Quanti-Scaling of Articulatory Defectiveness in Cleft Palate Speakers

DONALD A. HESS, ED.D.
Buffalo, New York 14222

In recent research (*3*) a new method of scaling articulatory defectiveness for cleft palate speakers was developed. This method, describable as quanti-scaling, allows ratings of articulatory proficiency on a *one* to *seven* scale. Although resultant ratings are similar to ratings obtainable from the equal-appearing-intervals rating procedure (*7*), the method used to obtain the ratings is entirely different and novel. Designed originally as a test of attitude, the equal-appearing-intervals rating procedure of Thurstone and Chave (*7*) provided descriptive terms for rating a given attitude on any one of several points separated by equal-appearing-intervals on a continuum scale. In past research this rating procedure has been utilized as a seven point scale for a number of studies of nasality, most of which used cleft palate speakers as subjects (*1*). It has provided a basis for assessment of harshness among non-cleft palate speakers (*5*) as well as harshness, breathiness, and hoarseness among cleft palate speakers (*2*). As a nine point scale, the method of equal-appearing-intervals rating has also been used to assess normal and articulatory defective speakers (*4*).

The quanti-scaling technique for rating articulatory defectiveness of cleft palate speakers (*3*) uses a different approach from the descriptive guidelines of the equal-appearing-intervals rating procedure. This new rating technique was developed experimentally as an attempt to obtain ratings of articulatory proficiency on the basis of explicit, less ambiguous guidelines than are available in conventional procedures (*7*). It was reasoned that improved criteria for rating articulatory proficiency might necessitate use of fewer judges (assuming proper qualification and training) than classic scaling procedures, which employ descriptive terms (*4, 6*), usually employ. A discrete numerical rating scale ranging from *one* to *seven* is employed. In the original procedure (*3*) each test sentence to be rated contained three consonants—/p/, /t/, and /k/ or /f/, /s/, and /tʃ/—for quanti-scaling. In each test sentence as it appeared on the rating sheet, each of the three appropriate consonants was underlined as a reminder to judges of the delimitation of the judgments. If all three conso-

Donald A. Hess, Ed.D., is Professor of Communication Disorders, State University College at Buffalo, Buffalo, New York.

This research was partially supported by a JAC/UAC faculty research fellowship-grant in aid award through the Research Foundation of the State University of New York.

nants were judged as correct, a rating of *one* was assigned to the sentence. For each consonant characterized by perceptible nasal emission, slighting, or slight distortion, a value of *one* was added. However, for each consonant judged to have lost phonemic entity because of omission, glottal substitution, velar or pharyngeal articulation, lateral emission, or otherwise severe distortion, a value of *two* was added. In this manner of rating, any given sentence could be rated from *one* to *seven,* depending on whether each of the three test consonants was correctly articulated or misarticulated (with degree of misarticulation evaluated for each consonant).

Median ratings of four qualified judges were employed as criterion measures in the original study (*3*). Although reliability for 40 repeated ratings was satisfactory (Pearson *r* of .91), a t of 2.00 (P, .05 level) for mean differences in 40 ratings and repeated ratings might raise a question regarding relative stability of the ratings. The mean Q value, an indication of dispersion of individual ratings and thus interjudge agreement, was 1.14. This mean Q value is considerably higher than a comparable measure (.78) reported by Spristersbach (*6*) for ratings of defectiveness of articulation of 50 cleft palate children by 38 judges. However, it is comparable to mean Q values reported by Morrison (*4*), who used 40 naive observers and 12 trained judges to rate severity of articulatory defectiveness of noncleft palate speakers on a nine point scale. Both of these studies (*4, 6*) employed equal-appearing-intervals rating procedures.

Further study of the quanti-scaling method with additional qualified judges appeared advisable. As in the original study (*3*), only well trained and qualified judges were desired, since the rating for each test sentence involves three separate ratings for individual consonants, quanti-scaled. With only five seconds available for judgment of each test sentence, the procedure is demanding and tiring. Of particular interest in this extended study were (1) replicability of criterion measures; (2) effect of number of judges on mean ratings; (3) relative stability of ratings over the entire judgment task; (4) effect of number of judges on mean Q values; and (5) replicability of main effects originally reported (*3*).

Procedures

The original procedures (*3*) for obtaining quanti-scaled ratings of articulatory proficiency from the four original judges were repeated for seven additional qualified judges (making data available for a total of 11 judges). These judges, five of whom were employed as speech clinicians in the 1971 Day Care-Residential Summer Cleft Palate Speech Program at the State University College at Buffalo, either achieved the master's degree in speech pathology or received this degree within one month of the time of the study. From the pool of eleven (four original plus seven additional) judges, five panels of judges were determined by random

elimination: 11 judges, nine judges, seven judges, five judges, and three judges. Median ratings for the entire sample of 320 sentences, and Pearson *r*s and t-tests for 40 repeated judgments were obtained for each of these judgment panels. A matrix of intercorrelations (Pearson *r*s) was computed for the five judgment panels. Mean judgments for every 80 measures of the 320 sentence sample were computed for the four original judges, seven additional judges, and total 11 judges. General means were obtained for each judgment panel for the entire 320 sentence sample, and mean Q values were computed for panels of five, seven, nine and 11 judges.

Results

Table 1 shows the mean ratings of articulatory proficiency by quanti-scaling for the five judgment panels. These measures varied from 4.04 to 4.07 and were not significantly different by t-test. They are somewhat higher than the mean ratings of the original four judges, 3.72.

How replicable are median ratings of articulatory proficiency by quanti-scaling technique for various numbers of judges? Table 2 provides a matrix of intercorrelations for median ratings among the various judgment panels. Correlations vary from .90 to .97, with highest agreement among panels of seven or more judges (*r*s of .95 – .97).

How well did each judgment panel agree with itself on 40 consecutive rejudgments? Did judgments remain stable on this test? Table 3 shows that panels of seven to 11 judges were comparably reliable, with *r*s of .93. Three to five judges were not much lower, with *r*s of .91. However, like the original four judges, each one of these judgment panels shifted significantly in its judgment of articulatory proficiency, with higher ratings on rejudgment (see Table 4). The mean differences were least for three

TABLE 1. Mean ratings of articulatory proficiency by quanti-scaling for various judgment panels.

	3 judges	*5 judges*	*7 judges*	*9 judges*	*11 judges*
mean ratings.....	4.05	4.07	4.07	4.05	4.04

TABLE 2. Matrix of Pearson *r*'s for median ratings of articulatory proficiency among the various judgment panels.

	11 judges	*9 judges*	*7 judges*	*5 judges*
9 judges.....	.97			
7 judges.....	.95	.95		
5 judges.....	.93	.93	.92	
3 judges.....	.91	.91	.91	.90

TABLE 3. Pearson *r*'s for repeated ratings of 40 test sentences by five judgment panels.

	3 judges	*5 judges*	*7 judges*	*9 judges*	*11 judges*
Pearson *r*	.91	.91	.93	.93	.93

TABLE 4. Mean differences for repeated ratings of 40 test sentences by five judgment panels.

	3 judges	*5 judges*	*7 judges*	*9 judges*	*11 judges*
mean difference	.25	.37	.40	.35	.40

judges (.25) and greatest for seven and 11 judges (.40). All differences were significant by t-test ($P < .05$).

These findings might suggest that the procedure for training and periodic retraining of the 11 judges (four original and seven additional judges) does not insure adequate stability of judgments. Therefore, for every 80 criterion measures of the total sample of 320 sentences, mean ratings of articulatory proficiency were computed for the four original judges, seven additional judges, and the complete judgment panel of eleven judges. The results are shown in Figure 1. The general means for

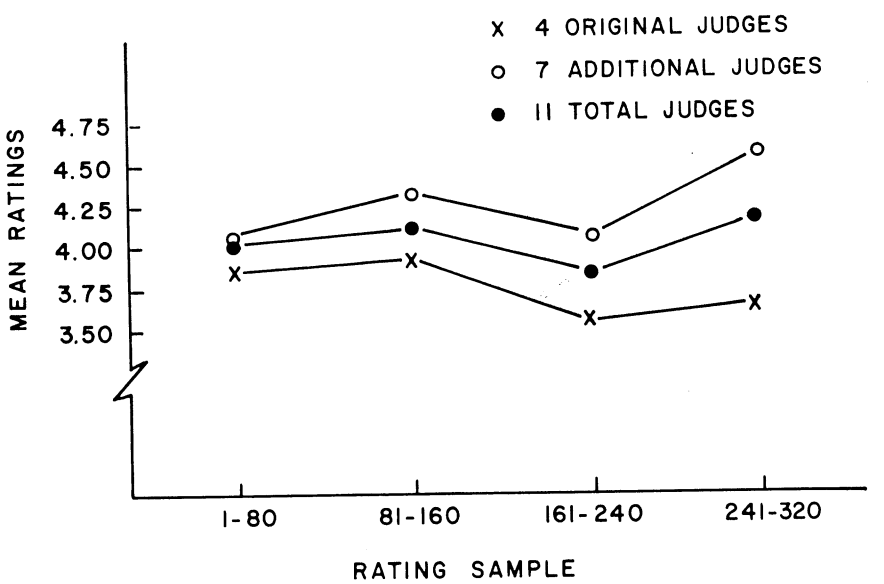

FIGURE 1. Distribution of mean ratings for every 80 measures among 320 sentences in the rating sample. Mean ratings are shown for four original judges, seven additional judges, and the total 11 judges.

TABLE 5. Mean Q values for ratings of articulatory proficiency by quanti-scaling for various judgment panels.

	5 judges	*7 judges*	*9 judges*	*11 judges*
mean Q.....	1.01	.96	.95	1.06

the four original judges (3.72) and seven additional judges (4.28) are markedly different, although the general mean for the total of 11 judges is almost centered on the seven point scale, 4.04. The original four judges tended to drift downward in their ratings in the second half of the total judgment procedure. The additional seven judges remained fairly stable throughout 240 ratings, then drifted markedly higher in their last 80 ratings. The distribution of mean judgments for the original four judges and additional seven judges remained essentially parallel, if increasingly divergent in the latter half of the judgment task. The composite picture for eleven judges shows a rather stable judgment throughout, with slight drift downward in the third quarter of judgments.

Table 5 shows mean Q values for quanti-scaling of articulatory defectiveness for four judgment panels. Surprisingly, the largest mean Q value was obtained from the entire panel of 11 judges, 1.06. Lower mean Q values were earned by panels of five to nine judges, and these ranged from .95 (for nine judges) to 1.01 (for five judges). All of these mean Q values are considerably lower than the comparable measure for the four original judges, 1.14.

The marked difference in mean ratings for the four original judges and the seven additional judges might lead one to question the relative performance of these two groups of judges with respect to the main effects under original study (*3*). Table 6 summarizes mean judgments of the original four judges (*3*) and additional seven judges for each of the main effects under study. The results are quite similar. Both groups of judges showed marked preference for better articulatory ability in stressed syllabic environments and for stop-plosives, among the test conditions. Differences for rhythm and rate, nonsignificant for four judges in the original study (*3*), are similar for the seven additional judges.

Discussion

The quanti-scaling technique for judging articulatory defectiveness of cleft palate speakers appears to be a viable procedure for research purposes. It yields comparable findings, regardless of numbers of judges used, within a range of three to 11 judges. For various judgment panels within this range, mean ratings of articulatory proficiency and distribution of median ratings of articulatory proficiency (the criterion measures) are highly similar. Although highest agreement occurs between panels of nine and 11 judges, lower dispersion of ratings is found for panels of fewer than

TABLE 6. Comparison of original four judges (*3*) and additional seven judges for mean judgments of articulatory proficiency under experimental conditions.

experimental conditions	*mean judgments of articulatory proficiency*	
	original 4 judges	*additional 7 judges*
rhythm		
iambic	3.77	4.27
trochaic	3.68	4.29
rate		
fast	3.84	4.38
slow	3.62	4.18
phoneme type		
stop-plosive	3.33	3.94
fricative-affricate	4.12	4.62
syllabic stress		
stress	3.28	3.81
reduced stress	4.17	4.87
(general means)	3.72	4.28

11 judges, with lowest dispersion for seven to nine judges (mean Q values of .96 and .95, respectively).

Regardless of numbers of judges employed in a range of three to 11 judges, correlations for repeated ratings are high (Pearson *r*s of .91 to .93). However, regardless of numbers of judges employed, significant differences by t-test, with higher ratings in the re-rating procedure, were found. This drawback was not found to be systematically related to stability of judgments in the distributions of the entire sample. Although the four original judges tended to drift toward lower ratings in the second half of the rating task, seven additional judges tended to drift upward in their ratings during the second quarter and particularly the fourth quarter of the rating task. When mean ratings for all eleven judges were considered, essential stability resulted. Tendency for all ratings to be lower in the third quarter of the sample is interpreted as an artifact of the random-rotation procedure for sequencing speakers and test conditions.

Why did the seven additional judges have ratings that averaged so much higher than the ratings of the original four judges? There are two possible reasons: (1) Among the four original judges, only one was a practicing speech clinician at the time of the study. Among the seven additional judges, five persons were involved as clinicians in the summer cleft palate speech program. Active involvement on the therapy level may make for a more critical ear. (2) The ratings from the four original judges were obtained in the evening hours, when ambient noise is minimal. The ratings from the seven additional judges were obtained in late afternoon hours, following the daily therapy schedule. At this time, the ambient noise level probably was a little higher.

For the purposes of the original study (*3*), in terms of main effects, number of judges employed was not too important. Highly similar results were obtained for four original judges and seven additional judges, regardless of sizeable differences in their general means. As previously noted, judgment panels of varying size (by random elimination) yielded highly similar criterion measures. Ideal judgment panel size probably is nine judges, if one wishes the lowest mean Q-value, stable distribution of criterion measures, and replicability of repeated measures. This would not insure against shift in mean ratings for 40 repeated measures, however. Perhaps adjustments in the procedures for training judges could provide this safeguard. However, consideration of results of such t-tests is academic if stable distribution of criterion measures over the entire judgment task is demonstrable.

The success of the quanti-scaling technique for judging articulatory defectiveness in cleft palate speakers would appear to rely heavily upon the use of qualified judges, in terms of training and clinical orientation. Present findings in this regard are in agreement with Counihan and Cullinan (1970), who found that nine highly qualified judges could achieve comparable reliability coefficients and mean Q values with those reported in other studies using 30 or more judges in ratings of nasality on a seven point equal-appearing-intervals rating procedure. Most of these studies employing large judgment panels have partly relied upon judges considerably short of those competencies expected in the minimally trained speech pathologist (e.g., a master's degree or its equivalent). Perhaps research of this type in the future might place more emphasis on clinical qualification of judges and less emphasis on numbers of judges employed.

It is the writer's opinion that the quanti-scaling technique for obtaining ratings of articulatory proficiency among cleft palate speakers is more powerful than the equal-appearing-intervals rating procedure in that it provides more explicit and less ambiguous guides for arriving at ratings. For the purposes intended in research employing the procedure (*3*), the technique yielded criterion measures that far outweighed the restraints that might ordinarily be posed by t-tests (as measures of stability of judgment) and mean Q values (as measures of dispersion of judgments). Although mean Q values of less than 1.00 may be obtained in the procedure with as few as seven to nine judges, it is after all replicability of the *criterion measure* that the researcher is primarily concerned with. Whatever the end result of any particular research effort, the criterion measure is the major determinant in the analysis. Assured replicability of criterion measures, demonstrable in this study, thus relegates mean Q values and results by t-test to a level of secondary importance. By this interpretation, the results of quanti-scaling of articulatory defectiveness by four qualified judges, as described in the original report (*3*) are viewed with confidence.

The quanti-scaling procedure for judging articulatory proficiency ap-

pears to offer a useful new research procedure. Modified as necessary to design of study, it should prove to be effective in assessment of noncleft palate articulatory defective persons as well as cleft palate speakers.

reprints: *Dr. Donald A. Hess*
223 Hennepin Road
Grand Island, New York 14072

References

1. COUNIHAN, D. T., and W. L. CULLINAN, Reliability and dispersion of nasality ratings. *Cleft Pal. J., 7:* 261–270, 1970.
2. HESS, D. A., Pitch, intensity, and cleft palate voice quality. *J. sp. hear. Res., 2:* 113–125, 1959.
3. HESS, D. A., Effects of certain variables on speech of cleft palate persons. *Cleft Pal. J., 8:* 387–398, 1971.
4. MORRISON, SHEILA, Measuring the severity of articulatory defectiveness. *J. sp. hear. Dis., 20:* 347–351, 1955.
5. SHERMAN, DOROTHY and E. LINKE, The influence of certain vowel types on degree of harsh voice quality. *J. sp. hear. Dis., 17:* 401–408, 1952.
6. SPRIESTERSBACH, D. C., Assessing nasal quality in cleft palate speech of children. *J. sp. hear. Dis., 20:* 266–270, 1955.
7. THURSTONE, L. L. and E. J. CHAVE, *The Measurements of Attitude.* Chicago: University of Chicago Press, 1929.