Reliability and Dispersion of Nasality Ratings

DONALD T. COUNIHAN, PH.D.
WALTER L. CULLINAN, PH.D.
Oklahoma City, Oklahoma

Reliable average scale values of nasality have been obtained in many studies using the psychological scaling method of equal-appearing intervals. In these studies, reliability has generally been evaluated through the use of correlation procedures. Many investigators have also computed a *Q* value, the semi-interquartile range, to measure the dispersion of ratings assigned by the judges to each stimulus. The mean *Q* value has then been reported as an index of interjudge reliability.

The speech stimuli presented to the judges for rating of nasality have typically been of one or more of four types: isolated vowels, CVC syllables, connected speech played forward, and connected speech played backward. In those studies (*10, 12, 14, 15*) in which more than one type of speech stimulus have been presented to the judges, there has been a tendency for the reliability coefficient and mean *Q* value to vary systematically with the stimulus type. There is evidence, for example, that smaller mean *Q* values are obtained for connected speech samples rated during forward play, than for connected speech samples rated during backward play, or for isolated vowels (*10, 12, 14, 15*). Differences in the dispersion of nasality ratings assigned to various types of speech stimuli suggest that the degree to which nasal voice quality can be defined reliably by judges varies as a function of the speech stimulus. In no study, however, have measures of dispersion or reliability been obtained for all four types of speech stimuli produced by the same speakers or rated by the same judges. Moreover, data regarding the reliability and dispersion of ratings of specific vowels, CVC syllables, and connected speech samples within each of the four types of speech stimuli are unavailable.

While mean *Q* values have been reported as measures of the reliability of average scale values of nasality, little interest has been shown in the *Q* values of subject groups or individual subjects. Thurstone and Chave (*17*) used *Q* (in this instance, the interquartile range) as a measure of the ambiguity of the sample, the sample in their case being a statement

The authors are affiliated with the Department of Communication Disorders, University of Oklahoma Medical Center.

This research was supported in part by Public Health Service Research Grant No. NB-04216-02 from the National Institute of Neurological Diseases and Blindness. The writers also wish to acknowledge the services received from the Computer Facility at the University of Oklahoma Medical Center.

about the church. They felt that there must be something about the statements with high Q values which made them more ambiguous and, therefore, made it more difficult for judges to agree on the ratings for these samples than for those samples with low Q values. Nasal speech samples might also be viewed as having differing degrees of ambiguity. That is, there may be characteristics of some speakers producing some samples which make it difficult for judges to agree on ratings of nasality. Nasality ratings, for example, may be influenced by such factors as proficiency of articulation (*8, 10, 14, 18*) and differences in vocal pitch and intensity level (*6, 7*). Moreover, the variety of definitions of nasality and nasality types in the speech literature suggests that cleft palate subjects may not present a single quality disturbance. Judges participating in rating experiments have commented that nasal speech samples seem to differ in kind as well as in severity and that it is difficult to apply a given severity scale with equal force to all samples. If the concept of nasality is unclear to the listener or if it is difficult to isolate nasality from other coexisting characteristics of the speech signal, a high mean Q value would be expected.

The purpose of the present study was to investigate how measures of reliability (reliability coefficient) and dispersion (Q) vary with respect to type of speech stimulus, and to assess the use of the Q statistic as a measure of sample ambiguity.

Procedure

Subjects. The subjects for this study were 20 male and 20 female cleft palate subjects between the ages of 15 and 50 years. All subjects had essentially normal hearing in the speech range in at least one ear. Clinical evaluation showed that few subjects effected an adequate velopharyngeal seal.

Speech Samples. Each subject was required to produce the following speech samples: the vowels /i/, /ae/, /u/, /ɑ/, /ɔ/, and /ʌ/ in isolation, each sustained for three seconds; the consonants /t/, /d/, /s/, and /z/ individually combined with each of the above vowels except /ʌ/ in CVC syllables, a total of 20 syllables per subject, and four short sentences devised by Bryan (*2*). For the purposes of a previous study, the sentences include no nasal consonants. Also for purposes of the previous study, the subjects peaked the production of each speech item at an intensity level of 75 dB SPL at a mouth-to-microphone distance of eight inches.

Rating Procedure. The recorded speech signals were reproduced for judgment of perceived nasality by means of a single-track, high-fidelity tape recorder (Ampex, Model 601) and an amplifier-speaker (Ampex, Model 620). Ratings of degree of nasality for each of the recorded speech samples were made by nine speech pathologists (graduate students and faculty) using a seven-point scale of equal-appearing intervals, with *one* representing the mildest and *seven* the most severe nasality. Judgments of nasality for the vowels and CVC syllables were made with the samples played forward. The sentences were rated twice by the judges, once with the samples played forward and once played backward. Each judge, therefore, rated 34 speech samples for each subject. All the samples representing one type of speech stimulus were presented together for rating. For example, first the judges rated all the samples of /i/, then all the samples of /æ/, et cetera. Within each of these stimulus types the samples were randomized with the male and female samples combined. To provide the judges with a common scaling reference, samples of the same type of speech stimulus, prejudged by the experimenters as representative of the *one* and the *seven* scale values, were played immediately prior to the presentation of each speech sample to be rated. Six of the samples for each stimulus type were randomly selected and were rated a second time following the first rating of each of the 40 samples.

Results and Discussion

Reliability of Ratings and Q Values. For each of the 34 types of speech stimuli (that is, 6 vowels, 20 syllables, 4 sentences played forward, and 4 sentences played backward) a Pearson *r* was computed for the ratings from each possible pair of judges. The mean correlation coefficient for each of the 34 stimulus types is presented in Table 1. The obtained mean *r*s range from .51 to .76. The mean correlation coefficient used here is essentially the same as the intraclass correlation coefficient with the between-judge variance removed (*5*). None of the mean coefficients is high enough to indicate sufficient reliability for the use of a single rating from a single judge as a predictor. Estimates of the reliability coefficients for the mean ratings of the nine judges, obtained using the Spearman-Brown formula, range from .90 to .97. Thus, reliable mean scale values of nasality for all 34 types of stimuli can be obtained from nine judges. This finding of satisfactory reliability for average scale values is in agreement with the findings of other studies summarized in Table 2.

A total of 1,360 *Q* values was computed, one for each of the 34 samples produced by each of the 40 speakers. The obtained *Q* values ranged from .25, the lowest possible value, to 2.06. The distribution is positively skewed with a mean *Q* value of .76 and a median *Q* value of .71. The mean *Q* value for each of the 34 stimulus types is presented in Table 1.

The limitations of *Q* as a measure of agreement among judges have been discussed in several recent articles (*3, 13, 20*). One of these limitations is the difficulty in determining what size *Q* value indicates "satisfactory" reliability. In several studies (*10, 12, 14*) in which various dimensions of speech have been rated using a seven-point equal-appearing intervals scale, mean *Q* values up to and including 1.00 have been interpreted as indicating 'satisfactory' reliability. These studies have also generally reported reliability coefficients in the .90's for the average scale

TABLE 1. Mean interjudge reliability coefficients (Pearson *r*s) and mean *Q* values for the 34 stimulus types.

vowels

	/i/	/ʌ/	/ae/	/ɑ/	/ɔ/	/u/	*mean*
M_r	.59	.56	.51	.70	.63	.64	.60
M_Q	.81	.80	.93	.70	.82	.79	.81

sentences forward

	#1	*#2*	*#3*	*#4*	*mean*
M_r	.76	.71	.74	.71	.73
M_Q	.63	.61	.68	.67	.65

sentences backward

	#1	*#2*	*#3*	*#4*	*mean*
M_r	.62	.67	.70	.51	.62
M_Q	.66	.68	.74	.75	.71

CVC syllables

	/tit/	/taet/	/tɑt/	/tɔt/	/tut/	*mean*
M_r	.67	.66	.60	.63	.72	.66
M_Q	.72	.84	.93	.83	.83	.83
	/did/	/daed/	/dɑd/	/dɔd/	/dud/	*mean*
M_r	.69	.70	.68	.55	.66	.66
M_Q	.78	.75	.82	.90	.70	.79
	/sis/	/saes/	/sɑs/	/sɔs/	/sus/	*mean*
M_r	.73	.70	.72	.68	.70	.71
M_Q	.70	.76	.72	.77	.71	.73
	/ziz/	/zaez/	/zɑz/	/zɔz/	/zuz/	*mean*
M_r	.74	.72	.68	.60	.72	.69
M_Q	.66	.67	.72	.67	.63	.67

values. Interjudge reliability as defined by *Q* and interjudge reliability as defined by a reliability coefficient are, of course, quite different concepts. While *Q*s and *r*s need not be highly correlated, if the dimension being rated is meaningful to the judges, a negative correlation between the two measures might be expected. An inspection of Table 1 indicates that for this study mean *Q* and mean *r* do tend to be somewhat negatively correlated. In a recent study, Cullinan and Counihan (*3*) obtained an intraclass correlation coefficient of .00 with a mean *Q* value of 1.48 for ratings of connected

TABLE 2. Reliability coefficients reported in studies of nasality using seven-point equal-appearing interval scales.

study	*vowels*	*syllables*	*connected speech*		*comments*
			backward	*forward*	
present study	.93	.95	.94	.96	est. *r* for mean ratings
Hess (*6*)	.74	—	—	—	for sum of ratings, with repeated ratings
Lintz and Sherman (*7*)	.89		—	—	for mdn scale values with repeated ratings, vowels and syllables combined
Sherman (*10*)	—	—	.89	.89	for mdn scale values for comparable passages
Spriestersbach (*14*)	—	—	.90	.96	for mdn scale values with repeated ratings
Spriestersbach and Powers (*15*)	.81	—	.97	—	for mdn scale values with repeated ratings
Van Hattum (*18*)	—	—	.91	—	for avg scale values with repeated ratings; used most reliable of judges

speech samples along an equivocal speech dimension ("resiliency") using a seven-point equal-appearing interval scale. It appears, therefore, that one can obtain a mean Q value as low as 1.48 with an almost chance assignment of ratings. The confidence with which one feels that data are 'satisfactorily' reliable might decrease quickly, therefore, with small increases in Q above about 1.00 for a seven-point scale.

The interpretation of a given mean Q value as indicating 'satisfactory' reliability will depend on the use to which the ratings are to be put and the amount of dispersion of ratings which the experimenter is willing to tolerate. One may determine whether an obtained mean Q value is relatively large or small by comparing it to mean Q values reported in other studies of the same speech dimension using similar speech samples and the same number of scale units (*3*). The data summarized in Table 3 indicate that the mean Q values obtained in this study tend to be smaller than comparable mean Q values reported in other studies.

DIFFERENCES AMONG STIMULUS TYPES. A Friedman Two-Way Analysis of Variance was performed to test the significance of differences in rank totals for Q values among seven groups of stimulus types (vowels, sentences forward, sentences backward, four CVC syllables) for each sex. For purposes of this analysis the 20 CVC syllables were arranged in four groups: tVt, dVd, sVs, and zVz. This was done to permit examination of the effect of consonant context on the dispersion of ratings assigned to CVC syllables. For both males and females the differences were

TABLE 3. Mean Q values reported in studies of nasality using seven-point equal-appearing interval scales.

study	*vowels*	*syllables*	*connected speech*		*comments*
			backward	*forward*	
present study	.81	.78	.71	.65	
Lintz and Sherman (*7*)	.99		—	—	vowels and syllables combined
Sherman (*10*)	—	—	.97	.81	
Sherman and Goodwin (*12*)	—	—	.98	.90	
Spriestersbach (*14*)	—	—	1.00 .93	.98 .90	samples rated twice
Spriestersbach and Powers (*15*)	1.07	—	.92	—	

significant beyond the 0.1% level, indicating that the dispersion of ratings is not equal for all stimulus types.

Reliability coefficients and mean Q values obtained in this study are presented in Tables 2 and 3 with comparable measures from other studies. All the investigations cited used taped speech samples and seven-point equal-appearing interval scales. The data evidence a consistent trend over studies for mean Q values to decrease and a nearly consistent trend for reliability coefficients to increase as the stimulus changes from isolated vowel to syllable to connected speech played backward to connected speech played forward. One exception to this trend is reported by Bradford, Brooks, and Shelton (*1*), who obtained higher reliability coefficients for ratings on a vowel test than on connected speech. Their procedures, however, were greatly different from the procedures of the studies cited in Table 2. The systematic changes in the reliability and dispersion of nasality ratings that occur as a function of type of speech stimulus suggest that judges are less successful in rating nasality when consonant cues are minimized or absent. It may also be that attempts to eliminate 'irrelevant' speech dimensions, such as consonant articulation, enhance neither the reliability nor validity of nasality ratings. Whatever nasality is, and as a perceptual construct it seems to be poorly defined, it is apparently scaled with better agreement when consonant cues are present and when consonant-vowel cues are heard in their normal sequence in speech, than when such cues are not available. While it might be argued that such cues contaminate the purity of nasality judgments, it is quite possible that acoustic cues associated with consonants form an important part of the nasality of cleft palate speakers (*19*).

To test the significance of differences in rank totals for Q values within each of the seven groups of stimulus types for each sex, fourteen additional Friedman Two-Way Analyses of Variance were performed. For females, none of the differences were significant at the 5% level. For males, significant differences were obtained among sentences played backward ($P < .01$), vowels in tVt contexts ($P < .01$), and vowels in dVd contexts ($P < .001$). These findings indicate that female subjects received a similar dispersion of ratings for specific speech types within each of the seven groups of stimulus types. Certain speech types produced by males, however, were associated with a greater dispersion of ratings than others. It may be that these differences in Q values reflect a greater difficulty in the perception of nasality in certain phonemes or samples produced by male speakers.

One of the more serious limitations in the use of Q is the fact that there is no available test to determine the value of Q that is needed to indicate greater than chance agreement among judges. Young and Downs (*20*) have proposed the range (R) as an index of agreement and they present information for testing the significance of an obtained range. The R was computed for each of the samples in this study and the test of significance applied. Using the percentage of R values which show significant agreement at the 5% level, results parallel those obtained for reliability coefficients and mean Q values. For example, for the four major types of speech stimuli, vowels, syllables, sentences backward, and sentences forward, the percentages of the R values which show significant agreement were 57%, 68%, 72%, and 76%, respectively. Again, it can be seen that there is a systematic difference in the dispersion of ratings according to the type of speech stimulus. The difficulties that are associated with judgments of isolated vowels are reflected in the fact that 43% of the vowel samples received ranges of ratings that could have occurred by chance. Of equal significance is the finding that, even in ratings of connected speech played forward, 24% of the samples rated evince a range of scale values that could have occurred by chance. That there are problems associated with scaling nasal voice quality can be seen when the R test is applied to ratings obtained in studies of other common speech dimensions. For instance, the R test was applied to the ratings obtained in three previously reported studies: Sherman and Cullinan (*11*), for 50 one-minute connected speech samples rated for articulation defectiveness; Cullinan, Prather, and Williams (*4*), for 27 twenty-second connected speech samples rated for severity of stuttering; and Sansone (*9*), for 40 vowels rated for degree of vocal roughness. The percentages of R values showing significant agreement at the 5% level were 100% for articulation defectiveness, 89% for severity of stuttering, and 95% for vocal roughness. These findings tend to underscore the difficulties inherent in the judgment of nasality.

INDIVIDUAL SUBJECT VARIATION AND STABILITY OF DISPERSION MEASURES. The Mann-Whitney U Test was applied to test for significance the differences between the Q values for male and female speakers for each of the 34 stimulus types. Of the 34 differences tested, only for the syl-

TABLE 4. Coefficients of concordance (corrected for tied ranks), with Q value for the criterion measure, for each of seven stimulus types and for each sex.

	vowels	*sentences*		*tVt*	*dVd*	*sVs*	*zVz*
		forward	*backward*				
males	.31	.56	.68	.55	.54	.53	.45
females	.19	.24	.27	.48	.38	.52	.22

lable /dɔd/ was the difference between the male and female Q values significant (at the 5% level). The difference for /dɔd/ was .25 scale units, all the other differences being .15 or smaller with a mean difference of only .07 scale units. In general, then, it appears that speech samples produced by cleft palate speakers of one sex have no greater dispersion of ratings of nasality than the same sample produced by cleft palate speakers of the other sex.

Using Q value as the criterion measure, a coefficient of concordance (corrected for tied ranks) was obtained for each of the seven groups of stimulus types for each sex. The obtained coefficients are presented in Table 4. These coefficients, the highest of which is only .68, indicate that the subjects in this study tended to rank order themselves, according to degree of dispersion of ratings, quite differently from one vowel to another, from one syllable to another, and from one sentence to another. The male speakers, however, tended to agree more closely in rank order than did the females.

The individual subject variability in Q value might be interpreted as reflecting considerable interaction between subject and speech stimulus type. The picture is confounded, however, by the fact that Q tends to be one of the least stable measures of dispersion (*16*). For the 36 vowel samples for which two ratings were obtained in this study, the mean Q values for the first and second ratings, respectively, were .81 and .73. Pearson rs for the two sets of Q values and for the two sets of median scale values were .61 and .95, respectively. For the 24 sentences-forward samples which were repeated for second ratings, the mean Q values were .61 and .60, with a Pearson r of .44 for the two sets of Q values and .96 for the median scale values. Although the computations were not performed, essentially similar results would be expected for the syllable and sentence-backward samples. In addition, one of the sentences (40 samples) was re-rated during forward play by a different group of 12 judges. While the median scale values for the two ratings yielded a Pearson r of .88, the Q values had an r of only .19, with mean Q values of .63 and .71. It appears, then, that while mean Q values for groups of subjects tend to be relatively stable, individual subject's Q values are not very stable. The use of Q as a measure of the ambiguity of individual nasal samples would, therefore, appear to be limited. Further study of measures of and variables responsible for ambiguity in the perception of nasal speech would appear to be useful.

Summary

Ratings of nasality were obtained from nine judges using a seven-point equal-appearing intervals scale for recordings of each of 6 vowels, 20 CVC syllables, and 4 sentences produced by each of 40 cleft palate speakers. All speech samples were rated during forward play and, in addition, the four sentences were rated during backward play. Median scale values of nasality and Q values, the semi-interquartile ranges, were computed for each of the 34 stimuli for each of the 40 speakers. Reliability coefficients and mean Q values indicate that reliability of ratings of nasality increases and the degree of ambiguity decreases from vowels to syllables to connected speech rated during backward play to connected speech rated during forward play. While mean Q values tend to be relatively stable for groups of subjects, individual subject's Q values are not very stable. The need for further study of the measures of dispersion of ratings is indicated.

reprints: *Dr. Donald T. Counihan*
Speech and Hearing Center
University of Oklahoma Medical Center
825 N. E. 14th Street
Oklahoma City, Oklahoma 73104

References

1. Bradford, L. J., Alta R. Brooks, and R. L. Shelton, Clinical judgement of hypernasality in cleft palate children. *Cleft Palate J., 1,* 329–335, 1964.
2. Bryan, G. A., Relationships among nasal and "oral" sound pressures and ratings of nasality in cleft palate speech. Unpublished Ph.D. dissertation, University of Oklahoma, 1963.
3. Cullinan, W. L., and D. T. Counihan, Some factors affecting the size of Q values for speech ratings. *Percept. motor Skills, 27,* 531–536, 1968.
4. Cullinan, W. L., Elizabeth M. Prather, and D. E. Williams, Comparison of procedures for scaling severity of stuttering. *J. speech hearing Res., 6,* 187–194, 1963.
5. Ebel, R. L., Estimation of the reliability of ratings. *Psychometrika, 16,* 407–424, 1951.
6. Hess, D. A., Pitch, intensity, and cleft palate voice quality. *J. speech hearing Res., 2,* 113–125, 1959.
7. Lintz, Lois, and Dorothy Sherman, Phonetic elements and perception of nasality. *J. speech hearing Res., 4,* 381–396, 1961.
8. McWilliams, Betty Jane. Some factors in the intelligibility of cleft palate speech. *J. speech hearing Dis., 19,* 524–527, 1954.
9. Sansone, F. E., Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult males. Unpublished Ph.D. dissertation, University of Oklahoma, 1969.
10. Sherman, Dorothy, The merits of backward playing of connected speech in the scaling of voice quality disorders. *J. speech hearing Dis., 19,* 312–321, 1954.
11. Sherman, Dorothy, and W. L. Cullinan, Several procedures for scaling articulation. *J. speech hearing Res., 3,* 191–198, 1960.

12. Sherman, Dorothy, and F. Goodwin, Pitch level and nasality. *J. speech hearing Dis., 19,* 423–428, 1954.
13. Silverman, F. H., Interpretation of mean *Q* values for sets of stimuli rated on a seven-point equal-appearing interval scale. *Percept. motor Skills, 24,* 842, 1967.
14. Spriestersbach, D. C., Assessing nasal quality in cleft palate speech of children. *J. speech hearing Dis., 20,* 266–270, 1955.
15. Spriestersbach, D. C., and G. R. Powers, Nasality in isolated vowels and connected speech of cleft palate speakers. *J. speech hearing Res., 2,* 40–45, 1959.
16. Tate, M. W., *Statistics in Education.* New York: Macmillan, 1955.
17. Thurstone, L. L., and E. J. Chave, *The Measurements of Attitude.* Chicago: The University of Chicago Press, 1929.
18. Van Hattum, R. J., Articulation and nasality in cleft palate speakers. *J. speech hearing Res., 1,* 383–387, 1958.
19. Westlake, H., and D. Rutherford, *Cleft Palate.* New Jersey: Prentice-Hall, 1966.
20. Young, M. A., and T. D. Downs, Testing the significance of the agreement among observers. *J. speech hearing Res., 11,* 5–17, 1968.