discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253–260.
REED, C., DURLACH, N., & BRAIDA, L. (1982). Research on tactile communication of speech: A review. *ASHA Monographs, 20,* 1–23.
SAUNDERS, F. (1974). Electrocutaneous displays. In F. A. Geldard (Ed.), *Cutaneous communications systems and devices* (pp. 20–26). Austin, TX: Psychonomic Society.
SAUNDERS, F. (1985, November). *Wearable multichannel electrotactile sensory aids.* Paper presented at the Second Tactual Communications Conference, Wichita, KS.
SHERRICK, C. (1984). Basic and applied research on tactile aids for deaf people: Progress and prospects. *Journal of the Acoustical Society of America, 75,* 1325–1342.
SPARKS, D., KUHL, P., EDMONDS, A., & GRAY, G. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech. *Journal of the Acoustical Society of America, 63,* 246–257.
SZETO, A. (1982). Electrocutaneous code pairs for artificial sensory communication systems. *Annals of Biomedical Engineering, 10,* 175–192.
YENI-KOMSHIAN, G., & GOLDSTEIN, M. (1977). Identification of speech sounds displayed on a vibrotactile vocoder. *Journal of the Acoustical Society of America, 62,* 194–198.

Received March 19, 1987
Accepted July 30, 1987

Requests for reprints should be sent to Rebecca E. Eilers, University of Miami, Mailman Center for Child Development, Miami, FL 33101.

Journal of Speech and Hearing Research, Volume 31, 131–136, March 1988

Research Note

INTEROBSERVER RELIABILITY AND PERCEPTUAL RATINGS: MORE THAN MEETS THE EAR

KEVIN P. KEARNS
VA Medical Center (126), North Chicago, IL

NINA N. SIMMONS
Touro Infirmary, New Orleans, LA

The purpose of this study was to examine the reliability of ratings of perceptual characteristics for 10 ataxic dysarthric subjects. The influence of the occurrence of "deviant" speech parameters on the calculation of reliability coefficients was also explored. Results indicated that overall interobserver agreement levels for minimally trained judges compared favorably to reliability coefficients reported in previous studies. Furthermore, levels of overall agreement were above levels of agreement expected on the basis of chance alone.

In contrast to overall interobserver agreement, much lower levels of interobserver agreement were obtained when "occurrence reliability" coefficients were calculated for deviant dimensions alone. However, occurrence reliability coefficients surpassed the level of agreement expected on the basis of chance alone for all subjects. Based on the results of this investigation, recommendations are made for modifying standard practices for obtaining interobserver reliability for perceptual ratings of speech characteristics.

Despite recent advances in the instrumental evaluation of communicative impairments, perceptual analysis of speech and language problems remains a primary tool for differential diagnosis and clinical management (Darley, 1984). The importance of the clinician's ability to discriminate normal from pathological states has been emphasized in all major content areas of speech pathology, including articulation, voice, fluency, language development, and the adult neuropathologies. It is surprising, therefore, that so little investigative attention has been given to the reliability of listener judgments and the need for specialized training in this area. Although clinical researchers have examined the reliability of perceptual judgments for trained or expert observers (Darley, Aronson, & Brown, 1969a, 1969b; Ludlow & Bassich, 1983, 1984), little information is available regarding perceptual judgments in the clinical setting.

The critical need for research that focuses on the reliability of our perceptual measures is supported by recent findings in the applied behavioral literature (McReynolds & Kearns, 1983). Hopkins and Hermann (1977), for example, have demonstrated that reliability coefficients are integrally related to the frequency of occurrence of target behaviors. Thus, high or low rates of production of clinically relevant behaviors, in a given sample, can artificially inflate reliability estimates. In some cases the response rate for deviant parameters can dramatically affect the probability of observer agreement resulting from chance alone. For example, even though Darley et al. (1969a, 1969b) rated nearly 40 perceptual characteristics in their dysarthria research, approximately one-fourth of these parameters were rated as deviant, whereas the remaining perceptual characteristics were judged to be within normal limits. Consequently, reliabil-

0022-4685/88/3101-0131$01.00/0

ity judges may have easily agreed on ratings of the 30 or so "normal" perceptual characteristics and had, perhaps, more difficulty reaching agreement on ratings of the relatively few deviant characteristics present in Darley et al.'s speech samples. As a result of the high percentage of essentially normal dimensions, it is possible that the overall reliability coefficients reported in such studies are artificially inflated. Given our reliance on perceptual analyses, and the fact that overall reliability of ratings for aberrant speech characteristics may be inflated by subjects' response rate, additional research is needed to assess the reliability of our perceptual measures.

The present study was designed to examine the reliability of perceptual characteristics of the speech of patients with Friedreich's ataxia. Friedreich's ataxia is a recessively inherited degenerative disease that is characterized by progressive, unremitting ataxia of limbs and gait, muscle weakness, and dysarthria (Barbeau, 1976). Although early cerebellar signs have been stressed in the literature, cerebral and brain stem involvement are also observed in Friedreich's ataxia.

The specific purpose of this investigation was to determine if speech-language pathologists could reliably rate the perceptual characteristics of the speech of dysarthric patients following minimal training. Of equal importance, the influence of the frequency of occurrence of relevant perceptual characteristics on the calculation of interobserver reliability was also assessed. The following questions were posed for the perceptual ratings conducted in this study:

1. Are overall interjudge reliability levels above the levels of agreement expected on the basis of chance alone?
2. What is the level of occurrence reliability for deviant perceptual characteristics?
3. Is the calculated level of occurrence reliability above the level that would be expected on the basis of chance?

METHOD

Subjects

Ataxic subjects. Ten subjects were randomly selected from a larger pool of 23 patients with a confirmed diagnosis of Friedreich's ataxia. Two neurologists completed the Ataxia Rating Scale (Barbeau, 1976) and both agreed on the diagnosis for each subject. The subjects, 6 men and 4 women, ranged in age from 11 to 56 years (M = 28:4 yr). Reported time post onset of ataxia ranged from 5 to 34 years (M = 15:3 yr). The subjects' sentence level intelligibility, as measured by the Assessment of Intelligibility of Dysarthric Speech (Yorkston & Beukelman, 1981) ranged from 16% to 90% with a mean intelligibility rating of 69.8% (Table 1).

Reliability judges. Five experienced speech-language pathologists who were working with neurologically impaired adults served as reliability judges. Clinical experience with dysarthria and related disorders ranged from 6 months to 9 years (M = 6:9 yr). All judges participated in three 1-hr training sessions that oriented them to definitions of the perceptual characteristics of interest (See Perceptual ratings). Speech samples from the audiotape seminar in *Motor Speech Disorders* (Darley, Aronson, & Brown, 1975) were rated and discussed during each training session.

Table 1. Subject Characteristics.

Subject	Age (Years)	Sex	Duration of Friedreich's Ataxia (years)	Ataxia Score* (Max 144)	Intelligibility Score %**
1	17	M	9	62	90
2	17	M	9	55	82
3	22	M	16	50	81
4	41	F	26	76	79
5	21	M	11	81	79
6	11	F	5	68	74
7	24	F	11	79	73
8	56	F	25	75	63
9	30	M	17	85	61
10	45	F	34	91	16
Overall Mean	28.4		16.3	72.2	69.8

*Higher ataxia scores indicate more severe involvement.
**(Yorkston & Beukelman, 1981).

Procedures

The speech samples used in this study were obtained during an ataxia screening clinic. Each subject was tested individually in a clinical examination suite with only the subject and examiner present. Speech samples were recorded on a high quality audio cassette tape recorder (Superscope CD 330) using a lapel microphone. A test battery consisting of both formal and informal assessments was administered. The battery included the following measures: The Assessment of Intelligibility of Dysarthric Speech (Yorkston & Beukelman, 1981); the Motor Speech Evaluation (Wertz et al., 1981); the Complex Ideational Materials subtest of the Boston Diagnostic Aphasia Examination (BDAE, Goodlass & Kaplan, 1972); a spontaneous description of the Cookie Theft picture from the BDAE; and a standard reading sample (Grandfather Passage, Darley et al., 1975). The standard reading passage provided the primary data for the present investigation.

Perceptual ratings. The perceptual characteristics rated in this study closely paralleled those used by Darley et al. (1975) and their colleagues (Appendix II). Throughout this study the perceptual characteristics of interest were organized into related categories on the judges' score sheets. For example, all loudness dimensions, including monoloudness and excessive loudness variation, were listed together on the score sheets. Similarly, judges also rated clusters of parameters relating to pitch, voice quality, nasality/air flow, speech rate, and articulation. Two additional characteristics, overall intel-

ligibility and bizarreness, were also rated. A 7-point severity rating scale (Darley et al., 1969a, 1969b) was used to evaluate the perceptual characteristics. The rating scale extended from 1, representing normal, to 7, representing severe deviation from normal.

The initial 30 s of each subject's reading of the "Grandfather Passage" were randomly dubbed onto an experimental listening tape. The five judges independently rated the samples in a quiet conference room within a speech pathology clinic during the listening sessions. The five judges were not permitted to discuss or share their ratings. Prior to each rating session the judges were given a list of definitions of the perceptual dimensions (Appendix) and they were given several minutes to review the definitions and ask questions regarding terminology. General instructions were also provided to remind the judges of the nature of the rating scales.

During the rating sessions each sample was played six times to permit separate ratings of the six categories of perceptual characteristics. Speech samples were played in their entirety and the judges rated each group of perceptual characteristics sequentially. The judges were allowed as much time as they desired between presentations of the speech samples.

Analyses. To assess interobserver reliability, overall and occurrence agreement levels were calculated along with the levels of chance agreement associated with these measures. During each analysis an "agreement" was tallied if the judges' severity ratings were within one scale value of one another on the severity rating scale. Overall reliability was calculated on the basis of point to point agreement between judges. Pairwise comparisons were made for each of the 40 perceptual characteristics so that a total of 400 data points were used for the computation of overall agreement for each of the 10 dysarthric subjects.

The formula used to calculate overall point to point reliability (R) was:

$$\text{Overall } R = \frac{\text{Total No. Agreements}}{\text{Total No. Agreements + Disagreements}} \times 100 \quad (1)$$

This is, of course, the standard formula used by most investigators to calculate overall agreement for observations by independent judges. To assess the inflationary effect of rate of responding on overall reliability, Hopkins and Hermann's (1977) formula for calculating the level of chance agreement was employed:

$$\text{Chance R} = \frac{(O_1 \times O_2 + N_1 \times N_2)}{(T)^2} \times 100 \quad (2)$$

The O_1 and O_2 in the formula designate the number of occurrences of deviant dimensions recorded by observers 1 and 2 respectively, and T designates the total number of observations. Similarly, N_1 and N_2 refer to nonoccurrence or, in this study, the number of *normal* dimensions recorded. At a minimum, overall reliability coefficients should exceed levels of agreement expected on the basis of chance.

A second method of controlling for the inflationary effects of the high rate of occurrence of normal dimensions is to calculate occurrence reliability and associated levels of chance agreement. The calculation of occurrence reliability was based solely on the perceptual characteristics recorded as deviant by the judges. In effect, occurrence reliability removes agreements on normal dimensions from the reliability computation.

Hopkins and Hermann's (1977) formula for calculating occurrence reliability is:

$$\text{Occurrence R} = \frac{O_1 \text{ and } O_2}{T_0} \times 100 \quad (3)$$

The symbol (O_1 and O_2) designates the number of agreements on deviant dimensions. For our purposes, the T_0 represents the number of agreements on deviant dimensions, plus the number of disagreements on whether or not a dimension was deviant. Occurrence reliability coefficients have only recently been used by applied investigators and guidelines have not been firmly established for determining the level of occurrence reliability that is acceptable in clinical research. One can, however, calculate the level of agreement expected *on the basis of chance* as a minimum standard of acceptability. Hopkins and Hermann's (1977) formula for computing chance agreement for occurrence reliability is:

$$\text{Chance Occurrence R} = \frac{(O_1 \times O_2)}{(T)^2} \times 100 \quad (4)$$

RESULTS AND DISCUSSION

The results of overall reliability and associated levels of chance agreement are presented in Table 2. Overall agreement among the five judges on the 40 dimensions ranged from 60% to 90% for the 10 ataxic subjects. The mean overall reliability level of 82% compares favorably with the 84% interjudge agreement reported for expert judges (Darley et al., 1969). At this level of analysis it appears that experienced clinicians can reliably rate the speech characteristics of patients with Friedreich's ataxia following a minimum amount of specialized training.

TABLE 2. Occurrence interjudge reliability and levels of chance agreement for perceptual analyses of the speech characteristics in Friedreich's ataxia (R = reliability).

Subject	*Overall Reliability (%)*	*Chance Overall R (%)*	*% Above Chance*
1	87	62	15
2	90	68	22
3	81	56	25
4	87	57	30
5	78	52	26
6	87	60	27
7	81	53	28
8	84	61	23
9	83	54	29
10	60	50	10
Overall Mean	82	57	24

A comparison of the overall reliability estimates and their associated levels of chance agreement reveals that the reliability coefficients were from 10% to 30% above chance for these subjects (Table 2, Column 3). These data indicate that the judges were reliable despite any deleterious effects of the disproportionate number of normal perceptual characteristics in our samples.

The second level of analysis, calculation of occurrence reliability and associated levels of chance, was undertaken to examine more closely those characteristics that judges agreed were aberrant. Results from these calculations are presented in Table 3. The levels of agreement reached for the deviant perceptual characteristics of the dysarthric samples were below levels that most clinical researchers find acceptable. Specifically, occurrence reliability ranged from 49% to 75% with a mean occurrence agreement level of 68% (Table 3, Column 1). Given the infrequent use of occurrence reliability in speech-language pathology, it is necessary to interpret these results with caution. Additional experience with this approach to reliability assessment for specific populations and behaviors is needed before we can firmly establish acceptable agreement levels for clinical research in speech-language pathology. Occurrence reliability estimates are generally lower than overall reliability coefficients and we may need to adopt new standards and guidelines for interpreting such data.

Despite our need for cautious interpretation, the occurrence reliability data indicate that additional training may have been useful for the clinicians who rated our speech samples. Perhaps more importantly, the occurrence reliability data revealed potentially serious problems relating to agreement levels for deviant characteristics that were not revealed by the overall reliability analysis. These data demonstrate that use of overall agreement as the primary or sole measure of reliability may mask lower agreement levels for deviant speech and language behaviors. Needless to say, low levels of agreement on impaired behaviors could lead to the misdiagnosis and mismanagement of communicatively impaired patients.

The final result of this study concerns the chance occurrence agreement levels obtained for each subject. Occurrence reliability coefficients were from 25% to 69% above the levels of chance agreement (Table 3, Column 3). Thus, interjudge agreements for deviant perceptual characteristics were well above minimal standards of acceptability. The clinical significance of this finding is somewhat diminished by the generally low levels of occurrence reliability reported earlier.

TABLE 3. Occurrence reliability and levels of chance agreement for perceptual analyses of the speech characteristics in Friedreich's ataxia.

Subject	*Occurrence Reliability (%)*	*Chance Overall R (%)*	*% Above chance*
1	75	6	69
2	68	4	64
3	65	11	54
4	71	10	61
5	71	16	55
6	76	7	69
7	68	14	54
8	63	7	56
9	72	12	60
10	49	23	57
Overall Mean	68	11	57

Overall percentage agreement provides an inadequate means of inferring the accuracy of judges' observations on perceptual ratings (Hopkins & Hermann, 1977; Kearns, 1981; McReynolds & Kearns, 1983). Agreement levels achieved with overall percentage agreement calculations are artificially inflated by high or low rates of responding and they may, therefore, provide an inaccurate measure of reliability. Despite these limitations, overall percentage agreement coefficients continue to be a primary means of establishing the reliability of perceptual analyses of speech and language behaviors.

In the present study the overall point to point reliability coefficients were spuriously inflated because fewer than half of the perceptual dimensions were perceived as aberrant. These data support the need to resort to alternative and augmentative reliability measures when reporting perceptual ratings of speech characteristics of dysarthric and other communicatively impaired patients.

Overall percentage agreement measures should, at a minimum, be supported by calculations of associated levels of chance agreement. In addition, whenever the response rate for target behaviors is significantly above or below the 50% level, occurrence or nonoccurrence reliability and their associated levels of chance agreement should be reported (Hopkins & Hermann, 1977). In the final analysis, the results of this study highlight the need to attend to those behaviors that are most critical in our clinical research activities: those dimensions that are deviant.

ACKNOWLEDGMENTS

This research was supported by Veterans Administration Research funds. The helpful assistance of the Speech-Language Pathology Staff of the New Orleans Veterans Administration Medical Center is gratefully acknowledged.

REFERENCES

BARBEAU, A. (1976). Friedreich's ataxia—An overview. *Le Journal Canadien des Sciences Neurologiques, 3,* 302–318.

DARLEY, F. L. (1984). Perceptual analysis of the dysarthrias. In J. C. Rosenbek (Ed.), *Current views of dysarthria. Nature, assessment and treatment: Seminars in speech and language* (pp. 267–278). New York: Thieme-Stratton.

DARLEY, F. L., ARONSON, A., & BROWN, J. R. (1975). *Motor speech disorders.* Philadelphia: W. B. Saunders.

DARLEY, F. L., ARONSON, A. E., & BROWN, J. R. (1969a). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research, 12,* 462–497.

DARLEY, F. L., ARONSON, A. E., & BROWN, J. R. (1969b). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research, 12,* 246–269.

GOODGLASS, H., & KAPLAN, E. (1972). *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger.

HOPKINS, B. L., & HERMANN, R. J. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10,* 2141–2150.

KEARNS, K. P. (1981). Interobserver reliability procedures in applied asphasia research: A review with suggestions for change. In R. H. Brookshire, (Ed.), *Clinical Aphasiology Conference Proceedings* (pp. 26–34). Minneapolis: BRK Publishers.

LUDLOW, C. L., & BASSICH, C. J. (1983). The results of acoustic and perceptual assessments of two types of dysarthrias. In W. Berry (Ed.), *Clinical dysarthria* (pp. 121–154) San Diego: College-Hill Press.

LUDLOW, C. L., & BASSICH, C. J. (1984). Relationships between perceptual ratings and acoustic measures in hypokinetic speech. In McNeil, M., Rosenbek, J., Aronson, A. E. *The dysarthrias: Physiology, acoustics, perception, management* (pp. 163–196). San Diego, CA: College-Hill Press.

MCREYNOLDS, L. & KEARNS, K. P. (1983). *Single-subject experimental designs in communicative disorders.* Austin, TX: Pro-Ed.

YORKSTON, K., & BEUKELMAN, D. (1981). *Assessment of intelligibility of dysarthric speech.* Tigard, OR: C. C. Publications.

WERTZ, R. T., COLLINS, M. J., WEISS, D., KURTZKE, J. F., FRIDEN, T., BROOKSHIRE, R. H., PIERCE, J., HOLTZAPPLE, P., HUBBARD, D. J., PORCH, B. E., WEST, J. A., DAVIS, L., MATOVICH, V., MORLEY, G. K., & RESSURRECCION, E., (1981). Veterans Administration cooperative study on aphasia: A comparison of individual and group treatment. *Journal of Speech and Hearing Research, 24,* 580–594.

Received July 31, 1985
Accepted May 22, 1987

Requests for reprints should be sent to Kevin P. Kearns, Ph.D., Chief, Audiology/Speech Pathology Service, VA Medical Center 126, 3001 Green Bay Rd., North Chicago, IL 60064.

APPENDIX

DIMENSIONS USED IN MAYO CLINIC DYSARTHRIA STUDY (DARLEY, ARONSON, & BROWN, 1975)

Dimension	*Description*
Pitch Level	Pitch of voice sounds consistently too low or too high for individual's age and sex.
Pitch Breaks	Pitch of voice shows sudden and uncontrolled variation (falsetto breaks).
Monopitch	Voice is characterized by a monopitch or monotone. Voice lacks normal inflectional changes. It tends to stay at one pitch level.
Voice tremor	Voice shows shakiness or tremulousness.
Monoloudness	Voice shows monotony of loudness. It lacks normal variations in loudness.
Excess loudness variation	Voice shows sudden, uncontrolled alterations in loudness, sometimes loud, becoming too loud, sometimes too weak.
Loudness decay	There is a progressive diminution in loudness.
Loudness Level (overall)	Voice is insufficiently or excessively loud.
Harsh voice	Voice is harsh, rough, and raspy.
Hoarse (wet) voice	There is wet "liquid sounding" hoarseness.
Breathy voice (continuous)	Voice is continuously breathy, weak, and thin.
Breathy voice (transient)	Breathiness is transient, periodic, and intermittent.
Strained-strangled voice	Voice (phonation) sounds strained (an apparently effortful squeezing of voice through glottis).
Voice stoppages	There are sudden stoppages of voiced airstream (as if some obstacle along vocal tract momentarily impedes flow of air).
Hypernasality	Voice sounds excessively nasal. Excessive amount of air is resonated by nasal cavities.
Hyponasality	Voice is denasal.
Nasal emission	There is nasal emission of airstream.
Forced inspiration-expiration	Speech is interrupted by sudden, forced inspiration and expiration sighs.
Audible inspiration	There is audible, breathy inspiration.
Grunt at end of expiration	There is a grunt at the end of expiration.
Rate*	Rate of actual speech is abnormally slow or rapid.
Short phrases	Phrases are short (possibly because inspirations occur more often than normal). Speaker may sound as if he has run out of air. He may produce a gasp at the end of a phrase.
Increase of rate in segments	Rate increases progressively within given segments of connected speech.
Increase of rate overall	Rate increases progressively from beginning to end of a sample.
Reduced Stress	Speech shows reduction of proper stress or emphasis pattern.
Variable rate	Rate alternates from slow to fast.
Prolonged intervals	There is prolongation of interword or intersyllable intervals.
Inappropriate silence	There are inappropriate silent intervals.
Short rushes of speech	There are short rushes of speech separated by pauses.
Excess and equal stress	There is excess stress on usually unstressed parts of speech, for example, monosyllabic words and unstressed syllables of polysyllabic words.
Imprecise consonants	Consonant sounds lack precision. They show slurring, inadequate sharpness, distortions, and lack of crispness. There is clumsiness in going from one consonant to another.
Prolonged phonomes	There are prolongations of phonemes.
Repeated phonemes	There are repetitions of phonemes.
Irregular articulatory breakdown	There is intermittent, nonsystematic breakdown in accuracy of articulation.
Distorted vowels	Vowel sounds are distorted throughout their total duration.
Unintelligibility (overall)	This is a rating of overall intelligibility or understandability of speech.
Bizarreness (overall)	This is a rating of degree to which overall speech calls attention to itself because of its unusual, peculiar, or bizarre characteristics.