DOCUMENT RESUME

ED 193 280 TM 800 595

AUTHOR McLean, James E.; Chissom, Brad S.
TITLE Evaluating Composition Skills: A Method and Example.
PUB DATE Sep 80
NOTE 22p.; Paper presented at the Annual Meeting of the Evaluation Network (6th, Memphis, TN, September 29-October 1, 1980).

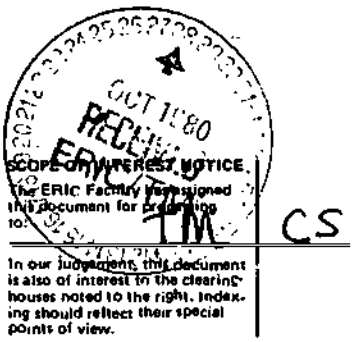
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Evaluation Methods; Informal Assessment; *Reliability; *Scoring; *Writing (Composition); *Writing Skills
IDENTIFIERS *Holistic Evaluation

ABSTRACT

Holistic evaluation is a reliable, valid, and cost-effective alternative to the usual mechanical assessment of writing. Writing samples are scored on a five-point scale against an overall impression of development, organization, and coherentness. The method was applied to the Communication Activities Skills Project (CASP) for grades 3-12. Writing samples were collected before and after the project from experimental and control groups, using a stimulus question. Three or four teachers rated each sample, after three days of holistic evaluation training. To eliminate possible bias, raters were not aware of the research group origin of the sample, and scoring was anonymous. Reliability coefficients were measured for: the pre-Project and post-Project samples, the difference between them, and choice of best sample. Coefficients ranged from .62 to .95; most were above .75. Results on the relative performance of control and experimental groups conflicted, depending on the method of comparison: analysis of variance vs proportion. Detailed information concerning the holistic method and the rating categories is appended. (GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193280



Evaluating Composition Skills:

A Method and Example

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

James E. McLean and Brad S. Chissom

The University of Alabama

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. McLean
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the

Sixth Annual Evaluation Network Conference

Memphis, Tennessee

September 30, 1980

TM 800595

Evaluating Composition Skills:

A Method and Example

The product evaluation of projects involving student composition has often neglected the use of actual writing samples in favor of more objective forms of measurement. This can easily result in the project concentrating on the teaching of the mechanics of composition rather than writing itself. The purpose of this paper is to propose a method for the product evaluation of a project using actual writing samples and provide an example of its application.

Method

The evaluation of a writing project brings to the surface two problems not usually associated with other types of projects. These are scoring or evaluating the writing samples themselves in a reliable and valid manner and developing a design which eliminates scorer halo effects. Each of these are addressed relevant to the evaluation of a Grade 3-12 writing project.

Scoring Writing Samples

Student writing samples were scored using a holistic method. A rater using the holistic method draws an over-all impression of the paper's worth. Detailed standards such as assigning weights to grammar, sentence structure, punctuation, etc. are avoided in favor of an impressionistic judgment of the paper's fluency. Fluency is judged in terms of the paper's development, organization, and most of all, its coherentness (Willig, 1979). Another

OCT 2 0 1980

discussion of the use of holistic scoring in evaluation can be found in Hendrickson (1980).

The holistic method of scoring writing samples requires multiple scorers. A minimum of three is recommended. Godshalk, Swineford, and Coffman (1966) indicated that holistic rating commonly provides intercorrelations among raters of .90, while the analytic technique often yields intercorrelations of .31 or lower even among trained professionals.

It has been reported that the holistic method increases speed as well as accuracy. Willig (1979) reported that a holistic rater rates approximately 30 papers per hour when assessing college level compositions. These same compositions would require approximately three hours of work if scored using an analytical method.

Training of raters proficient in the holistic rating method can often be accomplished in no more than three days. The training is based primarily on practice. Trainees were provided with a list of the categories and descriptions of these categories (see Appendix for those used in example project). The trainees were then provided with an example paper to rate using the categories. A rating was solicited from each trainee on the rated paper and displayed to the group. Any rater who differed from the majority by more than one category was asked to defend the rating. This practice helped trainees become more aware of the categories and standards (Willig, 1979). This training mode was continued over a three day period. At the end of the third day, virtual uniformity among the raters was noted.

Evaluation Design

The basic evaluation design was a pretest-posttest control group design. To eliminate possible rater bias, raters were not aware of whether a paper

was a pre-writing sample, post-writing sample, experimental group paper, or control group paper.

The writing samples were collected on a pre and post basis from the experimental and control groups using a stimulus question. Students were given 30 minutes to respond to the stimulus question in each case. The pre- and post-writing samples were assembled by grade and matched (pre and post for each student). From the matched writing samples, random samples were drawn from the experimental and control groups (100 papers each from grades 3-7, 200 from grades 8 and 9, and 200 from grades 10-12). Each pair of papers (a pre and post) were clipped together in random order with all distinguishing marks removed. The papers were then submitted for holistic scoring. In addition, each rater was to pick the "best" paper of the pair.

The data were analyzed in two ways. Analysis of variance was used to compare the gains between sum of pre-rating and sum of post-rating of the control groups with those of the experimental groups. Further, the proportion of students in the experimental group whose post-writing samples were rated better than their pre-writing samples were compared to the proportion in the control group.

Two aspects of the evaluation design need to be emphasized. These are the totally anonymous scoring using the holistic method and the sampling of writing samples. The anonymous scoring removed the chance of systematic bias in the rating of the writing samples. The random sampling from among all collected papers reduced greatly the labor in scoring them. Writing samples can be collected very economically (in terms of time and money) so not scoring a large proportion does not cause hardships on teachers or students. Not sampling until after completion of the project encouraged teachers to provide the full treatment to all experimental students.

Example

The method described above was applied to the writing portion of the Communications Activities Skills Project (CASP) in Augusta, Georgia. CASP was designed to improve communication skills in writing, reading, and listening, with the emphasis on writing, of students in grades 3-12. The results of the project need to be viewed with the fact in mind that a preliminary process evaluation indicated that the project was not implemented full at the teacher level. The purpose in presenting the project here is to illustrate the evaluation method and not the validity of the CASP curriculum.

Two aspects of CASP are presented. These are the reliability results of the holistic rating technique and the results of the analysis.

Reliability

Reliability coefficients were assessed for four aspects of the holistic rating system. These are pre-writing samples, post-writing samples, post minus pre differences, and choice of best writing sample.

The randomly selected pre and post samples of student essays were evaluated by a panel of language arts teachers with expertise in evaluation of writing using holistic grading. Each teacher on the grading panel read and rated every paper at his or her respective level. Using a five-point scale, the teachers assigned a rating to each paper. Three secondary English teachers rated the writing samples from grades 8-12, and four elementary teachers rated the writing samples from grades 3-7. The raters received training in the holistic approach to grading.

Pre-writing samples. For the elementary (grades 3-7) pre samples, the reliability coefficients of the ratings ranged from .87 to .93. The reliability coefficients were .62 for the junior high samples (grades 8-9) and .73 for the high school samples (grades 10-12).

Post-writing samples. For the posttest writing samples, the reliability coefficients for grades 3-7 ranged from .91 to .95. The reliability coefficients for the posttest were .65 for the junior high and .75 for the high school.

Post minus pre differences. The difference of the sum of post ratings and the pre ratings was calculated by subtraction. The reliability coefficient of the ratings of grades 3-7 ranged from .74 to .85, and the reliability coefficients for the junior high and high schools were .70 and .77, respectively.

Choice of best writing sample. The raters were asked to choose the better of the pretest and posttest writing samples. The reliability coefficients in grades 3-7 for choosing the better writing sample ranged from .71 to .83. The reliability coefficients for the junior high and high schools were .62 and .74, respectively. The raters for grades 3-7 were consistent in the rating of writing samples; however, the raters for the junior high did not agree on the ratings as consistently as the raters for grades 3-7.

The reliabilities are summarized in Table 1.

Results

The results of the writing sample analysis are presented in Tables 2 and 3. Table 2 presents the pre-writing sample results, post-writing sample results, and gain results for each grade or level. Table 3 presents the results for the comparisons of the proportions in the experimental and control groups showing improvement.

Table 1

Reliability of Ratings of Writing Samples

Grade	N	Pre	Post	Difference	Choice of Best
3	100	.93	.95	.85	.83
4	100	.92	.92	.82	.70
5	100	.87	.93	.82	.70
6	100	.87	.91	.77	.71
7	100	.90	.91	.74	.74
8 & 9	200	.62	.65	.70	.62
10-12	200	.73	.75	.77	.74

Table 2

Comparison of Pre- and Post-Writing Sample Ratings

Grade	Test Administered	Group	N	Mean	SD	F	p
3	Pre	CASP	50	8.600	3.325	.356	.5520
		Control	50	8.180	3.701		
	Post	CASP	50	12.540	3.887	6.250*	.0141
		Control	50	10.44	4.491		
	Gain	CASP	50	3.940	3.260	7.185*	.0086
		Control	50	2.260	3.002		
4	Pre	CASP	50	10.480	3.278	2.118	.1488
		Control	50	9.480	5.587		
	Post	CASP	50	13.640	3.367	3.711*	.0570
		Control	50	12.320	3.483		
	Gain	CASP	50	3.160	2.881	.336	.5637
		Control	50	2.840	2.637		
5	Pre	CASP	50	10.860	2.814	.010	.9208
		Control	50	10.920	3.193		
	Post	CASP	50	13.5200	4.087	.210	.6479
		Control	50	13.1800	3.293		
	Gain	CASP	50	2.660	2.804	.480	.4901
		Control	50	2.260	2.97		

*Significant at the .01 level in favor of the control group.

Table 2, Continued

Grade	Test Administered	Group	N	Mean	SD	F	p
6	Pre	CASP	50	10.460	2.971	.017	.8952
		Control	50	10.540	3.085		
	Post	CASP	50	11.840	3.353	2.421	.1230
		Control	50	12.900	3.400		
	Gain	CASP	50	1.380		3.926*	.0503
		Control	50	2.360			
7	Pre	CASP	50	11.480	3.940	.106	.7453
		Control	50	11.240	3.41		
	Post	CASP	50	12.360		.181	.6712
		Control	50	12.700			
	Gain	CASP	50	.880	2.577	1.321	.2531
		Control	50	1.460	2.468		
8 & 9	Pre	CASP	100	8.190	1.932	5.129*	.0246
		Control	100	8.820	2.00		
	Post	CASP	100	8.130	2.053	8.670*	.0036
		Control	100	9.010	2.172		
	Gain	CASP	100	- .0600	2.4070	.549	.4596
		Control	100	.1900	2.364		
10-12	Pre	CASP	100	8.140	2.070	9.796*	.0020
		Control	100	9.160	2.518		
	Post	CASP	100	8.30	2.464	8.578*	.0038
		Control	100	9.320	2.461		
	Gain	CASP	100	.160	2.415	.000	1.00
		Control	100	.160	2.135		

*Significant at the .01 level in favor of the control group.

Table 3

Proportion of Post-Writing Samples Rated Better Than Pre-Writing Samples

Grade	CASP (proportion)	Control (proportion)	z	p
3	.90	.76	.76	.22
4	.84	.84	.00	.50
5	.82	.82	.00	.50
6	.66	.76	- .48	.32
7	.60	.64	- .25	.40
8 & 9	.26	.41	-1.83*	.03
10-12	.34	.44	-1.13	.13

*Significant at the .05 level in favor of control group.

It is interesting to note that there are inconsistencies between the two methods of analysis. In grade 3, the analysis of variance method (ANOVA) indicated that the CASP group significantly outperformed the control group where the proportion comparison method was not significant. Also in grade 6, the ANOVA indicated the control group significantly outperformed the CASP group while the proportion comparison method was not significant. On the surface it seems that the ANOVA method is more sensitive. However, for grades 8 and 9, the proportion comparison method was significant in favor of the control group while the ANOVA method was not significant. It is not clear which method would be best but that the combination of the two should be better than either one by itself.

Discussion and Summary

Evaluation of student writing projects have often assessed only the mechanics of writing rather than writing itself due to the problems associated with grading writing samples. This, in turn, has lead to the emphasis of mechanics in the projects themselves. Braddock, Lloyd-Jones, and Schoer (1963) indicated "Study after study . . . confirms that instruction in formal grammar has little or no effect on the quality of student composition" (p. 37).

The method presented in this paper represents a viable alternative procedure with demonstrated reliability and validity. Further, costs associated with this design should be little more than those associated with traditional procedures.

References

Braddock, R., Lloyd-Jones, R., & Schoer, L. The state of knowledge about composition in Research in Written Composition. National Council of Teachers of English. Champaign, Illinois: 1963.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The Measurement of Writing Ability. Research Monograph, Number 6, College Entrance Examination Board, New York: 1966.

Hendrickson, L. Procedures and results of an evaluation of writing. Educational Evaluation and Policy Analysis. American Educational Research Association. Washington, D.C.: Vol. 2, Number 4, 1980.

Willig, C. Holistic evaluation and the CASP Project. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, November, 1979. Little Rock, Arkansas.

Appendix

Holistic Rating Categories*

1=FAILING. A hopeless paper in which the student fails to communicate. Often a paper in this category is too short to evaluate; usually, however, the paper contains so many grammatical, structural, punctuation, and/or spelling errors that it cannot be judged in terms of communication.

2=FAILING. A paper that shows some potential, that has a basic "thrust" toward fluency, but is so marred by mechanics that it cannot communicate even marginally.

3=BARE PASS. A paper that does communicate, though it contains serious errors of usage, grammar, punctuation, spelling, and/or structure.

4=SOLID PASS. A paper that contains elements that distinguish it, though it also may contain serious errors of usage, grammar, punctuation, spelling, and/or structure. The paper in this category is sufficiently fluent to communicate clearly.

5=SUPERIOR PASS. A paper that is organized, developed, and coherent; though it may contain mechanical deficiencies, it indicates both fluency and control.

*(Willig, 1979)

EVALUATING COMPOSITION SKILLS:

A METHOD AND EXAMPLE

A. METHOD
   SCORING WRITING SAMPLES
   EVALUATION DESIGN

B. EXAMPLE
   CASP
   COMMUNICATIONS
   ACTIVITY
   SKILLS
   PROJECT

SCORING WRITING SAMPLES

HOLISTIC METHOD:

- OVERALL IMPRESSION
- DEVELOPMENT
- ORGANIZATION
- COHERENTNESS

REQUIRES:

- MULTIPLE RATERS
- RATER TRAINING

HOLISTIC SCORING CATEGORIES

1 = FAILING--HOPELESS
2 = FAILING--POTENTIAL
3 = BARE PASS
4 = SOLID PASS
5 = SUPERIOR PASS

SEE APPENDIX FOR DETAILS

RATER TRAINING

1. PRESENT CRITERIA
2. HAVE RATERS RATE PAPER
3. DEFEND AND DISCUSS RATINGS
4. REPEAT PROCESS

EVALUATION DESIGN

PRE		POST
SAMPLE WRITING	CASP	SAMPLE WRITING
SAMPLE WRITING	REGULAR TREATMENT	SAMPLE WRITING

SPECIAL FEATURES
OF
EVALUATION DESIGN

1. ANONYMOUS SCORING
2. MULTIPLE RATERS
3. RANDOM SAMPLING

ANOVA

VS.

PROPORTION:

NEITHER CLEARLY

BEST

METHOD

AS DESCRIBED

REPRESENTS VIABLE

ALTERNATIVE

FOR

EVALUATION OF

WRITING PROJECTS