Research

# Some Reasons for the Grades We Give Compositions

Sarah Warshauer Freedman

Susan took the first semester of required college freshman writing from Ms. B. and failed. Second semester, she repeated the class with Ms. R. In a first week writing conference, Ms. R. routinely asked each student about past writing problems. Susan admitted, "I'm not quite sure what my problems are because it's kind of hard to interpret from one class to another class." Susan felt confused about how to interpret the responses of different teachers to her writing, but she understood that she would be more likely to pass the course if she knew how to compare her different teachers' grading and commenting practices.

Concerned about Susan's confusion, Ms. R. said, "It'll be interesting when you get your papers back from me and I mark them to see if they're somewhat the same kind of comments [as those of Ms. B.]." Responding to her teacher's openness, Susan made an unusual request: she asked Ms. R. to mark clean copies of the papers she had written for Ms. B. Ms. R. agreed.

From their own points of view, Susan and Ms. R. were exploring one of the most difficult and important problems in writing classes. Ms. R. wanted to discover the fairest and most useful ways to grade and comment on student papers. Susan wanted to discover how to interpret and learn from her teachers' responses.

Grades on compositions are often difficult for students to predict and interpret because students frequently do not understand many of their teachers' written comments (Hahn, 1981). Comments are often overgeneralized, ambiguous, and uninterpretable (Sommers, 1981). And when they are interpretable, students may not always read what we write carefully or thoughtfully. When students do not understand how we evaluate their work, they tend both to devalue advice and to have difficulty learning from us. Hirsch (1977) claims that writing assessment is "the single most important snag to practical progress in composition teaching and research."

In the 1950s, research on why evaluators respond as they do interested Educational Testing Service (ETS). ETS needed to solve a practical problem, how to reliably score the essays which had recently become part of their standardized tests. After an extensive review of the literature on reliability (agreement among raters), Edith Huddleston concluded "that the unreliability of essay examinations is most pronounced in the area of English composition" (p. 165).

For the first ETS project, Diederich, French, and Carleton (1961) had 300 papers written by college freshmen rated by 53 readers from six fields—college English teachers, social science teachers, writers, editors, lawyers, and business executives. These readers rated in their homes, much as English teachers evaluate papers. Readers were directed to write brief comments on as many papers as they could and to assign each paper a general merit score.

After finding gross disagreements among the evaluators, Diederich attempted to discover the cause. Statistical analyses revealed that raters giving similar scores fell into five clusters. Analyses of the comments of the raters in each cluster led Diederich to hypothesize that the clusters were formed around the part of the essay the cluster members valued most: (1) ideas, (2) usage, sentence structure, punctuation, and spelling, (3) organization and analysis, (4) wording and phrasing, or (5) personal qualities. Although most of cluster two, the mechanics group, was made up of English

teachers, the other clusters were not strongly associated with particular occupations.

After completing the study, the Diederich team developed an analytic rating scale so raters could judge a composition separately on each of its apparently salient dimensions. Each part of the scale was keyed to the values of one cluster of Diederich's readers. Diederich reasoned that on the new scale, the readers' different values could not influence their scores, and readers would rate more reliably, more consistently, more dependably. But the reliability problems remained. Because of methodological problems in the Diederich study (see Freedman, 1977, for further explanation), the theory that led to creating the scale was flawed. The scale never gained popularity because it was time consuming and limited to the expository prose of older students.

In a 1966 ETS study, Meyers, McConville, and Coffman found that a homogeneous group of experienced English teachers also fell into agreement clusters, but the clusters could be explained by the fact that some raters were more lenient than others. They suggest:

> in contrast to the Diederich, French, and Carleton study, that, although all of the readers may have assigned different weights to the various attributes of a composition (e.g., spelling, grammar, organization), *each of the individual attributes tended to be given the same weight by the readers* (italics mine). That is, it does not seem as though some judges rated the papers primarily upon grammar, while others rated the papers primarily upon style. Rather, this explanation suggests that the relative weighting between grammar and style is *roughly* the same for all judges (p. 52).

Unfortunately for teachers, without finding out anything else about why evaluators respond as they do, ETS solved their reliability problems empirically; by summing the scores of three raters with homogeneous backgrounds rating on a four-point holistic scale, they could get respectable reliability (Godshalk, Swineford, and Coffman, 1966). Later, ETS discovered that they could obtain even better reliability if the raters were trained to agree with one another. Miles Myers describes standard holistic rating practices today: "the most reliable scoring of writing samples takes place when the readers are trained together and read together in the same room under common direction" (p. 26).

The ETS research suggests that experienced English teachers can apply the same set of values to papers they evaluate and, with training, they can adjust one difference, their degree of leniency. To move toward reducing the frustration of students like Susan, we need to know precisely what properties of student essays experienced teachers value most.

Researchers have examined easily countable qualities within essays that correlate with teacher scores (e.g., Page, 1968; Nold and Freedman, 1977; Grobe, 1981). The findings are unremarkable. Consistently, essay length is found to correlate with scores for essays; the longer the essay, the better the score. Correlational research shows associations but cannot explain why raters evaluate as they do. Just because essay length is associated with essay quality, we would not advocate that students pad their essays. However, we might hypothesize that essay length is related to other factors that might cause teachers to score as they do.

I designed an experiment to try to discover what causes teachers to score as they do. I particularly wanted to know how experienced teachers weigh each part of an essay when they assign a score to the whole (Freedman, 1977, 1979-a, 1979-b; Freedman and Calfee, in press). I wanted to know whether English teachers value mechanics above development and organization as the Diederich study implied they might or whether teachers value development most, as the correlational studies suggest.

I selected a set of previously scored, average quality, expository essays by college freshmen and had them rewritten to be weaker or stronger in (1) development or content, (2) organization, (3) sentence structure, and (4) mechanics. Then I had these rewritten essays judged by groups of English teachers who did not know that the essays had been tampered with. Since I knew which essays were strong or weak in which qualities, I could examine scores to see what types of strength or weakness mattered most to the teachers.

In an earlier study using a partial rewriting technique, Harris (1977) examined two parts of the essay, (1) content and organization and (2) sentence structure and mechanics. She found that high school teachers would rank-order papers on the basis of strengths or weaknesses in content and organization rather than on the basis of the strengths or weaknesses of sentence structure and mechanics.

My study confirmed Harris' findings and yielded interesting additional results. I found

teachers gave significantly higher scores to papers rewritten to be strong in development, organization, and mechanics than they did to papers rewritten to be weak in these areas. That's to be expected, but not all the positive points are weighed in a parallel way, and *the weighing shows our hierarchy of values.* First, sentence structure rewriting, in and of itself, proved a negligible effect on the raters, unlike the other types of rewriting. The development rewriting was most influential, the organization rewriting next, and mechanics third.

The development rewriting proved so powerful that, given strong or weak development, other types of strength or weakness did not matter much.

However, if the organization was strong, then the quality of the mechanics mattered even more than it did on its own and the quality of the sentence structure mattered as well. Teachers in this study apparently felt that if the discourse level qualities were weak, the paper deserved a low score, and *strong sentence structure and mechanics could not redeem such papers.* But once the discourse level qualities were strong, the lower levels mattered; strengths or weakness of mechanics or sentence structure could then raise or lower a score. English teachers apparently care about sentence structure only under conditions in which the organization is already strong and about mechanics most under that condition.

The excellent research on the writing process lends theoretical support for our hierarchy of values: development, then organization, then mechanics and sentence structure once the higher levels of the discourse are under control. Remember that these are values which we apply to finished products; they do not necessarily dictate a curricular sequence.

Others have rewritten student essays to determine what influences the evaluator (e.g., Piche, Michlin, Rubin, and Turner, 1978; Nielsen and Piche, 1981; Hake and Williams, 1981). They have looked mostly at sentence level or mechanical parts of the student essay in some detail—e.g., black dialect features or nominal versus verbal style.

In an interesting correlational study, Thompson (1976) examined how several specific aspects of these discourse level categories were related to teacher's evaluations. He found unsupported statements, lack of unity, and independent judgment errors (flaws in arguments, oversimplifications of topics, and lack of proper inferences) accounted for holistic scores much better than problems with mechanics, coherence, or wordiness. Thompson's discourse level categories fall within the category that I labeled development.

None of this research took place in the classroom. In a later study, Thompson (1981) illustrates how successful we can be in the classroom if we communicate expectations to students. He found students could be trained to develop and reliably apply a set of standards about essays. Most interesting, he found students learned to understand and apply their teacher's standards. But until teachers articulate the bases of their scores and act consistently with one another, students like Susan still may have difficulty making the transition from one class to the next.

I would guess that one of the appeals of group editing is that the sessions help students clarify why papers are evaluated as they are—not just why they receive a particular grade but also why they receive particular comments.

Recently, a freshman at Berkeley, a participant in a research project, told me that she did not feel she learned anything in her writing class. She thought she began the class as an adequate writer. At the end of the course she received a *B*; she wanted an *A*. Although the grade was important to her, more important was the fact that she had no idea what she needed to do to write *A* papers. Later her teacher told me that he thought the student began the class as a good writer and her writing had improved, but not to *A* level.

We need to let competent, highly motivated writers know that they should reach higher and then show them how to reach. It is only too easy to praise such students and thereby fail to demand the achievement we might.

I recently began a descriptive study of how experienced teachers respond to student papers in individual conferences (Freedman, 1981). Besides providing detail about teachers' responses as they occur, I hope to develop hypotheses about how experienced teachers communicate values about writing to students. So far, I have found that students have a set of concerns about their writing when they enter a class and what they *hear* from the teacher depends on how well the teacher *listens* to them. Teachers seem to be most successful in helping students *hear* if they first listen and respond to students' concerns.

From yet another point of view, Hirsch (1977)

considers global assessment issues outside writing classrooms. He claims writing assessment must "be consistent with judgments about good writing in literate society at large" and that "no . . . professional assessment should be at odds with the verdict of that (society's) court" (p. 177). Before accepting Hirsch's view, two considerations are important. First, society often says it values mechanics most when in fact it may value larger levels of discourse more. Teachers in my rewriting study could not have told me their values as clearly as they revealed them to me under testing. We need to distinguish what society claims it values from what it actually values. Second, regardless of what society claims to value, as English teachers it is our place to guide society to see our point of view. We are professionals. It is crucial that we communicate our point of view clearly not only to society, but also to students.

**References**


Diederich, Paul, John French, and Sydell Carlton. *Factors in Judgments of Writing Ability.* Research Bulletin RB-61-15. Princeton: Educational Testing Service, 1961.

Freedman, Sarah. "Evaluation in the Writing Conference: An Interactive Process." *Selected Papers from the 1981 Texas Writing Research Conference,* Maxine Hairston and Cynthia Selfe, eds. Austin: University of Texas, 1981.

Freedman, Sarah. "How Characteristics of Student Essays Influence Teachers' Evaluations." *Journal of Educational Psychology* 71 (July 1979): 328-338.

Freedman, Sarah. "Influences on the Evaluators of Expository Essays: Beyond the Text." *Research in the Teaching of English* 15 (October 1981): 245-255.

Freedman, Sarah. *Influences on the Evaluators of Student Writing.* Unpublished doctoral dissertation, Stanford University, 1977.

Freedman, Sarah. "Why Teachers Give the Grades They Do." *College Composition and Communication* 30 (May 1979): 161-164.

Freedman, Sarah and Robert Calfee. "Holistic Assessment of Writing: Experimental Design and Cognitive Theory." *Research in Writing: Principles and Methods.* Peter Mosenthal, Lynn Tamor, and Sean Walmsley, eds. New York: Longman, in press.

Godshalk, Fred I., Frances Swineford, and William Coffman. *The Measurement of Writing Ability.* New York: College Entrance Examination Board, 1966.

Grobe, Stewart. "Syntactic Maturity, Mechanics, and Vocabulary as Predictors of Quality Ratings." *Research in the Teaching of English* 15 (February 1981): 75-85.

Hahn, J. "Students' Reactions to Teachers' Written Comments." National Writing Project Network Newsletter 4 (1981): 7-10.

Hake, Rosemary L. and Joseph M. Williams. "Style and Its Consequences: Do as I Do, Not as I Say." *College English* 43 (September 1981): 433-451.

Harris, Winifred Hall. "Teacher Responses 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* 11 (Fall 1977): 175-185.

Hirsch, E. Donald. *The Philosophy of Composition.* Chicago: University of Chicago Press, 1977.

Huddleston, Edith. "Measurement of Writing Ability at the College Level: Objective vs Subjective Testing Techniques." *Journal of Experimental Education* 22 (March 1954); 165-213.

Myers, A., Carolyn McConville, and William Coffman. "Simplex Structure in the Grading of Essay Tests." *Educational and Psychological Measurement* 26 (Spring 1966): 41-54.

Myers, Miles. *A Procedure for Writing Assessment and Holistic Scoring.* Urbana, Illinois: National Council of Teachers of English, 1980.

Nielsen, Lorraine and Gene Piche. "The Influence of Headed Nominal Complexity and Lexical Choice on Teachers' Evaluation of Writing." *Research in the Teaching of English* 15 (February 1981): 65-73.

Nold, Ellen and Sarah Freedman. "An Analysis of Readers' Responses to Student Writing," *Research in the Teaching of English* 11 (Fall 1977): 164-174.

Page, Ellis. "Analyzing Student Essays by Computer." *International Review of Education* 14 (1968): 210-225.

Piche, Gene, Michael Michlin, Donald Rubin, and L. Turner. "Teachers' Subjective Evaluation of Standard and Black Nonstandard English Composition: A Study of Written Language Attitudes." *Research in the Teaching of English* 12 (May 1978): 107-118.

Sommers, Nancy. "Responding to Student Writing." *National Writing Project Network Newsletter* 3 (1981): 7-11.

Thompson, Richard. "Peer Grading: Some Promising Advantages for Composition Research in the Classroom." *Research in the Teaching of English* 15 (May 1981): 172-174.

Thompson, Richard. "Predicting Writing Quality, Writing Weaknesses that Dependably Predict Holistic Evaluations of Freshman Compositions." *English Studies Collections, 1* (1976). (Available from Scholarly Publishers, 172 Vincent Drive, East Meadow, New York 11554).



*Sarah Warehauser Freedman teaches at The University of California, Berkeley.*

## References

**Why Do Teachers Give the Grades They Do?**
Sarah Warshauer Freedman
*College Composition and Communication*, Vol. 30, No. 2. (May, 1979), pp. 161-164.
Stable URL:
http://links.jstor.org/sici?sici=0010-096X%28197905%2930%3A2%3C161%3AWDTGTG%3E2.0.CO%3B2-5

**Style and Its Consequences: Do as I Do, Not as I Say**
Rosemary L. Hake; Joseph M. Williams
*College English*, Vol. 43, No. 5. (Sep., 1981), pp. 433-451.
Stable URL:
http://links.jstor.org/sici?sici=0010-0994%28198109%2943%3A5%3C433%3ASAICDA%3E2.0.CO%3B2-R