DOCUMENT RESUME

ED 157 079 CS 204 234

AUTHOR	Freedman, Sarah
TITLE	The Evaluators of Student Writing.
PUB DATE	78
NOTE	35p.; Research prepared at San Francisco State College
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Composition (Literary); Composition Skills (Literary); *Educational Research; *English Instruction; Essays; *Evaluation Criteria; *Evaluation Methods; Higher Education; *Research Methodology

ABSTRACT

In a study of factors that influence evaluators' ratings of student papers, 32 student essays were rewritten to make them stronger or weaker in content, organization, sentence structure, or mechanics; the essays were then submitted to evaluators in both their original and rewritten forms to determine the way in which the changes influenced the ratings. This paper discusses the procedures used in selecting the essays to be rewritten, rewriting them, and having them evaluated; it then reports the results of the study, examines the relationship between the holistic ratings and the raters' perceptions of the papers' strengths or weaknesses in each of the rewritten categories, and discusses the results and their pedagogical significance. Among the major findings were that the most important influences on raters' scores were the content and then the organization of the essays and that sentence structure and mechanics proved to be far less significant influences on holistic judgments. Seven tables are included. (GW)

Sarah Freedman

THE EVALUATORS OF STUDENT WRITING

This paper reports the results of an experimental study about the factors in college-level student papers that influence judges' rating of the quality of those papers. Most past research on this topic has been correlational rather than experimental. In a correlational study the researcher investigates natural occurrences. Students write papers; judges rate the quality of the papers. The researcher then examines the paper for traits associated with high and low ratings. One type of correlational study (e.g., Page, 1968; Hiller, Marcotte, and Martin, 1969; Slotnick and Knapp, 1971; Thompson, 1976; and , 1977) attempted to predict ratings with measures of characteristics in the student paper, such as the number of spelling errors or the length of the essay. Another type (e.g., Diederich, French, and Carleton, 1961; and Meyers, McConville, and Coffman, 1966) sought to account for ratings with characteristics of the judges, such as their personal biases or their degree of leniency. The past studies show that characteristics of papers and of judges are associated with or correlated with ratings. However, it is not possible for a correlational study to establish the causal influence of papers or judges on the ratings.

To establish causal relations it is necessary to turn to an experimental approach. For instance, in an experiment on the evaluation of composition the researcher might rewrite student papers to make them stronger or weaker along some dimension of content or form and then see how such rewriting influences the ratings. After the student writes the paper, the researcher, instead of observing natural occurrences as in the correlational paradigm, interferes with nature by experimentally manipulating the student paper. Judges then rate the quality of the rewritten paper. The researcher, who created certain characteristics in the rewritten essay, can determine the extent to which the manipulations influenced the ratings. Such an experimental approach is akin to one suggested by Hiller, Marcotte, and Martin (1969):

> if a given characteristic is present in an essay, does that characteristic affect the essay's quality as reflected in the grade assigned by expert graders? To answer this question we should have to manipulate the quality and quantity of relevant category items under an experimental procedure. (p. 274)

For my study, I decided to manipulate characteristics in essays to examine the influence of papers on ratings. My first problem was which characteristics to manipulate. I did not base my choice of characteristics on any one theory of discourse. Instead, I selected four very broad, but pedagogically interesting categories: content, organization, sentence structure, and mechanics. More precise features, which fall under these broad categories, such as the number of spelling errors or the length of the essay, had been the focus of many of the correlational studies in the first type cited earlier. However, for a first experimental study, I thought it wise to manipulate general characteristics so that in future studies on the influence of characteristics in papers on ratings the features of the influential general categories could be investigated.

I next rewrote essays of moderate quality to be either stronger or weaker in the four categories of content, organization, sentence structure, and mechanics. Exactly how to perform the rewriting proved to be a very complex problem which I discuss in detail in a separate section.

In almost every correlational study some aspect of content or a marker of content (e.g., essay length) predicted ratings. Based on this finding I posited one hypothesis about the effects of the rewriting: essays rewritten to be strong in content would be rated significantly higher than those rewritten to be weak in content. The findings of past studies about the relationship between judges' ratings and the quality of the organization, sentence structure, and mechanics were not so consistent, making it difficult to predict the potential effect of rewriting in these three categories. Nevertheless, my experiment would allow me to determine the effects of these pedagogically interesting characteristics on ratings too.

SELECTION OF ESSAYS TO BE REWRITTEN

College students in two different required writing sections at each of four Bay Area colleges wrote essays for the study. The colleges, which ranged in type from highly select, private schools to open-admissions, public schools, provided writers representing a wide range of abilities. According to Cass and Birnbaum's (1972) most recent descriptions of admissions criteria, the schools in order from most to least selective admissions requirements were: Stanford University,

University of Santa Clara, California State University at Hayward, and San Jose City College. The classes at each school were obtained on the recommendation of the department chair who was asked to suggest two "typical" classes taught by different teachers.

Students wrote the essays in class on one of eight topics designed to elicit essays in the argumentative mode of discourse. The topics either asked students to compare and contrast two quotations or to argue their opinion on a current, controversial issue. A sample of each type of topic follows:

1. A Founding Father said: "Get what you can, and what you get hold; 'Tis the Stone that will turn all your Lead into Gold."

   A contemporary writer said: "If it feels good, do it."[1]

   What do these two statements say? Explain how they are alike and how they are different.

2. President Ford gave Nixon an "unconditional pardon." Do you agree or disagree with Ford's decision? Give reasons for taking your position.

A student writing on one of each of the eight topics was selected from each class to participate in an earlier study.[2] The papers of these same eight students from each class were used as the basis for the rewriting in this study. In all, there were eight student essays on each of eight topics, a total of 64 papers. In the earlier study, four judges rated each essay holistically. Of the eight student essays on each topic, the four rated to be most average in quality in the earlier study were selected for experimental rewriting in this study. The other four, which were not rewritten, were the two which had been rated highest and the two which had been rated lowest on each topic in the earlier study. These non-rewritten essays served to establish the reliability of the ratings in this study.

REWRITING METHOD

Because of the dearth of operational definitions for strength and weakness of content, organization, sentence structure, and mechanics, I pondered, at first, how to undertake the rewriting task. I decided on the set of procedures in Table 1.

Insert Table 1 about here

To validate the rewriting procedures, I trained two different students to rewrite. If the two students and I as independent rewriters produced no significantly different results in essay ratings, I then could obtain a measure of the effects of rewriting the four categories to be weak or strong on the ratings of the essays. Furthermore, the fact that it would be possible to train others to follow the rewriting procedures consistently indicates that the rewriting could be replicated.

Rewriting the content category to be weak brought one major constraint: When the content was made weak, the organization could never be made strong. It would have been an exercise in absurdity to attempt to order illogical ideas logically or to order and transition appropriately a group of inherently unrelated ideas. Thus, there were twelve possible rewriting combinations:

C = Content
O = Organization
SS = Sentence Structure
M = Mechanics

+ = Strong
- = Weak

(1) +C, +O, +SS, +M
(2) +C, +O, +SS, -M
(3) +C, +O, -SS, +M
(4) +C, +O, -SS, -M
(5) +C, -O, +SS, +M
(6) +C, -O, +SS, -M
(7) +C, -O, -SS, +M
(8) +C, -O, -SS, -M
(9) -C, -O, +SS, +M
(10) -C, -O, +SS, -M
(11) -C, -O, -SS, +M
(12) -C, -O, -SS, -M

As rewriters we had a commitment to create a revised paper that retained, insofar as possible, the sense of the original essay. We attempted to highlight the strengths and weaknesses in each category in each paper. Nevertheless, the act of highlighting often produced a new paper substantially unlike the original. In spite of how unlike the original a rewritten version became, we remained committed to rewrite papers to be like the papers students actually produced. Still, the rewriting aimed to reproduce only the reasonable extremes of strength and weakness for each category. Papers were never rewritten to be average in any category.

The rewriting was performed in layers: content first, then organization, then sentence structure, and finally mechanics. When an earlier layer was rewritten as strong and a later one was rewritten as weak, the rewriters had to be extremely careful not to obscure the strength of the earlier category with the weakness of the latter. When rewriting content to be strong, weaknesses in organization, sentence structure, or mechanics were not allowed to obscure the ideas and the development of those ideas. Similarly, when rewriting sentence structure to be strong, weaknesses in mechanics were not allowed to obscure the strength of the sentences.

Finally, the four broad rewriting categories were defined to include all possible specific features in an essay that relate to its quality. Thus, if a composition was rewritten to be strong in every broad rewriting category, then it would have no residual weaknesses. Likewise, if a composition was rewritten to be weak in every category, it would have no residual strengths. Because I used only four category headings, some features related to essay quality did not fit under any particular category. For example, the feature word choice seemed to fit under none of the category headings. In fact, word choice fit under both the content and the sentence structure headings. Some changes in word choice affected the clarity of presentation of an idea; they were included under content. Other changes affected the parallel structure of a sentence; they were included under sentence structure. Other changes, which were purely matters of diction, were arbitrarily placed under sentence structure.

REWRITING DESIGN

This section discusses the plan for rewriting the four student papers on each of the eight topics. First, each of the papers was rewritten in three different versions each. Each original essay was keyed to three of the twelve possible rewriting combinations listed earlier. The four essays, each rewritten in three versions, made twelve versions on each topic. The twelve rewritten versions on each topic represented the twelve possible rewritten versions. Across the eight topics, with twelve rewritten versions per topic, there were 96 rewritten papers.

In the end, because of the constraint against combining weak content and strong organization, two-thirds of the 96 rewritten papers (N = 64) were strong in content; one-third (N = 32) were strong in organization. Half (N = 48) were strong in sentence structure, and half were strong in mechanics. Of course, the remainder for each category was weak in that category.

REWRITING PROCEDURE

Two Stanford University sophomores helped the investigator perform the rewriting in return for course credit. All rewriters first practiced applying the operational definitions for strength and weakness in the four categories (Table 1) to training essays, in order to establish and define common ground as readers and writers. During practice all rewriters independently rewrote the same essay according to the same rewriting combinations, then exchanged rewrites and discussed points of agreement and disagreement. During the actual rewriting one rewriter always wrote all three versions of an essay. A second rewriter checked the rewriting, and the third remained uninvolved.

EVALUATING DESIGN

Twelve evaluators were chosen according to the following criteria: (1) strength of professional recommendations, (2) quantity of teaching experience, and (3) educational background. All were highly recommended teachers on the staff of Stanford's freshman English program. I placed the evaluators into three types from most (type 1) to least (type 3) teaching the twelve possible rewritten versions. Across the eight topics, with twelve rewritten versions per topic, there were 96 rewritten papers.

In the end, because of the constraint against combining weak content and strong organization, two-thirds of the 96 rewritten papers (N = 64) were strong in content; one-third (N = 32) were strong in organization. Half (N = 48) were strong in sentence structure, and half were strong in mechanics. Of course, the remainder for each category was weak in that category.

REWRITING PROCEDURE

Two Stanford University sophomores helped the investigaton perform the rewriting in return for course credit. All rewriters first practiced applying the operational definitions for strength and weakness in the four categories (Table 1) to training essays, in order to establish and define common ground as readers and writers. During practice all rewriters independently rewrote the same essay according to the same rewriting combinations, then exchanged rewrites and discussed points of agreement and disagreement. During the actual rewriting one rewriter always wrote all three versions of an essay. A second rewriter checked the rewriting, and the third remained uninvolved.

EVALUATING DESIGN

Twelve evaluators were chosen according to the following criteria: (1) strength of professional recommendations, (2) quantity of teaching experience, and (3) educational background. All were highly recommended teachers on the staff of Stanford's freshman English program. I placed the evaluators into three types from most (type 1) to least (type 3) teaching experience and education. Evaluators were divided into four reading groups of three judges each. Each group rated essays on two of the eight topics. The different types of evaluators were balanced across the groups in order to avoid placing a group of less experienced evaluators together.

Training and reading packets were compiled for each rater for each topic. The training packets contained holistic scoring forms and two training essays typical of those in the experimental set. In the reading packets two supplemental training essays were followed by eight experimental student essays. Of the eignt experimental essays all three evaluators in each group received the four essays that had not been rewritten. The four remaining essays in the experimental set were selected for each judge from those that had been rewritten. Each of the three evaluators received one of the three versions of each of the four rewritten essays. The rewritten versions were assigned to evaluators according to a balanced plan. The order of the eight experimental essays was randomized for each evaluator.

EVALUATING PROCEDURE

The evaluations took place on four consecutive days. One group of three evaluators rated essays on two of the eight topics on the first day; a second group of three evaluators rated essays on another two of the eight topics on the second day, and so on. Each group of evaluators was informed that college students had produced the essays. The fact that some essays had been rewritten was concealed from the evaluators. All essays were typed.

Before rating any essays the group of evaluators discussed their expectations for a good essay on the first topic they would rate. Then, they rated the first two training essays from the training packet, using the four-point holistic scale. After rating the training essays, they discussed with the trainer and with each other their reasons for assigning the scores they did. If they evidenced a difference of two or more points on the four-point holistic scale, the trainer tried to guide them to understand and reconcile their differences. Raters were never forced to agree.

After discussing these training essays, the evaluators received their reading packet on the first topic and began the holistic ratings. If the judges disagreed with one another on scores for the optional training essays in the reading packet, the reading was interrupted to continue training with these optional training essays. This same procedure was repeated for the second topic.

The group of judges first gave holistic evaluations to all essays on both topics. After completing both holistic evaluation sessions, the judges were asked to provide a more detailed evaluation for the rewritten essays on each topic. For these essays, the judges had to determine whether the content, organization, sentence structure, and mechanics was weak or strong. The fact that these essays had been rewritten to be weak or strong in these four categories was still concealed from the judges.

RELIABILITY

To assess the reliability of the judges' ratings, I used the Cronbach alpha (Cronbach, 1970, p. 159; Calfee and Drum, Before rating any essays the group of evaluators discussed their expectations for a good essay on the first topic they would rate. Then, they rated the first two training essays from the training packet, using the four-point holistic scale. After rating the training essays, they discussed with the trainer and with each other their reasons for assigning the scores they did. If they evidenced a difference of two or more points on the four-point holistic scale, the trainer tried to guide them to understand and reconcile their differences. Raters were never forced to agree.

After discussing these training essays, the evaluators received their reading packet on the first topic and began the holistic ratings. If the judges disagreed with one another on scores for the optional training essays in the reading packet, the reading was interrupted to continue training with these optional training essays. This same procedure was repeated for the second topic.

The group of judges first gave holistic evaluations to all essays on both topics. After completing both holistic evaluation sessions, the judges were asked to provide a more detailed evaluation for the rewritten essays on each topic. For these essays, the judges had to determine whether the content, organization, sentence structure, and mechanics was weak or strong. The fact that these essays had been rewritten to be weak or strong in these four categories was still concealed from the judges.

RELIABILITY

To assess the reliability of the judges' ratings, I used the Cronbach alpha (Cronbach, 1970, p. 159; Calfee and Drum, 1976, p. 14). The reliability for the ratings given by each group of judges was determined by comparing the ratings the different judges in a group assigned to the four papers on each topic that had not been rewritten. All ratings proved highly reliable. The reliability scores within each group of raters ranged from .86 to .96. Thus, these reliability scores for the non-rewritten papers suggest that the ratings of the rewritten papers were also quite reliable.

RESULTS OF REWRITING

I first examined the main results of the experiment, the effects of rewriting strong and weak content, organization, sentence structure, and mechanics on the raters' holistic scores. With an analysis of variance, I measured whether each rewriting characteristic contributed significantly to the difference in the scores the raters gave (Table 2). My hypothesis, that essays rewritten to be strong in content would be rated significantly higher than those rewritten to be weak in content, was confirmed. The largest main effect of the rewriting was for the content variable. The organization variable also proved to have a highly significant effect on the judges' scores. Mechanics too had its effect. Additionally, there were significant interactions between organization and mechanics and between organization and sentence structure.

Insert Table 2 about here

Table 3 helps explain these main results. It presents the mean scores for papers rewritten to be strong and those rewritten to be weak in each of the four rewriting categories.

Insert Table 3 about here

It reveals that the difference between the average score given papers weak in content and the average score given papers strong in content differed by 1.06 points. Since the maximum possible difference between the average scores was 3 points (on the 1 to 4 holistic scale), a difference of over one point is quite large. Strong versus weak rewriting in organization also led to a difference of about 1 point. The effect of mechanics and sentence structure rewriting was about 1/2 and 1/4 point, respectively.

The interactions between organization and mechanics and organization and sentence structure in these main results show that only if the essay had strong organization did the strength or weakness of the mechanics and sentence structure matter (Table 4). If the organization was strong, the

Insert Table 4 about here

mechanics rewriting caused almost an entire point difference between the strong and weak essays' average scores. In the same situation, sentence structure rewriting caused about a 1/2 point difference. The relation between organization and mechanics was more significant than that between organization and sentence structure.

In summary, the main results of the rewriting showed that the most significant influence on raters' scores is the strength of the content of the essay. The second most important influence proved to be the strength of the organization of that content. The third significant influence was the strength of the mechanics. Furthermore, the strength of the mechanics was most important when the organization was strong, and because the sentence structure alone was insignificant, the strength of the sentence structure was important only when the organization was strong.

EVALUATORS' PERCEPTIONS REWRITERS' INTENTIONS

I next prepared to examine a secondary set of main results. Instead of using the actual rewriting as the independent variable, I wished to examine the holistic ratings according to the raters' perceptions of the strength or weakness of each of the rewritten categories. The raters' perceptions were determined by their indication of their judgment of the strength or weakness of the rewritten categories of the rewritten essays. However, before I could examine the results using the raters' perceptions it was first necessary to measure how well the raters' perceptions of the strength or weakness of the rewritten categories matched with the way the rewriters intended to rewrite them. If the match was exact, there would be no reason to seek these secondary results. Since the categories were rewritten to be extremely strong or weak, I expected the raters to perceive the rewriting accurately for the most part even though they were not given the criteria for the rewriting.

Table 5 specifies the overall percent of match and mismatch for each category. Raters usually judged the strength and weakness of the categories accurately, although they did

Insert Table 5 about here

not always. The content category proved most difficult for the raters to assess; organization was next in difficulty followed by sentence structure and then mechanics. This order seems quite logical; the evaluators' overall perceptions of the different categories matched with the rewriters' intentions a lower percent of the time for the more difficult to define, abstract categories than for the more objective, concrete categories.

EVALUATORS' PERCEPTIONS AND THEIR HOLISTIC EVALUATIONS

Since the evaluators' perceptions of the quality of the content, organization, sentence structure, and mechanics of the essay did not match the rewriters' intentions exactly, I next examined the secondary set of major results, the relationship between raters' perceptions and their holistic scores. The evaluators' perceptions of the strength or weakness of the content, organization, sentence structure, and mechanics became the independent variables in the analysis of variance rather than the actual rewriting for the categories. Table 6 shows

Insert Table 6 about here

that the results for content and organization were similar to those found in the main results detailed earlier. But other findings proved different. Perceived mechanics, this time, did not contribute significantly to the evaluators' scores; perceived sentence structure did. None of the perceived quality categories interacted significantly with one another. Table 7 shows a comparison of the average difference between

Insert Table 7 about here

ratings on the perceived strong and weak level of each category across all of the rewritten essays.

DISCUSSION

In the interpretation of the results, several areas deserve mention. First, all methods of analysis show the most important influences on the raters' scores were the content and then the organization of the essay. These two aspects of the written text merit the special attention of the writing student, teacher, and researcher. Sentence structure and mechanics proved much less significant influences on holistic judgments.

Because the influence of sentence structure and mechanics are neither as strong nor as consistent as the influences of content and organization, raters are probably less conscious of the effects of these less important influences. The effects of sentence structure and mechanics and the interactions of these categories with organization differ between the analysis using the actual rewriting as the dependent measure and the analysis using the judges' perceptions of the quality of the rewritten categories as the dependent measure. The differences suggest that the judges' perceptions would have them claim that their holistic ratings were not weighted on the rewriting categories in the ways the analysis according to the rewriting showed them to be. Raters seem to perceive that they give: (1) less credit for the conventions of standard edited English (mechanics); (2) more credit for well-formed, graceful sentences (sentence structure); and (3) discrete credit for the four rewriting categories.

Two raters were disqualified from the research because the frequency of their mismatch was more than two standard deviations above the mean. These raters also exhibited a different pattern of mismatch from the others. They mismatched on all categories, and they mismatched more than the others on the more objective categories, mechanics and sentence structure. The raters who did not show frequent mismatch tended to cluster their mismatch on content or organization, mismatching mostly on only one category. Perhaps raters' abilities to perceive the quality of rewritten categories within essays could be used to test their competence before choosing them to participate in evaluation projects.

The raters, both in their mismatch patterns and with their holistic scores, showed a significant tendency to evaluate students' writing negatively. In all categories when their perceptions did not match the rewriters' intentions, they judged strong rewriting as weak more of the time than not. Also, the distribution of the holistic scores was skewed toward the lower end of the scoring range. Conlan (1976) at Educational Testing Service corroborated this tendency of readers to rate negatively, "Unfortunately, no reader--experienced or inexperienced--seems to need assurance about giving out 2's and 1's [lowest scores on four-point holistic scale]; what all readers seem to need from time to time is the reminder that not all the papers are '2' papers or '1' papers" (p. 4). Perhaps evaluators should be less reluctant to compliment student writing.

One limitation of this study is the difficulty in interpreting the exact results of the rewriting. When each category was rewritten, several aspects of the category were rewritten at once. The exact aspects of the category which influenced raters' reactions to that particular category remain unknown, and are a topic for further study. It is possible that the raters reacted to the rewriting of all of the aspects for each category. It is equally possible that they reacted to some part or combination of parts of the rewriting. For example, perhaps order but not transitions was what influenced raters in the organization rewrite. Broad areas of influence on raters' judgments have been identified; the more precise influences need to be examined.

A second limitation is the homogeneity of the raters in this study. They were carefully defined as a select, homogeneous group of college writing teachers from a major university. It would be interesting to learn how other raters would react. Joseph Williams (1977), rewriting essays in nominal and verbal styles, compared the responses of several types of evaluators who thought they were evaluating for different reasons. His judges included new graduate students in a Master of Arts in Teaching program, experienced college English professors, and evaluators who regularly read essays for a state proficiency examination. Some evaluators thought they were helping a fellow graduate student with a research project;

others thought they were determining the reliability of a college writing examination. He found that different types of raters preferred different types of essays. Some groups preferred a nominal style; others perferred a verbal style.

PEDAGOGICAL SIGNIFICANCE

If society values content and organization as much as the raters in this project did, then according to the definitions of content and organization used in this study, a pedagogy for teaching writing should aim first to help students develop their ideas logically, being sensitive to the appropriate amount of explanation necessary for the audience. Then it should focus on teaching students to organize the developed ideas so that they would be easily understood and favorably evaluated. The interaction between organization and mechanics and organization and sentence structure, showing that the quality of the mechanics and sentence structure matter most when the organization is strong, points even more strongly to a pedagogy aimed at teaching the skills of organization before or at least alongside those of mechanics and sentence structure.

It seems today that many college level curricula begin with a focus on helping students correct mechanical and syntactic problems rather than with the more fundamental aspects of the discourse. It is important to supplement these curricula with carefully planned curricula for teaching content and organization. Certainly, because of the excellent research in the area of written sentence structure (Hunt, 1965; Mellon, 1969; O'Hare, 1971; Christensen, 1967) and because of the objective nature of the mechanical rules for standard edited English, sentence structure and mechanics have become easier to teach than content and organization. The English profession knows more about teaching, evaluating, and doing research on sentence structure and mechanics than on the less objective areas of content and organization. Conceivably, instruction in strengthening sentence structure or mechanics could result in strong content or organization. But such a hypothesis has not been tested.

Scholars like Donald Murray (1968), Ken Macrorie (1970), and Peter Elbow (1973) have advocated college writing curricula centered around the larger levels of the discourse. However, although Murray, Macrorie, and Elbow offer pedagogical suggestions for encouraging students to find and expand their ideas, they do not offer as complete or as well-defined a pedagogy as, say, Christensen does for syntax in <u>The Christensen Rhetoric Program</u> (1968). Other scholars, like Kenneth Burke (1945), D. Gordon Rohmann (1965), and Young, Becker, and Pike (1970) have contributed to developing a modern theory of invention. Young, Becker, and Pike, in particular, have developed heuristic procedures for helping students retrieve, analyze, and order their ideas for a particular audience. Besides such work in invention, with pedagogies focused primarily on idea generation, more research focusing on how to analyze, teach, and evaluate the logical development of the already generated ideas (content) and the techniques used for ordering and making transitions between those ideas (organization) is badly needed before more concrete pedagogies can evolve.

CONCLUSIONS The methodology employed in this experiment provides a framework for studying the evaluation of student writing in many other contexts. Certainly the following aspects of the evaluation process deserve attention:

(1) the more exact effects of the rewriting (what within the categories influences the evaluators, does the influence work in a continuum--if so, where are the critical spots on the continuum?);

(2) evaluations given by different kinds of evaluators (e.g., peers, classroom teachers with varying amounts of experience who teach different subjects to different ages, teachers from non-mainstream cultural groups, teacher trainers);

(3) the evaluation of papers written by students from other age groups (elementary through senior high school);

(4) the evaluation of papers written in other modes of discourse (at least narrative or some expressive modes of writing).

I believe a more in-depth and more precise investigation of the aspects within the two most influential rewriting categories, content and organization, is the most important and the most promising area for future research. In this study much of the rewriting in these categories was done intuitively. Now that some aspects of content and organization have been proven powerful influences on evaluators' judgments, the precise aspects of content and organization that influence evaluators must be explored more carefully. Schemes for the linguistic analyses of texts (e.g. Kintsch, 1974) might provide a foundation for more careful experimentation in these aspects of writing. Out of such explorations a sound basis for developing curricula focused on teaching the skills of content and organization can evolve.

By using experimental research to learn more about the evaluation process, educators will be able to develop more efficient and fairer means of evaluation. Teachers as well as researchers need to know how to evaluate the quality of student writing. Discoveries of the bases of evaluators' responses will contribute to a set of definitions of what evaluators see as good writing. These definitions then can be examined critically and those criteria of good writing that seem sound can be incorporated into pedagogy and into training evaluators of student writing. One of the first steps in improving the evaluation and teaching of student writing is understanding how evaluators evaluate as they do.

BIBLIOGRAPHY

Burke, K. A grammar of motives. New York: Prentice-Hall, 1945.

Calfee, R. Sources of dependency in cognitive processes. In D. Klahr (Ed.) Cognition and instruction: 10th annual Carnegie-Mellon symposium on cognition. Hillsdale, N.J.: L. Erlbaum, 1976.

_____. and P. Drum. How the researcher can help the teacher with classroom assessment, Stanford University, mimeograph. 1976.

Christensen, F. Notes toward a new rhetoric: six essays for teachers. New York: Harper and Row, 1976.

_____. The Christensen rhetoric program. New York: Harper and Row, 1968.

Conlan, G. How the essay in the CEEB English composition test is scored: an introduction to the reading for readers. Educational Testing Service, 1976.

Cronbach, L. Essentials of psychological testing (3rd ed.). New York: Harper and Row, 1970.

Diederich, P., S. French, and S. Carlton. Factors in judgments of writing ability. Research Bulletin 61-15. Princeton: Educational Testing Service, 1961.

Elbow, P. Writing without teachers. New York: Oxford University Press, 1973.

Hiller, J., D. Marcotte, and T. Martin. Opinionation, vagueness, and specificity distinctions: essay traits measured by computer. American Educational Research Journal, 1969, 6, 271-286.

Hunt, K. Grammatical structures written at three grade levels. Urbana, Ill.: National Council of Teachers of English, 1965.

Kintsch, W. The representation of meaning in memory. Hillsdale, N.J.: L. Erlbaum, 1974.

Macrorie, K. Uptaught. New York: Hayden, 1970.

Mellon, J. Transformational sentence-combining: a method for enhancing the development of syntactic fluency in English composition. Urbana, Ill.: National Council of Teachers of English, 1969.

Meyers, A., C. McConville, and W. Coffman. Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 1966, 26, 41-54.

Murray, D. A writer teaches writing. Boston: Houghton Mifflin Company, 1968.

[redacted] An analysis of readers' responses to essays. Research in the Teaching of English, 1977, 11, 164-174.

BIBLIOGRAPHY

Burke, K. A grammar of motives. New York: Prentice-Hall, 1945.

Calfee, R. Sources of dependency in cognitive processes. In D. Klahr (Ed.) Cognition and instruction: 10th annual Carnegie-Mellon symposium on cognition. Hillsdale, N.J.: L. Erlbaum, 1976.

_____. and P. Drum. How the researcher can help the teacher with classroom assessment, Stanford University, mimeograph. 1976.

Christensen, F. Notes toward a new rhetoric: six essays for teachers. New York: Harper and Row, 1976.

_____. The Christensen rhetoric program. New York: Harper and Row, 1968.

Conlan, G. How the essay in the CEEB English composition test is scored: an introduction to the reading for readers. Educational Testing Service, 1976.

Cronbach, L. Essentials of psychological testing (3rd ed.). New York: Harper and Row, 1970.

Diederich, P., S. French, and S. Carlton. Factors in judgments of writing ability. Research Bulletin 61-15. Princeton: Educational Testing Service, 1961.

Elbow, P. Writing without teachers. New York: Oxford University Press, 1973.

Hiller, J., D. Marcotte, and T. Martin. Opinionation, vagueness, and specificity distinctions: essay traits measured by computer. American Educational Research Journal, 1969, 6, 271-286.

Hunt, K. Grammatical structures written at three grade levels. Urbana, Ill.: National Council of Teachers of English, 1965.

Kintsch, W. The representation of meaning in memory. Hillsdale, N.J.: L. Erlbaum, 1974.

Macrorie, K. Uptaught. New York: Hayden, 1970.

Mellon, J. Transformational sentence-combining: a method for enhancing the development of syntactic fluency in English composition. Urbana; Ill.: National Council of Teachers of English, 1969.

Meyers, A., C. McConville, and W. Coffman. Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 1966, 26, 41-54.

Murray, D. A writer teaches writing. Boston: Houghton Mifflin Company, 1968.

[redacted] An analysis of readers' responses to essays. Research in the Teaching of English, 1977, 11, 164-174.

O'Hare, F. Sentence combining: improving student writing without formal grammar instruction. Urbana, Ill.: National Council of Teachers of English, 1971.

Page, E. Analyzing student essays by computer. International Review of Education, 1968, 14, 210-225.

Rohman, D.G. Pre-writing: the stage of discovery in the writing process. College Composition and Communication, 1965, 16, 106-112.

Slotnick, H. and J. Knapp. Essay grading by computer: a laboratory phenomenon? Educational Measurement, 1972, 9, 253-263.

Thompson, R. Predicting writing quality, writing weaknesses that dependably predict holistic evaluations of freshman compositions. English Studies Collections, Inc., 1976, 1.

Williams, J. Nominal and verbal styles: some affective consequences. The University of Chicago, mimeograph, 1977.

Young, R., A. Becker, and K. Pike. Rhetoric: discovery and change. New York: Harcourt, Brace, and World, 1970.

TABLE 1

REWRITING RULES

Content

Strong	Weak
1. Delete all misinterpretations of quotations; add sound reinterpretations.	1. Retain all misinterpretations of quotations; add one misinterpretation if none are present.
2. Delete ideas not relevant to the topic unless they can be made relevant. If no ideas in the paper are relevant, either justify their inclusion or pull together possible relationships.	2. Retain all ideas not relevant to the topic. Do not add extra irrelevant ideas.
3. Delete repetition of entire arguments.	3. * Include repetition of entire arguments.
4. Take remaining ideas and: develop, resolve logical contradictions within ideas, clarify (this involves changes in word choice).	4. Take remaining ideas and: delete development, include contradictions within ideas, make ideas unclear and ambiguous (this involves changes in word choice).

*"Include" is used throughout this Table to mean retain and/or add.

TABLE 1--continued

Organization

Strong	Weak
1. Paragraph appropriately.	1. Include three misparagraphings per 25Ø word page.
2. Order ideas logically. Respect rules of given-new information. Keep main ideas together.	2. Violate logical order by separating development of a main idea (three times per two pages). Violate given-new strategies.
3. Include appropriate inter and intra paragraph transitions: repeat key words and use transition words and phrases appropriately.	3. Delete inter and intra paragraph transitions: vary the lexical items chosen for key words and avoid using transition words and phrases appropriately.

TABLE 1--continued

Organization

Strong	Weak
1. Paragraph appropriately.	1. Include three misparagraphings per 25Ø word page.
2. Order ideas logically. Respect rules of given-new information. Keep main ideas together.	2. Violate logical order by separating development of a main idea (three times per two pages). Violate given-new strategies.
3. Include appropriate inter and intra paragraph transitions: repeat key words and use transition words and phrases appropriately.	3. Delete inter and intra paragraph transitions: vary the lexical items chosen for key words and avoid using transition words and phrases appropriately.

TABLE 1--continued

Sentence Structure

Strong	Weak
1. Combine and balance sentences to achieve a mature syntactic style: reduce number of compound sentences; untangle awkward and unclear sentences, include final free modifiers and graceful parallel structures.	1. Achieve an immature syntactic style: include simple, primer sentences (include much compounding) or include long, rambly, uncontrolled, awkward sentences delete graceful parallelism, include verboseness on the sentence level.
2. Vary sentence structure.	2. Include sentence fragments and run ons.
3. Include at least one advanced punctuation mark: semicolon or colon.	3. Delete advanced punctuation marks: semicolon or colon.
4. Use appropriate tense and reference between and within sentences.	4. Use inappropriate tense and reference between and within sentences.
5. Change any misused words. Do not alter overall vocabulary level.	5. Include misused words.

TABLE 1--continued

Mechanics

Strong	Weak
1. Follow conventions of standard edited English.	1. Commas. Violate at least three of the following rules: Comma before conjunction in compound sentence. Comma after introductory adverbial clause. Comma within quotation marks. Commas between words and phrases in series.
	2. Quotation marks. Overuse and use inconsistently. Use to emphasize words. Forget to either open or close quotations.
	3. Possessives. Misuse "'s." Omit when needed. Use structures like "their's."
	4. Capitalization. Omit for proper names. Forget to capitalize first word of sentences. Add inappropriately for emphasis.
	5. Underlining. Overuse and use inappropriately for emphasis.
	6. Spelling. Include four or five errors per page.

The operational definitions, the general rules we followed for rewriting all four categories to be weak and strong, were adapted from descriptions on analytic rating scales (Diederich, 1974; Adler, 1972; [redacted], 1977), were based on definitions used in past correlational research on readers' responses (Thompson, 1976), and also were based on critical analyses of the strengths and weaknesses within the student papers written for this study.

TABLE 2

ANALYSIS OF VARIANCE FOR HOLISTIC SCORES: REWRITING EFFECTS

Source	df	MS	F1	F2
Reader (R)	11	.448		
Content (C)	1	9.860	37.78***	31.70***
Organization (O)	1	5.195	29.69***	16.70***
Sentence Structure (SS)	1	1.5	2.54	4.82
Mechanics (M)	1	5.042	9.77**	16.21***
C X SS	1	1.960		6.30
C X M	1	.990		3.18
O X SS	1	3.767		12.11**
O X M	1	6.155		19.79***
SS X M	1	.001		0
Reader Interactions				
R X C	11	.261		
R X O	11	.175		
R X SS	11	.591		
R X M	11	.516		
Residual	31	.311		

** p < .01 1,11df F1 = 9.65 F1 based on R by Source
*** p < .001 1,11df F1 = 19.69

**p < .01 1,31df F2 = 7.56 F2 based on residual
***p < .001 1,31df F2 = 13.29 error variance

TABLE 3

MEAN HOLISTIC JUDGMENTS (4 = highest, 1 = lowest)

	Strong	Weak	Difference
Content	N = 64 2.375	N = 32 1. 313	1.06
Organization	N = 32 2.656	N = 64 1.703	.95
Sent. Str.	N = 48 2.146	N = 48 1.896	.25
Mechanics	N = 48 2.250	N = 48 1.792	.46

N = 96 rewritten essays

TABLE 4

EFFECTS OF INTERACTION BETWEEN
ORGANIZATION AND MECHANICS AND SENTENCE STRUCTURE
ON HOLISTIC SCORE (4 = highest, 1 = lowest)

		Organization		
		Strong	Weak	
Mechanics	Strong	$\bar{X}$ 3.124 SD (.957)	1.183 (.592)	organization X mechanics $p < .001$
	Weak	2.188 (.834)	1.594 (.615)	
	Differ.	.936	.219	

SentSStructure	Strong	3.000 (1.03)	1.719 (.581)	organization X sentence structure $p < .01$
	Weak	2.313 (.873)	1.688 (.644)	
	Differ.	.687	.031	

TABLE 5

READER-REWRITER MATCH/MISMATCH

	% Match	%Mismatch
Content (C)	80.2	19.8
Organization (0)	83.3	16.7
Sentence Structure (SS)	84.4	15.6
Mechanics (M)	90.6	09.4

TABLE 6

ANALYSIS OF VARIANCE FOR HOLISTIC SCORES: PERCEIVED REWRITING EFFECTS

Source	df	MS	F1	F2
Reader (R)	11	.377		
ContentPerceived(CP)	1	12.537	41.65***	31.74***
Organ, Perceived(OP)	1	5.566	19.81***	14.09***
SenSt.Perceived(SSP)	1	3.501	7.34*	8.86**
Mech. Perceived(MP)	1	1.132	3.48	2.87
CP X OP	1	.481		1.22
CP X SSP	1	.131		.33
CP X MP	1	.146		.37
OP X SSP	1	.939		2.38
OP X MP	1	.034		.09
SSP X MP	1	.368		.93

Reader Interactions

	df	MS
R X CP	11	.301
R X OP	11	.281
R X SSP	11	.477
R X MP	11	.325
Residual	31	.395

*p < .05	1,11df	F1 = 4.84	F1 based on R by Source variance
**p < .01	1,11df	F1 = 9.65	
***p < .001	1,11df	F1 = 19.69	
*p < .05	1,31df	F2 = 4.17	
**p < .01	1,31df	F2 = 7.56	F2 based on residual error variance
***p < .001	1,31df	F2 = 13.29	

TABLE 7

MEAN HOLISTIC JUDGMENTS: PERCEIVED REWRITING
(4 = highest, 1 = lowest)

	Strong	Weak	Difference
Content Percv'd	N = 64 2.578	N = 32 1.529	1.Ø5
Organ. Percv'd	N = 32 2.719	N = 64 1.672	1.Ø5
Sent.Str.Percv'd	N = 48 2.34Ø	N = 48 1.714	.63
Mechan.Percv'd	N = 48 2.356	N = 48 1.725	.63

N = 96 rewritten essays

Footnotes

[1] This topic was first developed by the California State Universities and College System for their Freshman English Equivalency Examination.

[2] The method of selection of students was extremely complex and is detailed in the author's dissertation.

[3] Two raters from one of the four groups of raters had more difficulty than any of the other raters in the sample in matching their judgments of the strength and weakness of the four rewriting categories with the rewriters' intentions. These two raters were type 3, previously judged to be among the least well qualified. Because they were two standard deviations above the mean in the amount of mismatch between their judgments and the rewriters' intentions, I replaced them with a better qualified pair: one type 1 and one type 2 rater. These replacement raters performed the evaluations together. Analyses are based on the rating given by the replacement raters.