of dysarthria—ataxic—which likely decreased the number of features rated as deviant, resulting in a disproportionate number of features rated as normal, thereby inflating the reliability measures. Kearns and Simmons (1988) reported no differences in rater reliability across perceptual features. Overall, their reported mean occurrence reliability rating of 82% for experienced speech-language pathologists is comparable to that reported for expert judges in the original studies (Darley, Aronson, & Brown, 1969a). Sheard et al. (1991), on the other hand, reported significant differences in rater reliability across the perceptual features. A strength of these studies was that they used speech samples collected from a new group of speakers (as opposed to the original Darley et al. database).

Given the clinical reliance on perceptual analyses for the evaluation of dysarthria and the suggestions from several studies that the reliability within current clinical practice may be generally inadequate (Duffy & Kent, 2001), additional research is needed to evaluate the reliability of the Mayo Clinic rating system. For purposes of generalization, it is also essential to obtain reliability estimates for a newly collected set of dysarthric speakers.

The studies discussed above (except Kearns & Simmons, 1988) used correlation as a measure of interrater reliability. Interrater reliability measures provide an indication of the extent to which the variance in the ratings is attributable to differences among the objects being rated (Tinsley & Weiss, 2000). A high interrater reliability means that the relation of one rated object to another rated object is the same across judges, even though the absolute numbers used to express this relation may differ from judge to judge. In other words, interrater reliability is sensitive to the relative ordering of the rated objects but does not tell us whether a particular score on the scale used by one rater is equivalent to the same score provided by a second rater. Interest in establishing the merit of the Mayo Clinic rating scale extends to whether or not the rating scale is reliable and if the scale values are meaningful independent of the rater. This is a question of interrater *agreement,* not interrater reliability (for a discussion, see Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Measures of interrater agreement represent the extent to which the different judges tend to assign exactly the same rating to each object. Therefore, an index of the interrater agreement provides critical information about the meaning of the particular ratings and allows us to answer questions about the consistency with which clinicians use the rating scale.

One additional important element of auditory-perceptual evaluation is the effect that the speech task

has on ratings. Differences in ratings of the 38 perceptual features across speech task (syllable repetition and passage reading) were reported by Zeplin and Kent (1996). This raises some concern about which task should be used for perceptual ratings or if multiple samples should be included. It is likely that some perceptual features will be identified in some tasks and not others. Evidence that certain features may be rated differently depending on the speech task judged has been reported by Kent, Kent, Rosenbek, Vorperian, and Weismer (1997) and Brown and Docherty (1995). The original work by Darley, Aronson, and Brown (1969a, 1969b) included three speech tasks: vowel prolongation, syllable repetition, and passage reading. Zyski and Weisiger (1987) and Zeplin and Kent (1996) included only samples from syllable repetition and reading tasks. Studies that focused on descriptions of, and performance during, conversational speech have been limited; however, certain perceptually appreciable disturbances may be most fully expressed in conversational speech (e.g., prosody, rate, articulatory precision). Conversational speech may also facilitate rating of the overall impression categories of intelligibility and bizarreness. Recent reviews summarizing differences in conversational speech across dysarthria type can be found in Kent, Kent, Duffy, and Weismer (1998) and Kent and Kent (2000). On the basis of these assertions, it seems that identification of perceptual features may be maximized through use of a conversational task. There has also been a recent shift in clinical focus to functional outcome measures. Thus, conversational speech is being targeted more directly during remediation programs. Examining listener agreement for ratings of perceptual features based on a conversational sample may be a first step in linking methods of assessment and treatment that focus exclusively on functional behaviors.

The present study sought to determine if raters have satisfactory agreement for the perceptual features of the Mayo Clinic rating system, as applied to a group of 47 speakers with dysarthria. Listeners, blind to speakers' neurologic diagnosis and type of speech disturbance, were asked to rate the 38 perceptual features of the Mayo Clinic system on a 7-point rating scale based on a short conversational sample.

Method

Recorded Samples of Dysarthric Speech

Speech samples were obtained from 47 individuals with various types of dysarthria due to a variety of neurologic conditions. The speakers were recorded as part of a larger study of dysarthria at the University of Wisconsin–Madison and were collected in conjunction with the Mayo Clinic and the William S. Middleton Memorial Veterans Hospital (Madison, WI). Dysarthria types were determined by expert speech-language pathologists (third author [JRD] and fourth author [JCR]) based on a complete clinical assessment. A complete assessment typically involved a history of the speech problem; judgments about strength, symmetry, range of motion, and adventitious movements during an oral mechanism examination (as described by Duffy, 2005); perceptual judgments about respiration, phonation, resonance, articulation, and prosody during conversation, reading, and vowel prolongation; and rapid alternating motion rates for /pʌ/, /tʌ/, and /kʌ/ and sequential motion rates for /pʌtʌkʌ/.

Individual speaker characteristics are summarized in Table 1. The dysarthria types examined in the current study included hypokinetic, mixed (spastic-flaccid), flaccid, spastic, and ataxic. Hyperkinetic dysarthria was not included in this study because of the small number of speakers in the database with this dysarthria type. In the current study, the neurologic disease and lesion sites included amyotrophic lateral sclerosis, Parkinson disease, left-sided stroke, right-sided stroke, bilateral stroke(s), multiple sclerosis, Guillian-Barre syndrome, and cerebellar degeneration. A neurologist confirmed lesion locations for the speakers with stroke. Speakers were diverse in several other respects, including severity of dysarthria, duration of the disease, and medical history. The dysarthria diagnosis made by the speech-language pathologist was never incompatible with what might be predicted from lesion loci or neurologic diagnosis.

Speech Sample

The speech samples selected for presentation were 40-s clips of conversational speech taken from an initial interview between a speech-language pathologist and the speaker being assessed. The samples were selected from periods of continuous speech and did not contain any speech produced by the speech-language pathologist. In addition, the samples did not contain any potentially leading information with regard to speech symptoms, underlying neurologic disease, or hospital stay. Open-ended questions about the speaker's family, work, or hobbies were used to elicit the speech sample. Speech samples were recorded on digital audiotape and digitized into CSpeech (Milenkovic, 1994) for presentation to the listeners. Samples were low-pass filtered at 9.8 kHz and digitized at a sampling rate of 22.05 kHz. Individual speaker files were coded for identification purposes.

Listeners

Two groups of listeners were included in the present study. The first group, inexperienced clinicians, included 10 listeners who had just completed their master's degree program but had not yet begun a clinical fellowship. These listeners had completed a course on dysarthria and had received 5 hr of classroom training on

Table 1. Individual speaker characteristics.

Speaker	Gender	Age (yrs)	Neurologic diagnosis	Duration of disease (mos)	Dysarthria type
1	M	72	Bilateral stroke	10	Ataxic
2	M	69	Bilateral stroke	1	Ataxic
3	M	78	Bilateral stroke	12	Ataxic
4	F	55	CBLR degeneration	6	Ataxic
5	F	69	CBLR degeneration	21	Ataxic
6	F	65	CBLR degeneration	1	Ataxic
7	F	50	CBLR degeneration	9	Ataxic
8	M	60	CBLR degeneration	10	Ataxic
9	F	53	MS	1	Ataxic
10	F	31	MS	29	Ataxic
11	M	57	MS	1	Ataxic
12	M	22	MS	6	Ataxic
13	F	81	brainstem stroke (VII nerve involvement)	1	Flaccid
14	M	44	CBLR aneurysm (X and XII nerve weakness)	16	Flaccid
15	M	42	Guillian-Barre (Miller-Fisher variant)	1	Flaccid
16	M	67	Postoperative resection of 4th ventricle tumor	8	Flaccid
17	F	70	Right stroke	1	Flaccid
18	F	69	Right stroke	13	Flaccid
19	M	26	TBI	1	Flaccid
20	F	71	XII nerve lesion-post right carotid endarterectomy	7	Flaccid
21	F	70	PD	48	Hypokinetic
22	F	63	PD	48	Hypokinetic
23	F	61	PD	120	Hypokinetic
24	F	63	PD	10	Hypokinetic
25	M	77	PD	48	Hypokinetic
26	M	67	PD	64	Hypokinetic
27	M	75	PD	12	Hypokinetic
28	M	68	PD	12	Hypokinetic
29	M	38	PD	12	Hypokinetic
30	M	51	PD	120	Hypokinetic
31	F	41	ALS	36	Mixed Spastic-Flaccid
32	F	52	ALS	9	Mixed Spastic-Flaccid
33	F	54	ALS	19	Mixed Spastic-Flaccid
34	F	47	ALS	8	Mixed Spastic-Flaccid
35	F	29	ALS	1	Mixed Spastic-Flaccid
36	M	67	ALS	36	Mixed Spastic-Flaccid
37	M	64	ALS	12	Mixed Spastic-Flaccid
38	M	56	ALS	12	Mixed Spastic-Flaccid
39	M	75	ALS	24	Mixed Spastic-Flaccid
40	M	42	ALS	24	Mixed Spastic-Flaccid
41	M	65	ALS	9	Mixed Spastic-Flaccid
42	F	84	Bilateral stroke	1	Spastic
43	F	73	Bilateral stroke	8	Spastic
44	M	49	Bilateral stroke	17	Spastic
45	M	22	Left stroke	3	Spastic
46	M	56	Left stroke	8	Spastic
47	M	68	Left stroke	1	Spastic

Note. MS = multiple sclerosis; CBLR = cerebellum; TBI = traumatic brain injury; PD = Parkinson disease; ALS = amyotrophic lateral sclerosis.

the perceptual evaluation of dysarthria using the *Audio Seminars in Speech Pathology: Motor Speech Disorders* tapes (Darley et al., 1975b) at the University of Wisconsin–Madison. The second group of listeners was considered experienced and included 10 speech-language pathologists with more than 7 years of clinical experience. This group of clinicians regularly diagnosed and treated individuals with dysarthria as part of their practice. All listeners passed a hearing screening at 25 dB for frequencies of .5, 1, 2, and 4 kHz (American Speech-Language-Hearing Association [ASHA], 1997) and had no self-reported history of speech problems.

Rating Procedure

Listeners were seated at a table in a sound-treated room, and speech samples were presented through a loudspeaker placed 1 m in front of the listener. Listeners were given a written description for each of the 38 perceptual features based on definitions presented in Darley et al. (1969a). They were asked to rate speech samples from the 47 speakers on all 38 features using a 7-point scale (where 1 was defined as *normal* and 7 represented a *very severe deviation from normal*).

Additional features of pitch and loudness were scaled differently. For these two features, scale anchors represented extremes in behavior (e.g., low–high pitch; soft–loud level). Therefore, for these two features, a rating of 4 was considered normal, and the values of 1 or 7 represented deviations. This rating scale replicates the one used by Darley et al. (1969a). Listeners were reminded of the scale anchors at the beginning of each listening session and after every third speaker during a single listening session. Listeners rated all 47 speakers on a single perceptual feature before proceeding to the next feature. Within each feature, speaker order was randomized. In addition, the order of the features to be rated was randomized across listeners. Each listener provided 1,786 ratings (47 speakers × 38 features). To control listener fatigue, rating sessions were limited to 1 hr in duration. This meant that each listener participated in 6–8 listening sessions over a span of no more than 2 weeks.

Intrarater Agreement

To determine intrarater reliability, eight quasirandomly selected perceptual features for each speaker were rated twice by each listener. For example, Listener 1 may have rated the perceptual features of pitch level, monoloudness, harsh voice, nasal emission, phrases short, inappropriate silences, vowels distorted, and intelligibility twice for Speaker 1, whereas Listener 2 may have rated eight different features twice for the same speaker. Using this balanced assignment, it was possible to get a measure of intrarater reliability for each perceptual feature across multiple listeners. Measures were taken to ensure that each perceptual feature was selected an equal number of times.

Interrater Agreement

To examine agreement in assignment of scale values across the group of listeners, the frequency with which two raters agreed with one another for each speaker and feature was calculated. In other words, the rating score given by Listener 1 for Speaker 1, Feature 1 was compared with each of the values assigned by the other nine judges for that speaker and feature in a pairwise manner. This procedure generated 1,710 comparisons per speaker (38 features × 45 pairwise comparisons), thereby yielding 80,370 (1,710 comparisons × 47 speakers) pairwise comparisons.

Differences in interrater agreement have reportedly varied depending on the distribution of ratings along a 7-point interval scale (Kreiman & Gerratt, 1998; Kreiman et al., 1993). To quantify such variability, the probability that two raters would agree exactly for a given speaker and feature was calculated (p-exact) as well as the probability of agreement within one scale value (p-one). To determine if the number of cases of exact agreement or agreement within one scale value exceeded chance levels, a binomial distribution was calculated based on the probability of agreement at the specified level. The probability of chance agreement for independent ratings based on a 7-point scale was calculated as 0.14 for exact agreement and 0.39 for within one scale value (Tinsley & Weiss, 2000). The probability of the observed outcome (pairs with exact agreement or agreement within one scale value) was compared with the alpha level. The hypothesis was tested as two-tailed with an alpha level of .05.

The amount of variance in ratings that was accounted for by differences among speakers was estimated based the sum of square values taken from an analysis of variance (ANOVA) summary table in which the independent variable was the speaker being rated and the dependent variable was the rating received. A separate variance was calculated for each of the 38 perceptual features; because the variance reported was a descriptive statistic, no adjustment of alpha was necessary (Kreiman & Gerratt, 1998; Young, 1993). In this analysis, the error term reflects all other sources of variability, which could include interrater variability and random error. To eliminate any artifact in the results due to high levels of listener agreement near the end of the scale values, analyses included only speakers with mean ratings between 2.5 and 5.5 for each perceptual feature.

Results

Intrarater Agreement

Measures of intrarater agreement were calculated by determining the difference between the first and second rating of eight quasirandomly selected features for each speaker and listener. Differences between the first and second rating were then averaged to provide a single measure of intrarater reliability for each feature. There were no differences between the two listener groups on measures of intrarater reliability based on an ANOVA, $F(1, 3758) = 3.68$, $p = .503$. Therefore, results were collapsed across the two listener groups. Results are presented in Figure 1. The 38 perceptual features rated are shown on the x-axis, and the values on the y-axis are the mean differences between ratings across all speakers and listeners who rated that particular feature twice. There were 148 data points used to calculate each mean. A value of 0 on the y-axis indicates that there was no difference between the first and second rating of a stimulus. A value of 1 indicates that the difference between the first and second rating was 1 scale value, and so on. One feature, grunt at the end of expiration, had a mean difference score of 0. This feature was rated a 1 by all 20 listeners for all 47 speakers. Eight features had mean difference scores equal to or greater than 1. This included 6 features related to voice quality—harsh voice, hoarse voice, breathy voice-continuous, breathy voice-transient, strain-strangle voice, and voice stoppages—and 2 features related to global deficits: intelligibility and bizarreness. The remaining 29 features had mean differences scores between 0 and 1. Overall, these results suggest that individual listeners were reliable in the magnitude of their ratings.

Interrater Agreement

Patterns of interrater agreement were calculated by determining the percent of pairwise agreements. No differences between the two listener groups were found based on an ANOVA, $F(1, 17858) = 2.68$, $p = .266$; therefore, data were collapsed across the two groups of listeners. Data are diplayed in Table 3. Data reported in the table for Columns 2, 3, and 4 are cumulative, meaning that the likelihood of ratings within ±1 scale value is the likelihood of that level of agreement or better (e.g.,

Figure 1. Mean test–retest agreement plus standard deviation bars for individual perceptual features across speakers and listeners. Irreg. artic = irregular articulatory breakdown.

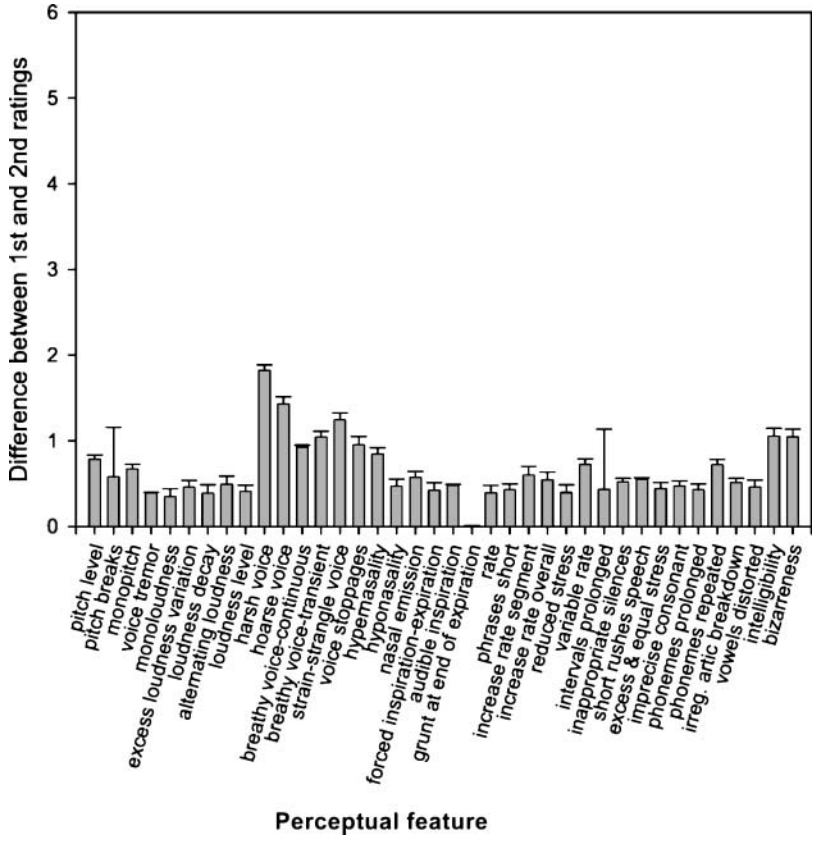


Table 2. Pairwise agreement among listeners.

Perceptual feature	Percent agreement			
	±0	±1	±2	±3 or more
Pitch level	32.78	58.58	77.32	22.68
Pitch breaks	51.44	69.64	82.84	17.16
Monopitch	38.60	67.78	83.54	16.46
Voice tremor	39.59	64.08	79.99	20.01
Monoloudness	34.79	64.42	82.74	17.26
Excess loudness variation	45.65	68.8	84.2	15.8
Loudness decay	38.60	67.45	83.31	16.69
Alternating loudness	49.27	68.74	84.71	15.29
Loudness level	40.80	71.3	87.43	12.57
Harsh voice	42.85	77.22	92.7	7.3
Hoarse voice	43.74	68.86	84.43	15.57
Breathy voice-continuous	42.31	64.14	80.25	19.75
Breathy voice-transient	40.97	62.51	78.7	21.3
Strain-strangle voice	43.55	68.38	82.07	17.93
Voice stoppages	51.94	72.19	85.41	14.59
Hypernasality	44.75	67.96	81.26	18.74
Hyponasality	47.01	69.72	82.7	17.3
Nasal emission	57.37	76.84	88.77	11.23
Forced inspiration-expiration	42.39	69.78	84.16	15.84
Audible inspiration	48.08	77.31	89.5	10.5
Grunt at end of expiration	100.00	—	—	—
Rate	39.14	72.62	89.98	10.02
Phrases short	42.76	68.8	83.67	16.33
Increase rate segment	59.99	82.03	91.79	8.21
Increase rate overall	54.39	75.63	85.17	14.83
Reduced stress	39.64	66.46	83.41	16.59
Variable rate	52.17	75.79	86.46	13.54
Intervals prolonged	43.47	69.81	83.62	16.38
Inappropriate silences	51.91	78.82	90.94	9.06
Short rushes speech	54.85	74.13	86.03	13.97
Excess and equal stress	44.30	65.98	80.19	19.81
Imprecise consonant	47.16	77.41	90.01	9.99
Phonemes prolonged	42.68	73.33	88.58	11.42
Phonemes repeated	72.03	90.77	96.86	3.14
Irregular articulatory breakdown	51.29	74.24	85.55	14.45
Vowels distorted	45.96	74.78	86.79	13.21
Intelligibility	53.76	81.4	90.43	9.57
Bizarreness	45.27	70.46	83.57	16.43

Note. The values in Columns 2, 3, and 4 are cumulative numbers. The numbers in Column 5, separated by a double line, are not.

Table 3. Standard error of estimate (se_e) for probability of agreement data presented in Figures 2 and 3.

Perceptual feature	se_e (p-exact)	se_e (p-one)
Pitch level	0.049	0.072
Pitch breaks	0.071	0.066
Monopitch	0.064	0.071
Voice tremor	0.060	0.060
Monoloudness	0.052	0.085
Excess loudness variation	0.057	0.060
Loudness decay	0.047	0.053
Alternating loudness	0.064	0.063
Loudness level	0.051	0.065
Harsh voice	0.065	0.077
Hoarse voice	0.059	0.059
Breathy voice-continuous	0.053	0.047
Breathy voice-transient	0.035	0.049
Strain-strangle voice	0.044	0.063
Voice stoppages	0.043	0.051
Hypernasality	0.048	0.050
Hyponasality	0.052	0.045
Nasal emission	0.050	0.050
Forced inspiration-expiration	0.040	0.053
Audible inspiration	0.050	0.057
Grunt at end of expiration	0.000	0.000
Rate	0.054	0.077
Phrases short	0.050	0.068
Increase rate segment	0.055	0.054
Increase rate overall	0.064	0.056
Reduced stress	0.053	0.054
Variable rate	0.061	0.054
Intervals prolonged	0.053	0.067
Inappropriate silences	0.052	0.059
Short rushes speech	0.069	0.046
Excess and equal stress	0.055	0.047
Imprecise consonant	0.069	0.073
Phonemes prolonged	0.054	0.056
Phonemes repeated	0.070	0.044
Irregular articulatory breakdown	0.065	0.051
Vowels distorted	0.058	0.059
Intelligibility	0.051	0.039
Bizarreness	0.040	0.036

for pitch level, the 58.58% value reported in Column 3 is the sum of the percent exact agreement [32.78%] and the agreement of ±1 scale value [25.8%]). Data in Column 5 represent gross disagreements of ±3 scale values. These data are not cumulative. The overall percent exact agreement (±0 scale points) calculated by collapsing across features was 47.8% (7,555 of 16,074 ratings were in exact agreement). For individual features, the p-exact agreement ranged from 32.8% to 100%. One feature, grunt at end of expiration, had perfect agreement by all listeners for all 47 speakers. There were 11 features for which the p-exact was greater than 50%. These features included pitch breaks, voice stoppages, nasal emission, grunt at end of expiration, increased rate in segments, increased rate overall, variable rate, inappropriate silences, short rushes of speech, phonemes repeated, irregular articulatory breakdowns, and intelligibility. The overall percentage of ratings that differed by 1 scale value or less (i.e., exact agreement ±1 scale value) was 67.0%. Results for individual features ranged from 58.6% to 90.8%. Ratings were within 2 scale points for 82.9% of the pairwise comparisons. For individual features, percent agreement within 2 scale points ranged from 77.3% to 96.9%.

Gross disagreements, those that differed by ±3 or more scale values, occurred on 17.1% of the pairs. These gross disagreements represent disagreements at or greater than half the length of the scale.

The probability that raters assigned the same scale value to a particular speaker for each of the 38 perceptual features is shown in the multiple panels of Figures 2 and 3. Figure 2 shows the probability of 2 listeners

Figure 2. The probability that 2 listeners had exact agreement in their rating for each perceptual dimension and speaker is plotted against the group mean rating for that speaker. Perceptual dimensions are clustered by category as outlined by Darley, Aronson, and Brown (1969a).

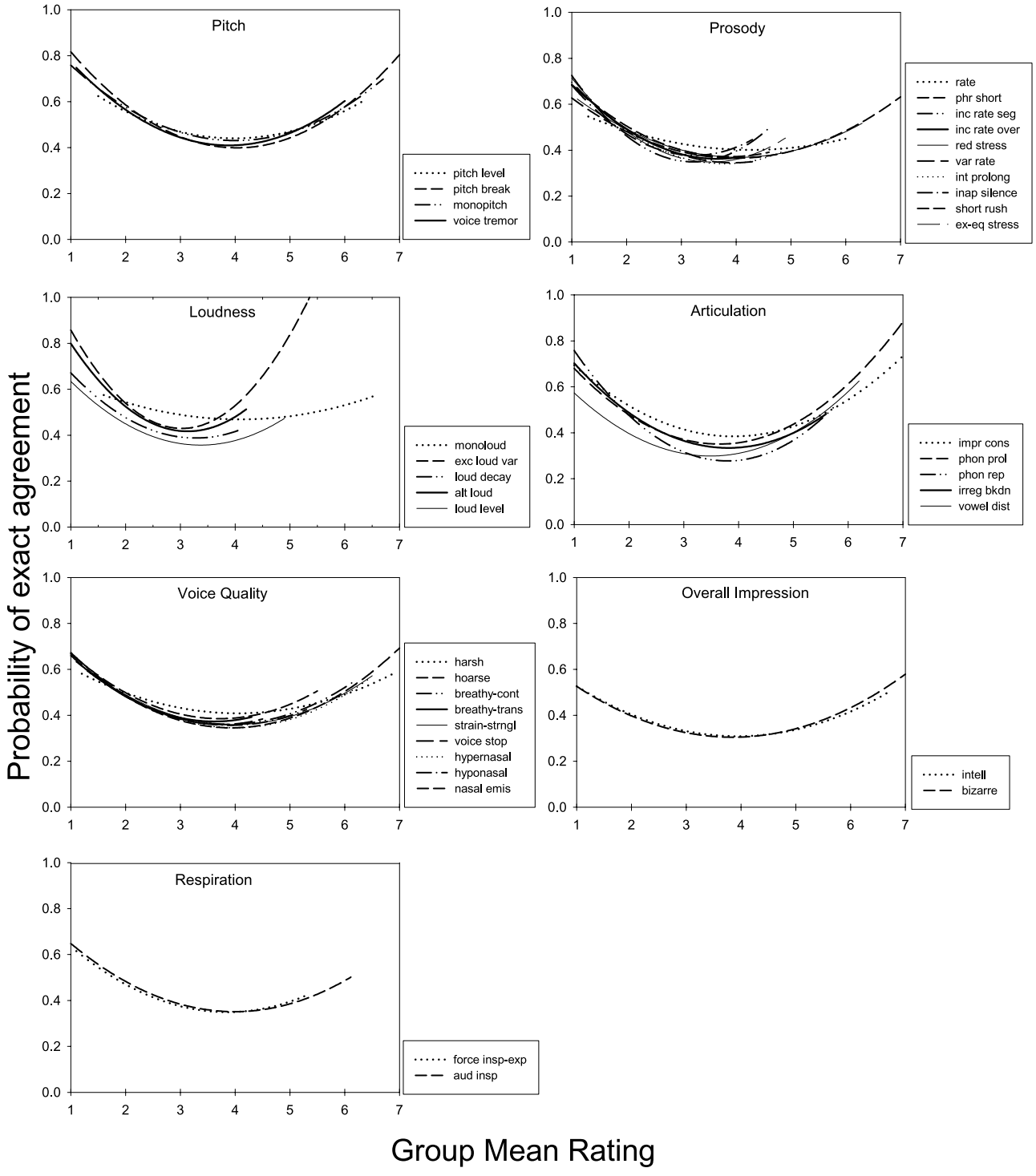
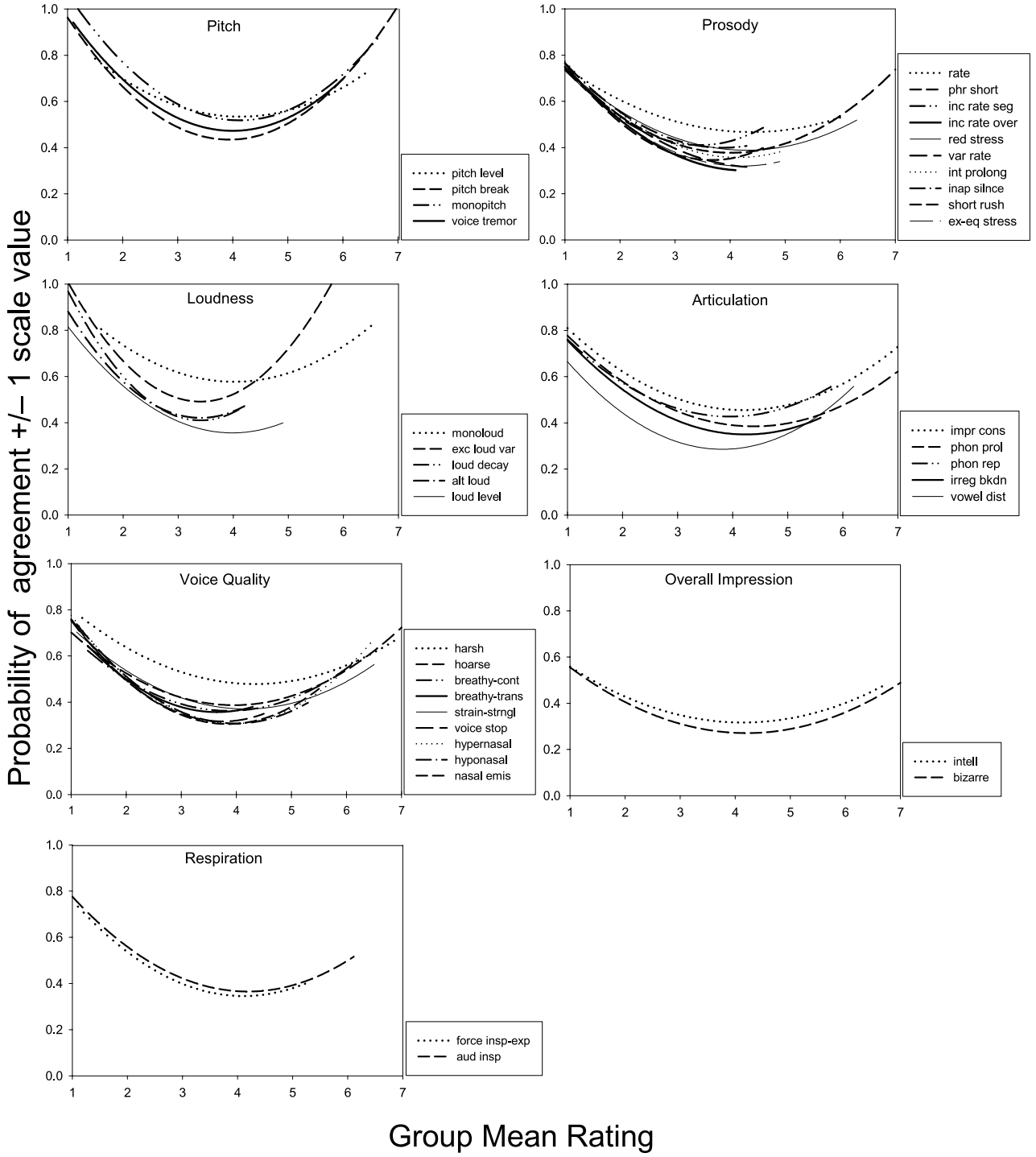


Figure 3. The probability that 2 listeners agreed within 1 scale value in their rating for each perceptual dimension and speaker is plotted against the group mean rating for that speaker. Perceptual dimensions are clustered by category as outlined by Darley, Aronson, and Brown (1969a).

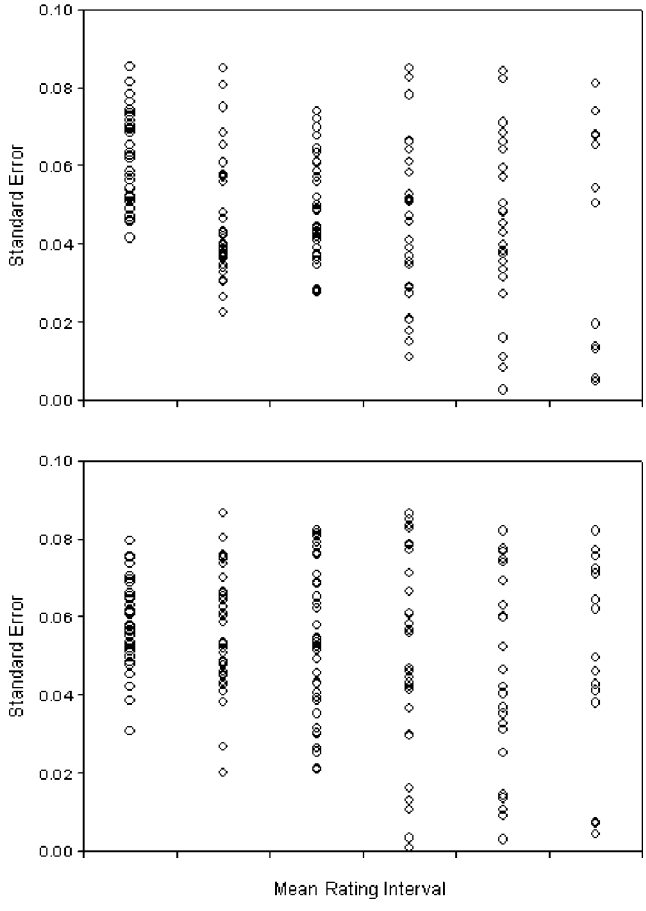


agreeing exactly, whereas Figure 3 shows the probability of listeners' ratings being within 1 scale value for each speaker. Mean standard errors of estimate for probability of agreement in Figures 2 and 3 are shown in Table 3. Previous research has raised concern about differences in interrater agreement level along the 7-point rating scale (e.g., Kreiman & Gerratt, 1998); therefore, the probability of agreement is plotted against the mean rating for each speaker across listeners for each perceptual feature. To efficiently present data for the 38 perceptual features rated

in the present study, panels within each figure represent the perceptual categories of pitch, loudness, vocal quality, respiration, prosody, articulation, and overall impression. These groupings are consistent with those used by Darley et al. (1969a). The large number of data points for each perceptual feature (47) necessitated collapsing the data; therefore, data for each perceptual feature are shown in the panel as a single line calculated using a second-order polynomial. In the plot for respiration (bottom left column), the probability of agreement for ratings of grunt at end of expiration was 1.0 because all listeners rated this perceptual feature as 1 for all 47 speakers; therefore, a best-fit line for this feature is not seen in the plot. The standard error of estimate (se_e) was calculated for intervals between discrete points in the rating scale to determine the dispersion of the data about the fitted line. In other words, a separate se_e value was calculated for the interval in which the mean rating fell between the values of 1 and 2 on the rating scale, 2 and 3, and so on. Data shown in Figure 4 are the se_e for each interval across all 38 perceptual categories. The upper panel is the se_e for the probability of exact agreement, and the lower panel is for the probability of agreement within 1 scale value. Values of se_e ranged from 0.0009 to .092. Data for one perceptual feature, grunt at end of expiration, are not plotted, as listeners were in agreement for all speakers. There were no visible differences in the spread of the data across the length of the rating scale.

Visual examination of the panels in Figures 2 and 3 show that the probability of agreement was consistently lower in the mid-range of the rating scales and higher near the ends of the scale. Mean p-exact value ranged from 0.33 to 0.89 (overall $M = 0.53$, $SD = 0.09$) across the 38 perceptual features. On the basis of the binomial distribution, the number of pairwise comparisons in which exact agreement was observed was above chance level. The features of vowels distorted, intelligibility, and bizarreness had the lowest p-exact values (0.33 or 0.34) at the mid-range of the scale. The features of pitch breaks and excess loudness variation had the highest p-exact values (0.46) at the mid-range of the scale. Probabilities calculated for agreement within 1 scale value ranged from 0.35 to 1.0 ($M = 0.67$, $SD = 0.12$) across the perceptual

Figure 4. Standard error of estimates calculated for intervals along the rating scale for each of the 38 perceptual dimensions.



features. For 32 of the 38 perceptual features, the number of pairwise comparisons in which agreement within 1 scale value was observed was above a level of chance agreement based on the binomial distribution. For 6 perceptual features, the number was not sufficient to reject the null hypothesis and could be attributed to chance. These included strain-strangle voice, hypernasality, vowels distorted, intelligibility, and bizarreness.

The amount of variance in the ratings for each of the 38 perceptual features that was accounted for by differences among speakers is shown in Table 4. Variance accounted for ranged from 36% to 62% (M = 56%, SD = 5.8%). On average, about 50% of the variance in ratings was due to differences among the speakers for the perceptual feature being rated. The remaining variance in ratings was due to other factors.

Table 4. Variance in ratings accounted for by differences among speakers with mean ratings in the mid-range of the scale (2.5–5.5, inclusive).

Perceptual feature	R^2
Pitch level	0.54
Pitch breaks	0.51
Monopitch	0.60
Voice tremor	0.58
Monoloudness	0.59
Excess loudness variation	0.51
Loudness decay	0.51
Alternating loudness	0.48
Loudness level	0.56
Harsh voice	0.62
Hoarse voice	0.60
Breathy voice-continuous	0.56
Breathy voice-transient	0.50
Strain-strangle voice	0.60
Voice stoppages	0.52
Hypernasality	0.54
Hyponasality	0.50
Nasal emission	0.56
Forced inspiration-expiration	0.59
Audible inspiration	0.61
Grunt at end of expiration	—
Rate	0.60
Phrases short	0.58
Increase rate segment	0.55
Increase rate overall	0.36
Reduced stress	0.58
Variable rate	0.39
Intervals prolonged	0.55
Inappropriate silences	0.55
Short rushes speech	0.51
Excess and equal stress	0.51
Imprecise consonant	0.61
Phonemes prolonged	0.57
Phonemes repeated	0.60
Irregular articulatory breakdown	0.59
Vowels distorted	0.60
Intelligibility	0.62
Bizarreness	0.61

Discussion

The purpose of this study was to examine listener agreement when using the auditory-perceptual rating system outlined by Darley et al. (1969a, 1969b) for a group of 47 speakers with dysarthria. To our knowledge, this is the first study to report on listener agreement for all 38 perceptual dimensions using a newly recruited group of speakers with different types of dysarthria. Results of the present study showed reasonable levels of listener agreement for all 38 perceptual features, with no significant differences in rater agreement between listener groups (inexperienced and experienced) or across individual perceptual features. These results are noteworthy in that they support the use of auditory-perceptual ratings of the multiple dimensions that characterize dysarthric speech, which encompasses essentially the full range of abnormalities in speech production. The results, however, may not support use of the auditory-perceptual rating system as the sole tool for clinical differential diagnosis; further research is needed to identify methods for increasing listener agreement for this purpose. Usefulness of auditory-perceptual ratings is discussed in more detail in the following sections.

Agreement Levels for Perceptual Features

Measures of agreement are important for rating scales with clinical application; a high level of inter-listener agreement would imply that listeners assigned identical meanings to each scale point. In other words, different speaker's definitions of what constitutes the endpoints of the scale (*normal* and *extreme*) and distances between intervening points on the scale would be similar. Accordingly, a report of the absolute level of the rating for a given perceptual feature would be meaningful to another rater. The level of interrater agreement found in the present study depended on the perceptual dimension being rated. Overall, exact agreement was obtained for 47.8% of the comparisons, and agreement was within 1 scale value or better for 67%. Percent exact agreement ranged from 32% to 100% for individual perceptual features, with 11 features having greater than 50% agreement (Column 2, Table 2). Cumulative percentages of agreement for ratings differing by 1 scale value or less showed that there were 21 features on which listeners agreed for more than 70% of the pairs (Column 3, Table 2). Agreement at this level ranged from 58% to 69% of the pairs for the remaining 17 features. These findings are encouraging because ratings were based on a variety of speakers with different dysarthria types and underlying

diseases; however, they were lower than levels of interrater agreement reported previously in the literature for the Mayo Clinic rating scale (Darley et al., 1969a; Kearns & Simmons, 1988; Sheard et al., 1991).

No hard rules exist for what constitutes a minimally accepted level of listener agreement, and what is interpreted as satisfactory for one purpose may not be for another. The Mayo Clinic rating scale has been used clinically for many purposes, including differential diagnosis, indexing severity, and determining change over time. For diagnostic purposes, we expect that the minimum level of agreement would be high, as consequences of using tools with low levels of agreement are great; such tools could lead to the misdiagnosis and mismanagement of a speech disorder. Alternatively, current clinical practice—as suggested by Duffy and Kent (2001) and Duffy (2005)—is for clinicians to rely on a pattern of deviant perceptual features rather than on an exact rating of all 38 features. Therefore, a high level of listener agreement may not be crucial to clinical practice. Instead, it would be important to establish interrater agreement for listener identification (presence vs. absence) of deviant perceptual features. Listener reliability and agreement have not been evaluated in this situation. Expectations for levels of agreement when the Mayo Clinic rating scale is used to index severity or change over time are equally important, as careful documentation of therapeutic progress is expected by patients and service providers and is stated in the American Speech-Language-Hearing Association Code of Ethics (2003). Agreement within 1 scale value or better was judged to be reasonable for purposes of the present study based on use of a 7-point scale. Using this criterion, results of the present study show good levels of listener agreement (range: 58.6%–90.8%). These levels of interrater agreement show promise for clinical usefulness of the Mayo Clinic rating scale.

Gross disagreements, defined as greater than 3 scale values, occurred for 17.1% of the trials. This is similar to levels reported by Kreiman and Gerratt (1998) for features related to voice quality. Disagreements of this magnitude would have little clinical utility, as raters did not agree that a sample possessed the same perceptual characteristics. Kreiman and Gerratt (1998) have argued that these findings suggest that the underlying assumption—that listeners can focus their attention on different perceptually distinct aspects of the stimuli and make the required judgments—may not be reasonable. In their case, the assumption was that the overall impression a listener receives from a voice could not be reliably broken down into attributes of breathiness or roughness. Gross disagreements, therefore, may be reflective of a listener's general perception of the signal and not necessarily a reliable quantification of features of the speech sample itself. Further analysis of samples where gross disagreements were found in the present study, however, showed that samples from 4 speakers accounted for all of the disagreements (Speaker 31, 35, 37, and 40). These 4 speakers were all diagnosed as having a mixed spastic-flaccid dysarthria. Further, all 4 samples were informally judged to contain opposite features in different parts of the sample. For example, a sample may have included breathy voice near the beginning and fry near the end; therefore, raters may have based their rating on different parts of the sample. The conversational speech sample used in the present study may also have contributed to these disagreements (see Effect of Speech Task section). In the present study, disagreements on these samples were judged not to be discouraging.

The likelihood that individual raters would agree with one another was calculated for each specific speaker. Kreiman and colleagues (1993, 1998) have argued that this method of calculation provides a more accurate picture of variability in individual listeners' internal standards for each of the perceptual features being rated. Probabilities of exact agreement or agreement within 1 scale value were compared to chance (Figures 2 and 3); all probabilities were above chance based on a 7-point scale. Exceeding chance, however, does not demonstrate an adequate level of listener agreement, particularly for a clinical tool. Large variability in calculated probabilities was found along the length of the scale, with values higher near the endpoints and lowest in the mid-range. This suggests that listeners agreed on what constituted "normal" or a "severe deviation from normal" but did not agree about the extent of mild-to-moderate deviation. Lower probabilities near the midpoints of the scale have been reported previously for ratings of voice quality and led the authors to question the validity of quality ratings (Kreiman & Gerratt, 1998; Kreiman et al., 1993). This finding seems to reflect poor interrater reliability or agreement; however, Shrivastav, Sapienza, and Nandur (2005) have suggested that this increased variability near the midpoints of the scale can be explained, at least partly, by the psychometric properties of the scale. The rating scale developed by Darley and colleagues (1969a, 1969b) is an ordinal scale—at least, as listeners use it. Ordinal scales are ordered in the sense that higher numbers represent higher values (except for the values of pitch and loudness in the present study). The intervals between the numbers, however, are not necessarily equal, even though the descriptor "equal-appearing intervals" may be applied to the scale. Therefore, the intervals between points on the scale do not necessarily translate to equal intervals in terms of perceptual magnitude. This may be especially true in the mid-region of the scale, where Shrivastav and colleagues (2005) have argued that stimuli may be perceptually more similar than those at the scale endpoints.

A more stringent approach to determining if agreement is above chance level, which would bolster confidence

in the level of interrater agreement calculated, would be to calculate the probability of chance agreement on fewer scale categories than were available to the raters (i.e., calculate chance agreement for a 5-point scale rather than a 7-point scale; Tinsley & Weiss, 2000). Using this method, chance was calculated as .20 for exact agreement and .52 for within 1 scale value. The lowest likelihood of 2 raters agreeing exactly for an individual speaker and perceptual feature was .33, above chance level. For agreement within 1 scale value, less than 1.5% of the samples had minimum probability levels that fell below chance level (range = .46–.49). The remaining 98.5% of the samples had probabilities above chance. The samples that were below chance level all had a mean rating between 4.0 and 4.5. Results should again be interpreted with appropriate caution, as the level of interrater agreement calculated in the present study indicates that listeners were generally able to agree in their assignment of a rating to a given perceptual feature, for purposes of identification of salient features but not for differential diagnosis of a dysarthria type.

Differences in the level of listener reliability and agreement among the 38 perceptual features have been reported previously for the Mayo Clinic rating scale (Zeplin & Kent, 1996); this is consistent with studies on perceptual ratings of voice quality (Kreiman & Gerratt, 1998; Kreiman, Gerratt, & Berke, 1994; Kreimen et al., 1993). Based on these reports, it was surprising that no statistically significant differences in levels of listener agreement between perceptual features were found in the present study. This was also surprising given the differences in the perceptual features that listeners were asked to rate. For example, several features involved countable attributes (e.g., pitch breaks, inappropriate silences); others involved ratings of severity (imprecise consonants, distorted vowels); and two scales asked listeners to rate more global perceptual impressions (intelligibility, bizarreness). The design of the present study may have contributed to lack of significant difference between perceptual features. Presentation of samples in the present study was blocked—in other words, listeners rated all speakers on a given feature before proceeding to the next feature. Several previous studies have suggested that listener internal standards for perceptual ratings are unstable and are influenced by other factors within and outside the acoustic characteristics of the speech sample being rated (Kreiman et al., 1993, 1994; Kreiman & Gerratt, 1998; Mackey, Finn, & Ingham, 1997; Shrivastav et al., 2005; Southwood & Weismer, 1993). This could include perceptual tradeoffs that listeners make when deciding what rating to use for individual features. Two features, such as harsh voice and hoarse voice, when rated at the same time, probably impact the each other's rating (Karnell et al., 2007). The blocked design, however, forced listeners to rate each feature independently, without regard for other related features. Because the primary interest for the present study was whether points along the scale were defined similarly across a group of listeners, this "blocked" approach was considered reasonable. This design, however, which is not typical of how ratings are done in a clinic situation, limits generalizations about the relation between perceptual features.

Effect of Speech Task

Previous studies of auditory-perceptual scaling methods have included multiple speech tasks, including vowel prolongation, syllable repetition, and passage reading. The current study focused exclusively on ratings based on a conversational speech sample. Results demonstrated that listeners were able to rate deviant perceptual features at levels above chance using this type of speech sample. Ratings of intelligibility and bizarreness, which represent listeners' overall or general impression, had the lowest calculated p-exact and p-one values. It is known from previous research that there is considerable variability in listener judgments of global features of speech, so this finding is not surprising (Southwood & Weismer, 1993). The relatively short duration of the sample (40 s) and /or the varied topics of conversation discussed by the speakers may also have influenced these ratings. It is possible that listener agreement would have improved with the inclusion of other speech tasks, although there are conflicting reports on whether differences exist between ratings based on connected speech versus sustained vowels for voice quality (de Krom, 1995; Revis, Giovanni, Wuyts, & Triglia, 1999; Wolfe, Cornell, & Fitch, 1995; Zraick, Wendel, & Smith-Olinde, 2005). The results of the present study suggest that use of even brief conversational speech samples in assessment can help clinicians identify perceptual features that play prominently in the presenting speech disorder.

Conversational speech is informationally rich in that it may be the only task other than reading that allows for the rating of all 38 perceptual dimensions used in the Mayo Clinic rating system. A recent acoustic study demonstrates that a brief conversational sample (2 min) is sufficient to distinguish hypokinetic dysarthria associated with Parkinson disease from normal speech (Rosen, Kent, Delaney, & Duffy, 2006). It appears that the fingerprint of dysarthria is expressed, to a significant degree, in relatively short samples of conversation. This is not to deny the importance of other tasks in the assessment of dysarthria but, rather, to emphasize the informational value of conversational samples.

Effects of Listener Experience

Two groups of listeners were included in the present study: listeners with limited experience and listeners

with more than 7 years' clinical experience working with individuals with dysarthria. These groups were included based on previous reports showing large differences in listener groups as a function of experience (Zyski & Weisiger, 1987). Zeplin and Kent (1995) included only experienced listeners in their study but reported that this level of experience likely contributed to the higher intra- and interreliability they found compared with Zyski and Weisiger (1987). In the present study, no statistically significant differences between listener groups were found for either intrarater reliability or interrater agreement. The lack of differences does not establish that the perceptual rating skills of the two groups were equivalent; rather, any differences between the groups were probably hidden by the variability related to the rating task. This variability may have included differences between speakers being rated, listener factors (sensitivity, bias, error), task factors (scale resolution, context effects), and interaction effects, among others (for a discussion, see Shrivastav et al., 2005). In addition, it is possible that the 5 hr of training on perceptual ratings of dysarthria that the inexperienced listeners received as part of their classroom training could have equalized the two listener groups in terms of rating skills.

Interestingly, none of the studies of reliability or agreement of auditory-perceptual ratings of dysarthric speech—the present one included—made any formal attempt to "train" listener reliability or agreement. This lack of training is likely reflected in some of the indices reported. A process of formally training listeners should rely on mutual discussion of representative sample cases to develop consensus about deviant features and magnitude of severity ratings on the 7-point scale. This is especially important given the complex acoustic features that listeners are asked to rate using on a perceptual severity scale. An individual's experience does not automatically guarantee interrater reliability of such judgments. Future investigations are needed to examine the benefits of training on auditory-perceptual ratings of dysarthric speech.

Future research efforts should focus on establishing how the Mayo Clinic rating scale is used clinically. If, indeed, current clinical practice—as suggested by Duffy and Kent (2001) and Duffy (2005)—is for clinicians to rely on a pattern of deviant perceptual features rather than a rating of all 38 features, then listener agreement in terms of exact ratings may not be crucial to clinical practice, and studies should focus on agreement for listener identification (presence vs. absence) of deviant perceptual features as a measure of the clinical reliability for the Mayo Clinic rating scale.

Conclusion

Despite its primacy as a tool for the auditory-perceptual assessment of dysarthria, the Mayo Clinic rating scale has not been adequately assessed to determine levels of within- and across-listener agreement for all 38 perceptual dimensions and a heterogeneous group of dysarthric speakers. In accomplishing such an assessment, this study leads to five major conclusions:

1. There were no significant differences in rater agreement for experienced and inexperienced listener groups.
2. There were no significant differences in rater agreement across the 38 perceptual dimensions.
3. Listener agreement was highest at the endpoints of the scale and lowest at the mid-range.
4. The fingerprint of dysarthria is generally expressed within a 40-s sample of conversational speech.
5. The agreement and reliability observed in this study are judged to be adequate for purposes of clinical assessment and research aimed at identifying perceptual features that play a prominent role in the presenting speech disorder.

Acknowledgment

This work was supported in part by National Institutes of Health Grants NIH R01 DC00319, awarded to Ray Kent, and NIH R03 DC005902, awarded to Kate Bunton.

References

American Speech-Language-Hearing Association. (1997). *Guidelines for audiological screening*. Rockville, MD: Author.

American Speech-Language-Hearing Association. (2003). *Code of ethics*. Retrieved from www.asha.org/docs/html/ET2003-00166.html

Brown, A., & Docherty, G. (1995). Phonetic variation in dysarthric speech as a function of sampling task. *European Journal of Disorders of Communication, 30,* 17–35.

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

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969b). Clusters of deviant speech features in the dysarthrias. *Journal of Speech and Hearing Research, 12,* 462–496.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975a). *Motor speech disorders*. Philadelphia: W. B. Saunders.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975b). *Audio seminars in speech pathology: Motor speech disorders*. Philadelphia: W. B. Saunders.

de Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38,* 794–811.

Duffy, J. (2005). *Motor speech disorders: Substrates, diferential diagnosis, and management*. St. Louis, MO: Mosby.

Duffy, J., & Kent, R. (2001). Darley's contributions to the understanding, differential diagnosis, and scientific study of the dysarthrias. *Aphasiology, 15,* 275–289.

Gray, W. S. (1936). *Standard oral reading paragraphs: My Grandfather*. Bloomington, IL: Public School Publishing Company.

Karnell, M., Melton, S., Childes, J., Coleman, T., Daily, S., & Hoffman, H. (2007). Reliability of Clinician-Based (CRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21,* 576–590.

Kearns, K., & Simmons, N. (1988). Interobserver reliability and perceptual ratings: More than meets the ear. *Journal of Speech and Hearing Research, 31,* 131–136.

Kent, R. D., & Kent, J. F. (2000). Task-based profiles of the dysarthrias. *Folia Phoniatria et Logopaedica, 52,* 48–53.

Kent, R. D., Kent, J. F., Duffy, J., & Weismer, G. (1998). The dysarthrias: Speech-voice profiles, related dysfunctions, and neuropathology. *Journal of Medical Speech Language Pathology, 6,* 165–211.

Kent, R. D., Kent, J. F., Rosenbek, J., Vorperian, H., & Weismer, G. (1997). A speaking task analysis of the dysarthria in cerebellar disease. *Folia Phoniatrica et Logopaedica, 49,* 63–82.

Kreiman, J., & Gerratt, B. (1998). Validity of rating scale measures for voice quality. *The Journal of the Acoustical Society of America, 104,* 1598–1608.

Kreiman, J., Gerratt, B., & Berke, G. (1994). The multidimensional nature of pathologic voice quality. *The Journal of the Acoustical Society of America, 96,* 1291–1302.

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

Mackey, L., Finn, P., & Ingham, R. (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.

Milenkovic, P. (1994). *CSpeech* [Computer program]. Madison, WI: University of Wisconsin.

Revis, J., Giovanni, A., Wuyts, F., & Triglia, J. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica, 51,* 108–116.

Rosen, K. M., Kent, R. D., Delaney, A. L., & Duffy, J. R. (2006). Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. *Journal of Speech, Language, and Hearing Research, 49,* 395–411.

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

Shrivastav, R., Sapienza, C., & 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.

Southwood, H., & Weismer, G. (1993). Listener judgments of bizarreness, acceptability, naturalness, and normalcy of the dysarthria associated with amyotrophic lateral sclerosis. *Journal of Medical Speech-Language Pathology, 3,* 151–161.

Tinsley, H., & Weiss, D. (2000). Interrater reliability and agreement. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling*. San Diego, CA: Academic Press.

Wolfe, V., Cornell, R., & Fitch, J. (1995). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9,* 297–303.

Young, M. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research, 36,* 644–656.

Zeplin, J., & Kent, R. D. (1996). Reliability of auditory perceptual scaling of dysarthria. In D. Robin, K. Yorkson, & D. R. Buekelman (Eds.), *Disorders of motor speech: Recent advances in assessment, treatment, and clinical characterization*. Baltimore: Paul H. Brookes.

Zraick, R., Wendel, K., & Smith-Olinde, L. (2005). The effects of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574–581.

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

Received August 3, 2005

Revision received February 20, 2006

Accepted May 28, 2007

DOI: 10.1044/1092-4388(2007/102)

Contact author: Kate Bunton, who is now with the Department of Speech, Language, and Hearing Sciences, University of Arizona, P.O. Box 210071, Tucson, AZ 85721-0071. E-mail: bunton@u.arizona.edu.

John C. Rosenbek is now with the University of Florida in Gainesville.