DOCUMENT RESUME

ED 044 399 TE 002 034

AUTHOR	Barker, Larry L.; And Others
TITLE	Two Investigations of the Relationship among Selected Ratings of Speech Effectiveness and Comprehension.
INSTITUTION	Speech Association of America, New York, N.Y.
PUB DATE	Aug 68
NOTE	7p.
JOURNAL CIT	Speech Monographs; v35 n3 p400-406 Aug 1968
EDRS PRICE	EDRS Price MF-$0.25 HC Not Available from EDRS.
DESCRIPTORS	*Comparative Testing, Comprehension, *Educational Research, *Public Speaking, *Rating Scales, Speech Education, Speeches, *Speech Evaluation, Speech Tests

ABSTRACT

This research project examined the relationship between measures of speaker effectiveness obtained from rating scales and those obtained from objective comprehension tests of speech content. Two studies were used in order to provide independently derived results which could be compared. In the first study, 49 undergraduate public-speaking students judged 6 speeches using both a modified Baird-Knower rating scale and an objective comprehension test. Approximately half of the subjects listened to audio tapes of the speeches and half to video tapes with four of the six speeches used for final analysis. In the second study, 1190 students in 54 basic speech classes each judged one speech using five rating scales and a three-item comprehnsion test. Results from these studies indicated that (1) relationships among ratings on individual scales were high, (2) comprehension measures correlated to a modest degree (first investigation only), and (3) neglibible relationships existed between ratings and comprehension scores. These findings suggest that rating scales and comprehension scores are not measuring the same degrees and forms of speaker effectiveness. (JM)

Speech Monographs;
Vol. 35, No. 3,
August 1968

TWO INVESTIGATIONS OF THE RELATIONSHIP AMONG SELECTED RATINGS OF SPEECH EFFECTIVENESS AND COMPREHENSION

LARRY L. BARKER, ROBERT J. KIBLER and RUDOLPH W. GETER

RATINGS of speakers' effectiveness have traditionally been used by researchers and classroom instructors to assess speaking ability.[1] These ratings are based on theoretically, empirically, and/or observationally derived criteria which, it is assumed, reflect a valid measure of speaking skill. The use of such scales is based on the assumptions (1) that there is some absolute standard or model of excellence with which a given speech may be objectively compared; (2) that the comparison between an objective standard and the speech under observation may be made in numerical terms on an interval scale ranging from effective to ineffective; (3) that actual ratings are primarily a function of the stimulus (speech) rather than the internal subjective state of a competently trained judge of speaking.

Some researchers have proposed use of behavioral measures derived from audience reactions to assess speech effectiveness.[2] Examples of such measures include comprehension as determined by an objective test; attitude change as determined by a shift-of-opinion ballot; observable actions such as voting, buying, donating blood or charitable contributions; physiological measures of changes as in heart rate, blood pressure, pupillary dilation, or palmar sweat. Investigators proposing such measures of a speaker's effectiveness have contended that effective speeches do not necessarily adhere to set theoretical standards yet change the behavior(s) of audiences in manners desired by the speakers.

A first step toward clarifying these matters is to examine both ratings and behavioral measures to determine whether both are measuring the same

Dr. Barker is Assistant Professor of Speech and Assistant Director of the Communication Research Center, Department of Speech, Purdue University. Dr. Kibler is Associate Professor of Speech and Associate Director, Communication Research Center, Department of Speech, Purdue University. Mr. Geter is Instructor in Speech at the Purdue University Regional Campus, Fort Wayne, Indiana.

The first investigation here reported was supported cooperatively by the Educational Research Bureau, the Office of Research and Projects, and the School of Communications at Southern Illinois University. The second investigation comprises a portion of Mr. Geter's M.A. thesis (Purdue University, 1967). The investigators are indebted to Eugenia Hunter, David Petersen, and William Smith, all of Southern Illinois University, for assistance in this research.

[1] For examples of research and reviews of problems related to assessing speaking ability with ratings see Samuel L. Becker, "The Rating of Speeches: Scale Independence," *SM*, XXIX (March 1962), 38-44; Samuel L. Becker and Carl A. Dallinger, "The Effect of Instructional Methods upon Achievement and Attitudes in Communication Skills," *SM*, XXVII (March 1960), 70-76; Robert N. Bostrom, "Dogmatism, Rigidity, and Rating Behavior," *Speech Teacher*, XIII (November 1964), 283-287; Keith Brooks, "Some Basic Considerations in Rating Scale Development: A Descriptive Bibliography," *Central States Speech Journal*, IX (Fall 1957), 27-31; Theodore Clevenger, Jr., "Influence of Scale Complexity on the Reliability of Ratings of General Effectiveness in Public Speaking," *SM*, XXXI (June 1964), 153-156; Gerald R. Miller, "Agreement and the Grounds for It: Persistent Problems in Speech Rating," *Speech Teacher*, XIII (November 1964), 257-261.

[2] For example, see Paul D. Holtman, Robert E. Dunham, and Richard E. Spencer, "Direct Assessment of Effectiveness of Student Speakers," *The Journal of Communication*, XVI (June 1966), 126-132; Charles R. Gruner and Marsha W. Gruner, "Do Grades Awarded Classroom Speeches Indicate Effectiveness of Impact upon Audiences?" paper presented at the Speech Association of America Convention, Chicago, December 28, 1966.

degrees and forms of speakers' effectiveness. The present investigations focus on this problem by providing comparative data regarding the relationship between selected rating scales and a measure of one behavior—comprehension. Two studies are reported here. Different rating scales, comprehension measures, types of speeches, and subjects were used in the two investigations in order to provide independently derived results regarding the problem being examined.

INVESTIGATION I.

PROCEDURE

Subjects

Subjects for the first investigation were randomly selected from available public speaking classes at Southern Illinois University ($N = 49$). Participating subjects were inexperienced raters in that they had received only general classroom training in evaluation and they had limited experience in evaluating speeches in the classroom.

Criterion Variables

The variables under consideration were (1) a comprehension test and (2) a modified Baird-Knower rating scale.[3] The comprehension test contained twenty-five multiple-choice and fill-in items over five of six speeches presented to subjects in a series. Content validity was determined in the following manner. Manuscripts of the six speeches accompanied by sixty questions (ten items per speech) were distributed to thirty graduate students. The graduate students read each speech and then attempted to answer the questions about the speech. When answers were not apparent from the first reading, they were allowed to read through the manuscript again to find them. The tests completed by the graduate students were scored and test items for which answers were not identified by at least ninety percent of the graduate students were discarded. The remaining items were those determined to be answerable by reading the speeches. It was inferred that the same information could be obtained through listening carefully to the speeches. A split-half reliability estimate corrected by the Spearman-Brown prophecy formula was found to be .36 ($N = 56$ for basic speech course students), and the test was, consequently, judged sufficiently reliable for the purposes of the investigation.

The Baird-Knower scale is an instrument frequently used in classroom speech evaluation. Several modifications were made in the original Baird-Knower scale in the present investigation. (1) "Voice" and "Articulation," which appear as separate criteria on the original scale, were combined into one criterion requiring a single rating. (2) An "Audience Interest and Adaptation" scale was added as a criterion to be rated on the modified scale. (3) "Physical Activity" was eliminated as a scale because some subjects heard the speeches via audio tape. (4) Descriptive words and comments which are listed under each criterion on the original scale were changed from negative statements to positive statements on the modified scale. (5) The 1-9 rating scales used on the original Baird-Knower form were changed to 1-5 scales for each variable. Thus, the following scales were included on the modified evaluation form: speech attitudes and adjustments, voice and articulation, language, audience interests and adaptation, ideas, organization, and general effectiveness. In addition, a total for these ratings was computed.

[3] See the original Baird-Knower rating scale, published in A. Craig Baird and Franklin H. Knower, *General Speech*, 3rd ed. (New York, 1963) p. 84.

Stimulus Speeches

A series of six, three- to five-minute informative, video taped and audio taped speeches was shown to subjects. The speeches had been assessed in a previous investigation and judged to represent a wide range of speaking effectiveness. Nine faculty evaluators had judged two speeches to be above average, two average, and two below average. Test-retest reliability estimates on the Baird-Knower scales for a series of nine speeches (six used here plus three others), for nine faculty judges, ranged from .62 to .86. For eight of the nine scales reliability estimates were above .70, and five of the nine were above .75.[4]

The six speeches were recorded on audio and video tape in two different, randomly assigned orders with two-minute pauses between speeches. The pause allowed time for subjects to rate the speech before the next speech began, thus reducing the possibility of an adverse "overlap" effect.[5] The orders of presentation and the two modes were used to control for possible order and/or mode effects. Complete data were obtained for analysis for four of the six speeches presented. These were the speeches common to the two orders of presentation. Each speech omitted from the analysis appeared as the first speech in one of the two orders.

Administration of Speeches and Evaluative Instruments

Two weeks prior to the beginning of the investigation, subjects were given sample copies of the modified Baird-Knower rating scales and were instructed in their use by individual course instructors. In most cases subjects were allowed to practice using the scales by rating their classmates during regular class speeches.

The six speeches were presented to the subjects during a two-day period. On the day the speeches were presented, individual class instructors introduced a Research Associate, telling the subjects that the Associate was a member of the speech department attempting to assess the ability of students to evaluate speeches. Evaluation forms and instructions were distributed by the Research Associate and the instructions for using the rating scales were read aloud. One series of six speeches was then presented by video or audio tape, approximately half of the subjects receiving the stimulus speeches by each mode of presentation. Immediately after being exposed to each speech in the series, subjects evaluated it on the modified Baird-Knower rating form. At the conclusion of the entire series of speeches, evaluation forms were collected by the Research Associate. The comprehension test (immediate post test) was next distributed, and subjects were instructed to complete it. The test did not include questions on the first speech in the series. The orders of questions otherwise corresponded to the orders of presentation, and questions pertaining to a specific speech were identified by a heading which provided a cue to the content of the speech (e.g., "Air Defense Command"). Test booklets were collected at the end of the session and subjects were told they would receive their test scores at a later date.

Three weeks after the initial administration of treatments, but before subjects learned of their scores on the immediate post test, the same comprehension test (delayed post test) was ad-

[4] Robert J. Kibler, Larry L. Barker, and Roy H. Enoch, "The Development and Preliminary Assessment of a Set of Video-Taped Informative Speech Models," *Central States Speech Journal*, XVIII (November 1967), 268-275.

[5] Larry L. Barker, Robert J. Kibler, and Eugenia C. Hunter, "An Empirical Study of Overlap Rating Effects," *Speech Teacher*, XVII (March 1968), 160-166.

ministered to all subjects. Students were told by their individual instructors that the test was to determine how much information had been retained either as a result of initially viewing the speeches or taking the initial comprehension test.

Statistical Analysis

Pearson product-moment correlation coefficients for use with paired, ungrouped data were computed among rating scales on the modified Baird-Knower form and the comprehension tests.[6] The result of the analysis was a ten by ten matrix of inter-variable correlations.

The two different orders of presenting the series of six speeches resulted in four of the last five speeches in each series being the same, though they were heard in different orders. Only data for the four common speeches were included in the analysis of speech comprehension and ratings.

Results

The results of the investigation are reported in Table 1 and indicate that (1) there was a relatively high correlation among most scales on the modified Baird-Knower rating form (all r's $\geq$ .6117); and (2) the correlations among the scales on the Baird-Knower form and either immediate or delayed comprehension test scores were so low (all but three r's $\leq$ .17) as to suggest negligible relationships exist among these variables.

The study indicates, as has previous research, that most individual scales on the Baird-Knower rating form correlate highly with "General Effectiveness" and "Total Rating." The scale which correlated least with other scales was "Ideas," but the coefficients obtained were still relatively high.

Immediate and delayed comprehension tests correlated with each other to a modest degree. Information regarding the normal forgetting curve suggests an extremely high correlation should not be expected between these scores. The correlation obtained here supports this observation ($r = .60$).

INVESTIGATION II.

Procedure

Subjects

Subjects were students ($N = 1190$) enrolled in Purdue University's basic speech course and instructors (teaching assistants) for 72 sections of the course. Subjects in experimental groups ($N = 898$) were students assigned by the registrar to 54 sections of the course; the

[6] J. P. Guilford, *Fundamental Statistics in Psychology and Education*, 4th ed. (New York, 1965), pp. 91-112.

TABLE 1
CORRELATION MATRIX AMONG SELECTED SPEECH-EFFECTIVENESS RATING SCALES AND COMPREHENSION SCORES*

Variables	1	2	3	4	5	6	7	8	9	10
1 Speech Attitudes and Adjustments	1.00	.75	.79	.75	.62	.68	.80	.86	.07	.05
2 Voice and Articulation		1.00	.79	.75	.61	.72	.80	.87	.08	.09
3 Language			1.00	.79	.68	.76	.80	.90	.03	.06
4 Audience Interest and Adaptation				1.00	.77	.79	.85	.92	.13	.19
5 Ideas					1.00	.80	.78	.85	.06	.17
6 Organization						1.00	.83	.90	.09	.18
7 General Effectiveness							1.00	.94	.09	.20
8 Total Rating								1.00	.09	.15
9 Immediate Post-test Comprehension									1.00	.60
10 Delayed Post-test Comprehension										1.00

*These correlations are based on 196 observations on each variable by 49 subjects on 4 speeches. With N = 200 an r of .18 is necessary for significance at the .01 level.

control group consisted of students (N = 292) assigned to 18 different sections. The instructors for the 54 sections served as experimental subjects; instructors for the 18 different sections served as control subjects. In addition, 54 students (from other than the experimental or control sections) served as speakers to be evaluated by experimental groups.

Criterion Variables

Rating scales developed by Price[7] were modified for use in this investigation. The modifications included deleting one scale (Is the speaker intelligible?) and adding a "general effectiveness" scale. This was done on the basis of Clevenger's research.[8] Reliabilities for the Price scales in conjunction with general effectiveness have been reported by Clevenger (reliability coefficients ranged from .61 to .63 with a maximum of seven judges).[9] The following scales were included on the rating form as it was used in this investigation. (1) Does the speaker sound reasonable? (2) Does the speaker communicate well through bodily action? (3) Is the speaker socially acceptable? (4) Does the speaker use language vividly and imaginatively? (5) Does the speaker have a pleasing and expressive voice? (6) General, overall effectiveness. An average of these six scales was also computed as a criterion measure.

A three-item, multiple-choice, comprehension test was developed for each of the 54 persuasive speeches. Items for a comprehension test on each speech were drawn from those submitted by the student speakers but were modified to meet three criteria: (1) questions were to pertain to material at the beginning, middle, and end of the speech; (2) questions were to be phrased in multiple-choice form with five apparently reasonable choices (one correct answer and four foils); (3) the correct answers to the questions were to have been stated obviously in the speech and the language of the speech exactly duplicated in each correct answer.

Stimulus Speeches

The student speakers were assigned to present their speeches to one of 54 sections (experimental groups). The speakers had received minimal assistance from their course instructors in preparing persuasive speeches to be delivered from manuscript as the sixth assignment in the course. Students received extra credit in their own classes for presenting the speeches to the experimental groups and were informed that they were participating in a department-wide evaluation program.

For purposes of the investigation, comprehension score was defined as the sum of right answers for the three test items administered in experimental groups after the speeches had been heard. A comparison of the comprehension results for experimental groups (subjects who received the speeches and took the comprehension test) and control groups (subjects who did *not* receive the speeches but took the comprehension test) indicated that the experimental groups comprehended significantly more information than subjects in the control groups. When instructors' scores from experimental and control groups were compared by a t test, a significant t ($\leq .05$ level of confidence) of 6.65 ($df = 71$) was obtained, indicating the instructors in the experimental groups comprehended significantly more information from the

[7] William K. Price, "The University of Wisconsin Speech Attainment Test," unpubl. diss. (University of Wisconsin, 1964).

[8] Theodore Clevenger, Jr., "Influence of Scale Complexity on the Reliability of Ratings of General Effectiveness in Public Speaking," *loc. cit.*

[9] *Ibid.*

speeches than those in the control groups. Similar findings were obtained when students in experimental and control groups were compared. A significant t ($\leq .05$ level of confidence) of 11.342 ($df = 1189$) indicated that subjects in experimental groups comprehended significantly more information from the speeches than those in the control groups. These t test results show that subjects receiving the speeches obtained significantly more information than those who did not receive the messages. The results further indicate that the speeches contained information not generally available and that they were, in fact, informative.

Administration of Speeches and Evaluative Instruments

Student speakers were instructed to report five to ten minutes early to the classroom where they were scheduled to speak. The instructor of the section read instructions to each class at the beginning of the period. This information described the nature of the project on speech evaluation and indicated that the class had been selected as an "evaluation section" in the project. The students in the class were told that they would evaluate the speaker following his speech, and that their evaluations would not affect the speaker's grade in any way but might affect the evaluation techniques applied to students who would take the basic speech course in the future. Brief instructions concerning the concepts of specific scales on the rating form were also presented.

Each speaker was introduced by name, presented his speech, and left the room immediately afterward. The instructor then re-emphasized that evaluations being sought would be helpful in revising the techniques used in the course but would not affect the grade of the speaker. Following this reminder, the instructor distributed two-page booklets consisting of the general effectiveness scale, five specific rating scales (from which a mean score was procured), and the comprehension-test items. The evaluating students and instructor completed booklets which were then returned to one of the investigators after the class period ended.

Statistical Analysis

Pearson product-moment correlation coefficients for use with paired, ungrouped data were computed among the general-effectiveness rating scale, the average (arithmetic mean) of the five rating scales on the modified Price rating form, and the comprehension test. The result of the analysis was a three-by-three inter-variable correlation matrix.

Results

The results of the investigation indicate (1) that the correlations between the average of the five scales and the general effectiveness scale were rather high among both instructors ($r = .89$) and students ($r = .96$); (2) that the correlations between the average of the five scales on the modified Price form and comprehension-test scores were low ($r = -.06$ for instructors and $r = .10$ for students); and (3) that the correlations between general-effectiveness ratings and comprehension-test scores were also low ($r = .005$ for instructors and $r = .10$ for students). This study shows, as did the first investigation, that the rating measures were highly intercorrelated but that rating measures did not correlate meaningfully with the measure of comprehension.

Discussion

The results obtained from these two investigations indicate (1) that relationships among ratings on the individ-

ual scales were reasonably high, (2) that the two comprehension measures correlated to a modest degree (first investigation only), but (3) that negligible relationships existed between ratings on the various scales and the comprehension measures. These findings are interpreted as indicating that the two types of criterion measures, rating scales and comprehension scores, are probably not measuring the same degrees and forms of speakers' effectiveness.

If this observation is substantiated in subsequent research, it will be necessary for researchers to clarify the nature of the particular form of "speaker effectiveness" which is appropriate for any given research problem. Furthermore, those using rating scales for such purposes as assessing the effectiveness of classroom speech behavior and contest speaking may wish to weigh such practical concerns as convenience and ease in using rating scales against the more fundamental question of what is *really* being measured by such ratings of a speaker's effectiveness.

Additional research is required using different speech samples, different audiences, and different types of behavioral measures to ascertain whether the findings reported here are generalizable across other types of communicative events. Among the behavioral measures which might be correlated profitably with ratings in future research are: voting, nonverbal behavior, attitude change, and various types of recall. It would also be well to explore the consequences of altering evaluators' understandings of *why* they are furnishing evaluations and the influences of "set to rate" upon comprehension of messages.