AN ANALYSIS OF SOME OF THE SOURCES OF VARIATION INVOLVED IN RATING SPEECHES

Thesis for the Degree of M. A.

MICHIGAN STATE COLLEGE
Margaret Mary Anderson

1945

This is to certify that the

thesis entitled

An Analysis of Some of the Sources of Variation Involved in Rating Speeches

presented by

Mary Margaret Anderson

has been accepted towards fulfilment of the requirements for

M. A. degree in Education

L. J. Lubker
Major professor

Date September 1, 1945

AN ANALYSIS OF SOME OF THE SOURCES OF VARIATION INVOLVED IN RATING SPEECHES

By

MARGARET MARY ANDERSON

A THESIS

Submitted to the Graduate School of Michigan State College of Agriculture and Applied Science in partial fulfilment of the requirements for the degree of

MASTER OF ARTS

Division of Education

1945

THESIS

ACKNOWLEDGMENT

I wish to express my sincere appreciation to Dr. Paul L. Dressel for his interest, suggestions, and helpful guidance throughout this study.

CONTENTS

TABLES

AN ANALYSIS OF SOME OF THE SOURCES OF VARIATION INVOLVED IN RATING SPEECHES

INTRODUCTION

In connection with the Basic College Written and Spoken English course which was introduced last fall at Michigan State College, a six-hour Comprehensive Examination was given at the close of the fall quarter. A year's credit for freshman English was granted to those students who satisfactorily passed this examination while the other students were obliged to complete their year of work in English and take another Comprehensive Examination at a later date. Since there were some students with one and two terms of English completed under the old program, eligibility for taking this examination was automatically granted such students while students with only one term of English under the new program and with entrance test scores and high school English records meeting the required standards, were granted permission by the Dean upon the recommendation of their counselor or instructor. Of the one hundred and sixty-nine students who took the examination, one hundred and twenty were given full credit for the course.

Speech being one of the important phases of freshman English, the second half of the first test session was devoted to preparing and delivering a two-minute speech on some aspect of the library. To

obtain a random grouping in each of the eight rooms where the speeches were to be given, a card containing a room assignment and a speech number was handed each student as he left the earlier session. Twenty-one such cards had been made previously for each of the eight rooms. A list of suggested topics was distributed as the session began and the students were allowed twenty minutes to prepare their speeches.

Three instructors were assigned to act as faculty raters in each of the eight rooms, this rating being used to decide the student's grade on the speech. Each student was also rated by every other student, but on only one quality, the five qualities being taken in successive order, thus giving approximately four different ratings per quality per room.

Purposes of Study

The purposes of this study are (1) to compare the faculty and student ratings; (2) to study the independence of the five qualities used in rating and the ten-point scale; (3) to determine the reliability of the number of raters; (4) to determine the major sources of variance in ratings; and (5) to make possible suggestions for the improvement of speech testing.

EARLIER STUDIES REVIEWED

Studies of speaking skill in which the participants are brought into the speaking situation are far from numerous. Nichols (7, pp.385-391) developed a written test which appears to correlate more closely with oral performance than other written tests but it was designed more for courses in which the knowledge and application of the principles of speech were the main objectives and no speeches were given. As yet, it is only in the experimental stage as far as actual results are concerned.

Thompson (12, pp.87-91) realized that the accuracy of judgment could be increased by (1) a panel of raters, (2) a training program for increasing raters' skill, and (3) by giving the raters a better yardstick for measuring speaking skill. Turning his attention particularly to the improvement of the third item, he conducted three experiments to determine the accuracy of various rating techniques with the following general conclusions:

1. The grading system and the linear system are approximately equal in accuracy, with the slight margin apparently in favor of the linear scale but not significant statistically. Nine different letter grades were used here, however, and the linear scale included nine points (0 - 8).

2. Comparing the use of letter grades and the Bryan-Wilke Scale, each technique was used with approximately equal degree of accuracy, although the letter system is more practical because of simplicity.

3. Paired-comparisons method of evaluating speaking skill is superior to the rank-order method and should be used when the problem is one of ranking speakers. Because the ratings must be made after all the speeches have been delivered,this method is limited to small groups.

Experiments have brought various results concerning the number of points used, number of raters, and the types of ratings. Guilford (4, pp. 263-283) made the general statement that the number of points used on the scale depends upon the raters, their ability to discriminate, and their motivation in making the ratings. Conklin (1) found that for untrained persons the maximum of five points should be used while Symonds (11, pp. 456-461) states that seven is the optimal number for greatest reliability.

Rugg, (9, pp. 425-438) states that pooled ratings of not less than three independent judges should be used while Symonds (11, pp. 456-461) demands at least eight. Much depends, of course, on the particular trait and the manner of securing ratings.

Symonds (11, pp. 456-461) concludes that the results of ratings are as reliable as those obtained from the ranking method and Conklin and Sutherland (2, pp. 44-57) found ratings were less variable from one judge to another than were rankings.

PROCEDURE AND ORGANIZATION OF DATA

The speakers, identified by number only, were rated by three raters on five qualities according to the scale given below:

Table I. Speech Rating Scale

Points on which Speaker is Rated	High 10	9	8	7	6	5	4	3	2	Low 1
Physical Control										
Vocal Control										
Point (Controlling Idea or Theme Sentence)										
Sense of Communication										
Achievement of Purpose (Development of point—specific, appropriate, interesting, relevant)										

Although the main topics for each room were the same, the judges and students were not. As mentioned above, the students also rated every other student but only on one quality at a time, the five qualities being considered in successive order. Comparisons between the two groups of raters were thus based on a single quality for each student and not on the total score.

Medians were computed and used for comparison as well as for giving the students a numerical score. Means for each group were used in studying room-to-room variation among the qualities for both faculty and student results and, together with the standard deviations, gave an indication of the relationship between the standards of the two rating groups. Correlations between student and faculty ratings for each room were also computed.

The analysis of variance involved the setting up of individual tables for each of the one hundred and sixty-nine students. An example of one appears below:

Table II. Individual Table of Test Results

Raters	Qualities 1	2	3	4	5	Totals
1	8	5	7	7	7	34
2	5	6	6	5	5	27
3	5	4	4	2	2	17
Totals	18	15	17	14	14	78

This shows the scores for one student on all five qualities as given by the three raters. Combining these tables within each of the eight rooms and computing both the variances and the interactions, tables similar to Table VII on page 12 were set up. It is from these tables that the analysis of the sources of variation are found.

CONCLUSIONS

1. Room-to-Room Variation Among Faculty and Student Ratings:

Since the students were chosen at random for each of the eight rooms, there was reason to expect that a comparison of their average ratings among the eight rooms would reveal no significant

Table III. Room-to-Room Variation Among Student Ratings

Quality	Variance	Sum of Square of Deviations	Degrees of Freedom	Mean Square Deviation	
Physical Control	Within Rooms	36.00	31	1.16	
	Among Rooms	11.90	7	1.70	N.S.[1]
	Total	47.90	38		
Vocal Control	Within Rooms	20.33	25	.81	
	Among Rooms	13.73	7	1.96	S.5[2]
	Total	34.06	32		
Point	Within Rooms	33.38	24	1.39	
	Among Rooms	17.10	7	2.44	N.S.
	Total	50.48	31		
Communi-cation	Within Rooms	40.11	23	1.74	
	Among Rooms	11.74	7	1.68	N.S.
	Total	51.85	30		
Achieve-ment of Purpose	Within Rooms	31.19	24	1.28	
	Among Rooms	2.05	7	.29	N.S.
	Total	33.24	31		

[1]Non-significant difference.

[2]Significant difference at 5 per cent level.

differences. The results appear in Table III and Table IV, with the right-hand column indicating the level of significance or non-significance

as the case may be.

Table IV. Room-to-Room Variation Among Faculty Ratings

Quality	Variance	Sum of Square of Deviations	Degrees of Freedom	Mean Square Deviation	
Physical Control	Within Rooms	39.60	32	1.24	S.[1]
	Among Rooms	30.18	7	4.11	
	Total	69.78	39		
Vocal Control	Within Rooms	72.00	25	2.88	S.5[2]
	Among Rooms	49.64	7	7.09	
	Total	121.64	32		
Point	Within Rooms	40.25	24	1.68	S.5
	Among Rooms	31.22	7	4.76	
	Total	71.47	31		
Communi-cation	Within Rooms	61.25	24	2.55	N.S.[3]
	Among Rooms	36.97	7	5.28	
	Total	98.22	31		
Achieve-ment of Purpose	Within Rooms	41.75	24	1.74	N.S.
	Among Rooms	17.72	7	2.53	
	Total	59.72	31		

[1]Significant difference at 1 per cent level.
[2]Significant difference at 5 per cent level.
[3]Non-significant difference.

From these tables, it appears that the groups of students in the various rooms had more nearly uniform grading standards than did the faculty. Although the "among rooms" variance for the faculty raters is significant in only three of the five qualities, it is noticeable in every case that this variance is considerably greater

than the "within room" variance. In other words, the students were more in agreement as to the rating a speaker should get on these qualities while the faculty varied in their judgments. A further study of each room gave no evidence that the faculty raters in any particular room caused this great variation.

2. Total Room-to-Room Variation Among Faculty Ratings:

It had been planned to combine the faculty ratings and assign grades on the basis of all one hundred and sixty-nine ratings, but when an analysis of the room-to-room variation of the faculty ratings, given in Table V, indicated a significant difference of

Table V. Total Room-to-Room Variation Among Faculty Ratings

Variance	Sum of Square of Deviations	Degrees of Freedom	Mean Square Deviation
Within Rooms	5794.87	160	36.22
Among Rooms	2208.28	7	315.47
Total	8003.15	167	

the variance among rooms over the variance within a room, it was necessary to make grade assignments separately from the distributions within each room. This large variation indicated that a student's luck in drawing a room assignment was more important than giving a good speech.

3. Comparison of Means and Standard Deviations:

Students ranked the speeches higher than did the faculty in most cases, as exemplified by Table VI. Here we have included the averages for only two qualities, Physical Control and Vocal Control, but the other three show similar results. Although the amount of

Table VI. Means and Standard Deviations on Two Qualities

Room	Physical Control				Vocal Control			
	Means		Standard Deviations		Means		Standard Deviations	
	Student	Faculty	Student	Faculty	Student	Faculty	Student	Faculty
120	6.21	6.74	.77	.75	7.17	7.32	.19	.90
124	7.62	6.64	1.10	1.11	7.32	6.13	1.28	2.26
125	7.71	7.30	1.43	1.22	8.34	7.32	.33	.75
128	7.21	6.97	.39	.57	6.39	4.66	.57	.86
140	7.16	7.33	.81	.88	7.02	6.24	.69	1.65
144	7.65	5.39	.58	.68	7.10	4.72	.65	.44
145	7.66	7.26	.39	.90	7.41	6.57	.45	1.46
146	7.51	7.66	.93	1.11	8.29	7.84	.88	.83

difference between the means of the two groups varies, the greatest difference for all five qualities appears in Room 144. Two out of the three raters in this room were speech instructors who had not participated in the teaching of the English course. Since the variance among rooms is no more than a measure of the variation among the room means,

a comparison of the range of faculty and student means in the above table bears out the significant results obtained in Tables III and IV.

The average deviations from the mean within each room as measured by the standard deviation varies from room to room for both students and faculty but in most cases the faculty deviations are the larger. Hence, the faculty not only rated the speeches lower on the average but also showed greater variation in their ratings. Large variation is generally desirable since it results from finer discrimination in the quality measured.

4. Correlation between Room Means:

The correlation between faculty and student ratings ranges from .49 to .86, with most of them being above .60. Here it was necessary to pair the mean of twenty student ratings on Quality 1 with the mean of three faculty ratings on Quality 1 and so on for all the students, making certain that the ratings were for the same quality and the same student in each case. With correlations of this size, it appears that the students were consistently rating higher than the faculty.

5. Reliability of a Rater:

Although we would have liked to have a satisfactory method for computing the reliability of a rater, this seems impossible with the present study since identical speeches were not and never could be given. By means of correlations, the relationship between the ratings of three raters and one rater and between the ratings of

three raters and two raters was computed. The three-to-one comparison gave correlations from .55 to .74 and the three-to-two comparison gave correlations ranging from .79 to .88. These ranges do not include Room 144 where the results were quite different from the other rooms.

This method is based on the assumption that if two raters correlated very high with three raters, it would be useless to use three raters. From the results, however, it is conclusive that two raters are better than one but that two do not correlate high enough with three raters to warrant accepting the hypothesis and using two raters.

6. Analysis of Variance Results:

In order to make an investigation of the sources of variation leading to the discrepancy in the various ratings, an analysis of variance technique was employed and the computed results for each room set up in tabular form as shown in Table VII. Details of the

Table VII. Analysis of Variance Results for Room 145.

	Sum of Squares of Deviations	Degrees of Freedom	Mean Square Deviation
Raters	468.25	2	234.125
Students	236.55	20	11.827
Qualities	34.73	4	8.682
Raters x Qualities	28.45	8	3.556
Students x Qualities	230.34	80	2.879
Students x Raters	401.89	40	10.047
Students x Raters x Qualities	272.08	160	1.701

computation, analysis, and test of significance are not given here but may be found in such references as Rider (8, pp. 117-161) and Snedecor, (10, pp. 179-248). The sum of squares of deviations divided by the number of degrees of freedom, one less than the number of persons or qualities involved, gives the mean square deviation for each category.

Ideally, it would seem that the variance should be spread about as follows:

1. Low variance among the Raters would exist if they were in agreement on the various ratings.
2. Large variance among the Students would show that the raters were recognizing the difference in ability and ranking the students accordingly.
3. Low variance among the Qualities would result from the fact that quality variations would be eliminated in averaging over a large group of students.
4. Low variance should exist in the interaction of Qualities and Raters to show consistency of all raters in the ratings of the five qualities.
5. The interaction of Students and Qualities should be high since individual students would be expected to show differences on the various qualities.
6. The interaction of Students and Raters should be low since good raters should rate each student in the same manner.

7. The interaction of Students and Raters and Qualities should be small since most sources of variation are already accounted for.

With this brief explanation of what we would like to find in our study, let us examine the results. In every room the amount of variance among the raters greatly exceeded that of the students and qualities, as shown by Table VII, a typical example. This is exactly what we would not expect if the raters were in agreement on standards and qualities.

Although the variance among students, ranging from 7.939 to 37.375 is not large in comparison with the variance among the raters, it is significant in most cases, thereby indicating some spread among the students but far from the amount needed to compare favorably with the variance among the raters.

The amount of variance among the qualities ranges from .587 to 31.265, with two rooms showing significant results. With a group of students selected at random as this group was, it seems plausible that the average of all students on each quality should fall somewhere near the center of the scale--that is, result in a small amount of variance. Such favorable results were found in six of the eight rooms. It may be that the students in the other two rooms were quite different groups and should show a group average away from the center on some of the qualities, or it may be that the raters were emphasizing one quality more than another in making their ratings.

The variance due to the interaction of raters and qualities, which should ideally be low, shows a range from 1.367 to 7.631, and these amounts are significant in seven of the eight rooms. Here a tendency on the part of the rater to rate one quality high and another low is revealed.

The student and quality interaction variance ranges from 1.087 to 2.879. This is significant in a majority of the rooms but still rather low when a large amount of variance is necessary to show the expected individual differences on the various qualities.

The variance due to the interaction of students and raters is highly significant in all cases, again showing that the raters did not agree well on the ratings of individual students.

These analysis of variance results may be summarized in a few general statements:

1. The variance among the raters far exceeds that among the students although it is the latter group that should have a large spread.
2. The interaction variance among the qualities and students is not large enough to assure us that the raters were distinguishing among the five qualities.
3. The raters also show no consistent standard for rating the students on the five qualities nor do they rank the students in the same manner.

SUGGESTIONS

1. Because of the great variance among the faculty ratings, it would seem advisable to attempt some method for increasing the raters' skill.

2. From our reliability results, the number of raters in each room should not be reduced but increased if possible. The raters should be chosen from among the instructors of the course or at least all raters should be very clear on the standards appropriate to the course.

3. Although the experiment by Thompson (12, pp. 87-91) shows that ratings by grades and by numbers are approximately equal in accuracy, his study had nine points in each technique tested. Conklin (1) found that for untrained raters no more than five points should be included on the scale while Symonds (11, pp. 456-461) states that seven is the optimal number for greatest reliability. Since there is a tendency not to use the two end scores, thinking that possibly some later speaker will be a little better or even worse than the extreme speaker now being rated, the customary number of divisions on the scale probably should be increased; hence, the five points, corresponding to the five letter grades, probably could be increased by two without causing error. But, if the scale of ten points is to be continued, the correspondence between the five letter grades ordinarily used and the ten points should be thoroughly

understood by the raters.

4. The five qualities do not seem to have identical meanings to all raters so a more complete explanation of each quality and possibly a revision of the list might lower this variance.

5. As pointed out by Thompson (12, pp. 87-91), judges evaluations and interpretations are bound to differ somewhat but both techniques and qualities can be controlled to lessen the difference.

BIBLIOGRAPHY

1. Conklin, E. S., "The Scale of Values Method for Studies in Genetic Psychology," University of Oregon Publication, Vol. II (1923), No. 1.

2. Conklin, E. S. and J. W. Sutherland, "A Comparison of the Scale of Values Method with the Order of Merit Method," Journal of Experimental Psychology, Vol. VI (1923), pp. 44-57.

3. Guilford, J. P., Fundamental Statistics in Psychology and Education, McGraw-Hill Book Company, New York, 1942, pp. 273-284.

4. Guilford, J. P., Psychometric Methods, McGraw-Hill Book Company, New York, 1936, pp. 263-283.

5. Lindquist, E. F., Statistical Analysis in Educational Research, Houghton Mifflin Company, New York, 1940, pp. 173-179.

6. Newcomb, T., "An Experiment Designed to Test the Validity of a Rating Technique," Journal of Educational Psychology, Vol. XXII, (1931), pp. 279-289.

7. Nichols, Ralph G., "Case Method of Speech Examination," Quarterly Journal of Speech, Vol. XXVII, (Fall, 1941), pp. 385-391.

8. Rider, P. R., An Introduction to Modern Statistical Methods, John Wiley and Sons, Inc., New York, 1939, pp. 117-161.

9. Rugg, H. O., "Is the Rating of Human Character Practicable," Journal of Educational Psychology, Vol. XII,(1921) pp. 425-438, 485-501.

10. Snedecor, George W., Statistical Methods, Collegiate Press, Inc., Ames, Iowa (1938), pp. 179-248.

11. Symonds, P. M., "On the Loss of Reliability in Ratings Due to Coarseness of the Scale," Journal of Experimental Psychology, Vol. VII (1924), pp. 456-461.

12. Thompson, Wayne, "Is There a Yardstick for Measuring Speaking Skill," Quarterly Journal of Speech, Vol. XXIX, (1943), pp. 87-91.