Journal of Educational Psychology
1979, Vol. 71, No. 3, 328-338

How Characteristics of Student Essays Influence Teachers' Evaluations

Sarah Warshauer Freedman
San Francisco State University

This article explores the question of why competent evaluators award the ratings they do to college students' expository essays. Essays were rewritten to be stronger or weaker in four categories: content, organization, sentence structure, and mechanics. Twelve evaluators first used a 4-point holistic rating scale to judge the essays' quality. Then they rated whether each of the four rewriting categories in each rewritten essay was strong or weak (*perceptions*). Analyses of variance revealed content and organization to affect ratings most ($p < .001$). Mechanics and sentence structure had smaller effects, which differed when measured by the actual rewriting versus by the *perceptions*. Mechanics and sentence structure were significant in their interaction with organization ($p < .001$ and $p < .01$, respectively).

Very little is known about the process of evaluating students' writing. Most past research on composition evaluation has been correlational rather than experimental. In the usual correlational study in this area, students write papers and teacher-judges rate the quality of the papers. The researcher then examines the paper or the judges for traits associated with high and low ratings. One type of past correlational study (e.g., Hiller, Marcotte, & Martin, 1969; Nold & Freedman, 1977; Page, 1968; Slotnick & Knapp, 1971; Thompson, Note 1) attempted to predict ratings with measures of characteristics in the student papers, such as the number of spelling errors or the length of the essay. Another type (e.g., Diederich, French, & Carlton, 1961, and Meyers, McConville, & 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.

One study (Harris, 1977), which was published while this study was in progress, used a quasi-experimental design to discover what qualities within student papers most influenced high school teachers' responses. Harris had teachers rank order 12 student papers that were fixed so that the rank order would come out one way if the teachers based their judgments mostly on the category, content and organization, and another way if they cared more about a second category, sentence structure and mechanics. For the most part, the papers were chosen because of their natural strengths and weaknesses, but in some cases errors were introduced into the papers. Harris found a definite but statistically insignificant tendency for teachers to give the most weight to the content and organization category when assigning the rankings. In a questionnaire, the teachers reaffirmed the priority they gave to content and organization; however, in their comments to students they paradoxically emphasized mechanics.

To determine what within the paper influences the judge, I manipulated characteristics in student essays in a more systematic and more refined way than Harris (1977) did. Hiller et al. (1969), after completing their correlational study of student essays, first articulated the general question motivating the following experiment and

This article is based on my doctoral dissertation, *Influences on the Evaluators of Student Writing*, Stanford University, 1977. The research was supported in part by a grant from the Procter and Gamble Foundation. Special thanks go to Robert Calfee, who spent many hours helping me with all aspects of this study, from its inception to the reporting of the results.

Requests for reprints should be sent to Sarah Warshauer Freedman, School of Humanities, English Department, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132.

 0022-0663/79/7103-0328$00.75

then called for experimental research to provide a satisfactory answer:

> 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. (Hiller et al., 1969, p. 274)

I decided to manipulate four characteristics in essays: *content, organization, sentence structure,* and *mechanics.* In effect, I created four categories, whereas the Harris (1977) study had two. More precise features, which fall under these still general categories, such as the number of spelling errors or length of essay, had been the focus of many of the correlational studies in the first type cited earlier. However, for this first completely experimental study, I thought it wise to manipulate general yet pedagogically interesting 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. According to Cass and Birnbaum's (1972) most recent descriptions of admissions criteria, the colleges ranged in type from highly select, private schools to open-admissions, public schools, providing writers representing a wide range of abilities.

Students wrote the essays in class on one of eight topics designed to elicit essays in the argumentative mode of discourse. The topics asked students either 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." 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.[1]

The papers of eight students from each class, one on each of the eight topics, 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 an earlier study, four judges had rated each essay holistically (Freedman, 1977). 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 that had been rated highest and the two that had been rated lowest on each topic in the earlier study. These non-rewritten essays served as anchors for studying the reliability of the ratings in this study.

Method

Rewriting

Because of the dearth of operational definitions for

[1] Topic 1 was first developed by the California State Universities and College System for their Freshman English Equivalency Examination. Both of these topics as well as two more of the eight topics were also used in the Nold and Freedman (1977) study.

Table 1
Rewriting Rules

Strong	Weak
Content	
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.[a]
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).
Organization	
1. Paragraph appropriately.	1. Include three misparagraphings per 250-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 intraparagraph transitions: repeat key words and use transition words and phrases appropriately.	3. Delete inter- and intraparagraph transitions: vary the lexical items chosen for key words and avoid using transition words and phrases appropriately.
Sentence structure	
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, rambling, uncontrolled, awkward sentences, delete graceful parallelism, include verboseness on the sentence level.
2. Vary sentence structure.	2. Include sentence fragments and run-together sentences.
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.
Mechanics	
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 either to open or close quotations.
	3. Possessives. Misuse *'s*; omit when needed; use structures like *their's*.

Table 1 (*continued*)

Strong	Weak
	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.

Note. The operational definitions—the general rules followed for rewriting all four categories to be weak and strong—were adapted from descriptions on analytic rating scales (Adler, 1971; Diederich, 1974; Freedman, 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.
[a] Throughout the table, "include" is used to mean retain and/or add.

strength and weakness of content, organization, sentence structure, and mechanics, I pondered, at first, how to undertake the rewriting task. Based on both a study of actual student papers and on guidelines in rhetoric texts, I decided on the set of procedures in Table 1. 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 provide appropriate transitions for a group of inherently unrelated ideas. Thus, there were 12 possible rewriting combinations (C = content; O = organization; SS = sentence structure; M = mechanics; + = strong; − = weak), as follows:

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.[a] 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, arbitrarily were placed under sentence structure.

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 original essay was keyed to 3 of the 12 possible rewriting combinations listed earlier. The 4 essays, each rewritten in 3 versions, made 12 versions on each topic. The 12 rewritten versions on each topic represented the 12 possible rewritten versions. Across the 8 topics, with 12 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 were strong in content; one

third were strong in organization. Half were strong in sentence structure, and half were strong in mechanics. Of course, the remainder for each category were weak in that category.

Procedure. 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 and 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: (a) strength of professional recommendations, (b) quantity of teaching experience, and (c) educational background. All were highly recommended teachers on the staff of Stanford University'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 eight 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.

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 rated two training essays on the first topic in order to practice using the 4-point holistic scale and to practice rating essays on the topic. Then 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 supplemental 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 were 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.[2]

Reliability

To assess the reliability of the judges' ratings, I used the Cronbach alpha (Cronbach, 1970, p. 159; Calfee & Drum, in press). 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. These reliabilities are quite high but may represent the upper bounds of the reliability because they are based on papers taken from the extremes of the original distribution.

Results of Rewriting

An analysis of variance (Table 2) supports the hypothesis that essays rewritten to be strong in content would be rated significantly higher than those rewritten to be weak in content. 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.

Table 3 helps explain these main results.

---

[2] 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 2 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.

Table 2
Analysis of Variance for Holistic Scores: Rewriting Effects

Source	*df*	*MS*	*F*
Reader (R)	11	.448	
Content (C)	1	9.860	37.78**
Organization (O)	1	5.195	29.69**
Sentence structure (SS)	1	1.5	2.54
Mechanics (M)	1	5.042	9.77*
C × SS	1	1.960	6.30
C × M	1	.990	3.18
O × SS	1	3.767	12.11*
O × M	1	6.155	19.79**
SS × M	1	.001	0

Note. *F* for main effects is based on *R* by source ($df = 11$); *F* for interactions is based on residual error variance ($df = 31$). *$p < .01$. **$p < .001$.

It reveals that the difference between the average score given papers weak in content and the average score given papers strong in content was 1.06 points, a difference quite large in relation to the 4-point scale. Strong versus weak rewriting in organization also led to a difference of about 1 point. The effects of mechanics and sentence structure rewriting were 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 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.

Table 3
Mean Holistic Judgments

	Strong		Weak		
Variable	Judgment	*n*	Judgment	*n*	Difference
Content	2.375	64	1.313	32	1.062
Organization	2.656	32	1.703	64	.953
Sentence structure	2.146	48	1.896	48	.250
Mechanics	2.250	48	1.792	48	.458

Note. Scale: 4 = highest; 1 = lowest. Total $N = 96$ rewritten essays.

Table 4
Effects of Interaction Between Organization and Mechanics and Sentence Structure on Holistic Scores

	Organization			
	Strong		Weak	
Variable	*M*	*SD*	*M*	*SD*
Mechanics				
Strong	3.124	.957	1.183	.592
Weak	2.188	.834	1.594	.615
Difference	.936		−.411	
Sentence structure				
Strong	3.000	1.03	1.719	.581
Weak	2.313	.873	1.688	.644
Difference	.687		.031	

Note. Scale: 4 = highest; 1 = lowest. For Organization × Mechanics, $p < .001$; for Organization × Sentence Structure, $p < .01$.

In summary, the main results of the rewriting showed that the most significant influence on raters' scores was 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 and 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

Table 5
Reader–Rewriter Match/Mismatch

Variable	% match	% mismatch
Content	80.2	19.8
Organization	83.3	16.7
Sentence structure	84.4	15.6
Mechanics	90.6	9.4

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 percentage of match and mismatch for each category. Raters usually judged the strength and weakness of the categories accurately, although they did 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 percentage 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 that the results for content and organization were similar to those found in the main results detailed earlier. The findings for mechanics and sentence structure, this time, appear to be different. However, all of the results based on the perception data are inconclusive because some of the pairs of variables are correlated with one another (Table 7).

Table 6
Analysis of Variance for Holistic Scores: Perceived Rewriting Effects

Source	*df*	*MS*	*F*
Reader	11	.377	
Content perceived	1	12.537	41.65**
Organization perceived	1	5.566	19.81**
Sentence structure perceived	1	3.501	7.34*
Mechanics perceived	1	1.132	3.48

Note. F is based on R by source variance ($df = 11$).
$^{+}p < .05$. $^{**}p < .001$.

The chi-square analysis reveals significant correlations between the following pairs of

Table 7
Chi-Square Pairwise Correlations: Perceived Rewriting Frequencies

	Content perceived		
Variable	Strong	Weak	$\chi^2(1)$
Organization perceived			
Strong	24	7	15.97*
Weak	22	43	
	Sentence structure perceived		
	Strong	Weak	
Mechanics perceived			
Strong	31	14	13.47*
Weak	16	35	

Note. $N = 96$ rewritten essays.
$^{+}p < .001$.

Table 8
Mean Holistic Judgments: Perceived Rewriting

Variable	Strong		Weak		
	Judgment	*n*	Judgment	*n*	Difference
Content perceived	2.565	46	1.520	50	1.045
Organization perceived	2.742	31	1.677	65	1.065
Sentence structure perceived	2.340	47	1.714	49	.626
Mechanics perceived	2.356	45	1.725	51	.631

Note. Scale: 4 = highest; 1 = lowest. Total $N = 96$ rewritten essays.

main effects: perceived content with organization and perceived sentence structure with mechanics. The original rewriting design built in the content/organization correlation by not including strong organization with weak content; however, sentence structure and mechanics were originally independent variables. The raters, not being privy to the rewriting definitions, confounded these categories when giving their perceptions. When they perceived sentence structure as strong, they perceived mechanics as strong too; likewise, when they perceived sentence structure as weak, they perceived mechanics as weak. Finally, raters' perceptions of organization and mechanics were correlated to a lesser extent; raters tended to perceive mechanics to be weak when they perceived organization as weak, but mechanics was perceived as strong independent of organization. Because some of the main effects are correlated, it is not possible to assess how the interactions between them contribute to the variance in the holistic score.

Table 8 shows a comparison of the average difference between ratings on the perceived strong and weak level of each category across all the rewritten essays. Notice that the average difference between scores for strong versus weak sentence structure is almost identical to that between strong versus weak mechanics. The significance of perceived sentence structure and lack of significance of perceived mechanics in contributing to the holistic score is most likely an artifact of the nonorthogonal, correlated design.

Discussion

In the interpretation of the results, several areas deserve mention. First, for these argumentative essays all methods of analysis show that the most important influences on the raters' scores were the content and then the organization of the essay. These two aspects of the argumentative 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 influences 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 correlation between the judges' perceptions of the categories *sentence structure* and *mechanics* suggests that the raters either could not consistently distinguish between these two categories or could not correctly perceive one category as weak and the other as strong. It would be interesting to see whether raters could be trained to apply a set of definitions to the two categories and then perceive them discretely. The analysis according to the actual rewriting showed mechanics to contribute more significantly than sentence structure to the holistic score; it is unknown whether or not this finding would hold according to the raters' perceptions.

Two raters were disqualified from the research because the frequency of their mismatch was more than 2 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 mismatches 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 4-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 that 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 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 (Note 2), 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 preferred a verbal style.

Pedagogical Significance

If society values content and organization as much as the raters in this project and many of the earlier studies apparently 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 will be easily understood and favorably evaluated. The interactions between organization and mechanics and between organization and sentence structure, showing that the quality of the mechanics and sentence structure matter most when the organization is strong, point even more strongly to a pedagogy aimed at teaching the skills of organization before or at least alongside those of mechanics and sentence structure. At the very least, teachers should not value content and organization while commenting to students mostly on mechanics as those in the Harris (1977) study did.

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 (Christensen, 1967; Hunt, 1965; Mellon, 1969; O'Hare, 1971) 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 *The Christensen Rhetoric Program* (1968). Other scholars, like Kenneth Burke (1945), D. Gordon Rohman (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 nonmainstream 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.

Reference Notes

1. Thompson, R. *Predicting writing quality, writing weaknesses that dependably predict holistic evaluations of freshman compositions.* English Studies Collections, Series 1, No. 7, 1976. (Available from

Scholarly Publishers, 172 Vincent Drive, East Meadow, New York 11554.)

2. Williams, J. *Nominal and verbal styles: Some affective consequences.* Chicago: University of Chicago, 1977. (Mimeograph)

References

Adler, R. *An investigation of the factors which affect the quality of essays by advanced placement students.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1971.

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

Calfee, R., & Drum, P. How the researcher can help the reading teacher with classroom assessment. In L. Resnick & P. Weaver (Eds.), *Theory and practice of early reading.* Hillsdale, N.J.: Erlbaum, in press.

Cass, J., & Birnbaum, M. *Comparative guide to American colleges* (5th ed.). New York: Harper & Row, 1972.

Christensen, F. *Notes toward a new rhetoric: Six essays for teachers.* New York: Harper & Row, 1967.

Christensen, F. *The Christensen rhetoric program.* New York: Harper & Row, 1968.

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

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

Diederich, P. *Measuring growth in English.* Urbana, Ill.: National Council of Teachers of English, 1974.

Diederich, P., French, S., & Carlton, S. *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton, N.J.: Educational Testing Service, 1961.

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

Freedman, S. *Influences on the evaluators of student writing.* Unpublished doctoral dissertation, Stanford University, 1977.

Harris, W. Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis for teacher judgment of student writing. *Research in the Teaching of English,* 1977, *11,* 175–185.

Hiller, J., Marcotte, D., & Martin, T. 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.: 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., McConville, C., & Coffman, W. 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, 1968.

Nold, E., & Freedman, S. 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., & Knapp, J. Essay grading by computer: A laboratory phenomenon? *Educational Measurement,* 1971, *9,* 253–263.

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

Received June 15, 1978 ■