Contributors should follow the format of the APA *Publication Manual*, Second Edition. Each manuscript should include a cover sheet with the author's name, position, telephone number, and address; identifying information should not appear elsewhere in the manuscript. Captions should be appropriately placed, and tables should be typed on separate sheets. An abstract approximately 150 words in length should also be included. In order to facilitate the review process, three copies of each manuscript should be submitted; only one copy will be returned to the author. A self-addressed manila envelope, to which stamps or Universal Postal Union coupons are clipped, should accompany each manuscript. All editorial correspondence should be addressed to Roy C. O'Donnell, P.O. Box 1991, Athens, Georgia 30603-1991.

The journal staff makes every conceivable effort to be honest with its readership. Accordingly, those submitting advertisements for world literature and American literature texts are asked to include information which accurately describes the content and scope of the material, and to submit *only* those advertisements which truthfully describe the materials. Criteria for determining this standard of truth are available on request from NCTE. It is the policy of NCTE in its journals and other publications to provide a forum for the open discussion of ideas concerning the content and the teaching of English and language arts. Publicity accorded to any particular point of view does not imply endorsement by the Executive Committee, the Board of Directors, or the membership at large, except in announcements of policy, where such endorsement is clearly specified.

RESEARCH IN THE

TEACHING OF ENGLISH

VOLUME 15, NUMBER 3, OCTOBER 1981

Contents

ARTICLES

Research in the Teaching of English, Vol. 15, No. 3, Oct. 1981: 245-255

Influences on Evaluators of Expository Essays: Beyond the Text

SARAH WARSHAUER FREEDMAN
University of California, Berkeley

ABSTRACT

This study examines the relative effects on holistic scores given to college students' expository essays of three types of variables—essay variables, reader variables and environment variables. Sixty-four essays by students at four colleges were judged by four readers using a holistic scale and by two readers using a Diederich-type analytic scale. The essays were on eight topics. Readers were trained by two trainers. Of the three types of variables, the essay contributed most significantly to the variance in the holistic scores ($p<.001$). One of the environment variables, the trainers, contributed next most significantly ($p<.01$). Another environment variable, the topic, also affected the holistic scores. The readers judged consistently, their traits not affecting the variance in the scores. Finally, a comparison of ratings on the holistic and analytic scales revealed that the only additional information over the holistic score that the analytic scale yielded was a usage score.

In recent years, the student essay has become an increasingly common item on many standardized writing tests (e.g., the College Entrance Examination Board's *English Composition Test*, Educational Testing Service's *English Placement Test* in California and New Jersey, and local proficiency tests around the country such as the *Junior English Proficiency Essay Test* at San Francisco State University). Unlike multiple choice items on paragraph organization, sentence structure, and the like that accompany many of these essays, the essays cannot be scored by machine. Rather, they usually are scored in holistic evaluation sessions by large groups of hand-picked, trained, teacher-readers. Since the essay provides the only direct measure of writing and since the scoring is not objective in the way that scoring a multiple choice item is, exact knowledge about why the essays are scored as they are yields information about how to interpret their very special scores. This study aims to provide information about such essay scoring, in particular about what influences readers to give the scores they do to expository, argumentative essays written as part of a writing test.

Several types of variables can influence the score that this type of essay receives. First of all, variables within the *essay* itself should influence the score —does the essay show sensitivity to the reader's needs by supplying enough background information, is the essay well organized, does the essay adhere to the surface conventions demanded in formal written prose? Second, variables within the *reader* may influence the score—is the reader strict or lenient, does

the reader comprehend well, is the reader biased in any way? And finally, variables within the *environment* of the rating may influence the score—is the room well lit, is it morning or afternoon, who is doing the training and how is the training being conducted?

Past research on holistic rating falls into two types. In the first type, the researcher attempts, just as writing testers do, to hold the readers constant. Readers are chosen because they are a homogeneous group of English teachers who can discriminate well between papers with different characteristics and who can be trained to adjust their values with respect to those characteristics and thus rate reliably, that is agree well with one another. Once this homogeneous group of readers, who will probably score reliably, has been selected, the researcher attempts to determine what variables within the essays influence the readers to raise or lower their scores. The following studies take this approach: Page, 1968; Hiller, Marcotte, and Martin, 1969; Slotnick and Knapp, 1971; Thompson, 1976; Nold and Freedman, 1977; Harris, 1977; Freedman, 1979 a, b.

The second type of researcher on holistic rating attempts to study variables within readers. Most important here is the work of Diederich, French and Carlton (1961) who, by studying heterogeneous, untrained readers, proved, as Diederich wrote later, "how commonly and seriously teachers disagree in their judgments" (1974, p. 5). By analyzing the comments of the disagreeing readers Diederich and his colleagues contributed to knowledge about the basis of readers' natural disagreements and provided the field with ideas on how to train readers to lose their biases and to agree better with one another. He also developed an analytic rating scale (Diederich, 1974) which is based largely on the types of comments his readers made. Meyers, McConville, and Coffman (1966) examined how 25 readers rated 25 of the 20-minute essays written for the 1963 *English Composition Test*, observing whether the readers actually rated globally and checking to see whether or not their scores reflected different values about the qualities requisite to good compositions. Meyers et al. found that some raters were more lenient than others but that otherwise their values about the essays seemed to be similar.

Both types of past research only partially answer the fundamental question: why do evaluators award the scores they do to student papers? It should be noted that no one has even begun to examine the effects of the rating environment on raters' scores.

In this study, I explore the relative effects of the three sources of variables —essay, reader, and environment (including essay writer)—on the essay score and the particular effects within the environment that affect the score.

METHOD FOR WRITER AND TOPIC SELECTION

Writers

Four San Francisco Bay Area colleges, providing writers representing a wide range of abilities, participated in the study. 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. A sample of eight students was taken from each of two classes at each institution—16 students per school, 64 writers in all. The classes were obtained on the recommendation of the department chair who was asked to suggest two "typical" classes taught by different teachers. In order to assess further the typicality of the classes, the investigator also interviewed the teachers. Writers were selected from within the two classes at each school.

Topics

I selected eight essay topics, four asking students to compare and contrast two *quotations* and four asking students to argue their *opinion* on a current controversial issue. The following topics illustrate each type:

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.[1]

2. The Supreme Court has ruled that no state may deny a woman an abortion within the first six months of her pregnancy. Do you agree or disagree with this decision? Give reasons for taking your position.

The other eight topics, modelled after those above, can be found in Freedman, 1977.

This study is limited to evaluations of papers on topics in the argumentative mode of discourse so that differences in evaluations caused by mode of discourse would not have to be examined. The argumentative mode was chosen because it is so frequently demanded of college students in test situations.

PROCEDURE FOR COLLECTING ESSAYS

Each student in the class received one of the eight topics and wrote during forty-five minutes in his or her class. Only eight essays, one on each topic, were used in the study. Teachers distributed the essay topics in all except one Stanford class in which the investigator administered the essay.

After all essays were collected, the investigator coded them and had them typed to conceal the identity of the writers and to facilitate the evaluation process. Typing also avoided the inevitable effects of handwriting on raters (Markham, 1976). The essays were transcribed exactly, including all errors.

METHOD FOR EVALUATING ESSAYS

I aimed to examine *how* well-qualified readers judge essays. Therefore, for this first evaluation of the essays, I selected the four most qualified readers available, whom I paid to perform the evaluations. All expected to complete their doctorates in English literature at Stanford in the immediate future; all had taught writing requirement classes for at least three years. Their superiors in the English department at Stanford recommended them as excellent teachers and as able evaluators of college student prose.

Table 1 contains the plan for collecting the holistic ratings during four sessions on two mornings. During each of the four sessions on each day, two

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

readers rated one topic while the other two rated a different topic. One pair rated an opinion topic; the other pair rated a quotation topic. Each pair was trained by a different trainer. The pairs of readers changed, rotating in a balanced way from one session to the next. The balancing insured that both quotation and opinion topics were read first and last, so that the effect on the readers' scores of the order of judging each type of topic could be examined. The balancing also insured that both trainers would train all raters, so that the effect of a particular trainer on a particular rater could be examined. Furthermore, the same readers did not always evaluate together, so that agreement between readers could be determined not to be merely pair specific. No time limits were set for the length of the sessions beforehand so that the evaluators would feel no pressure to rush through their task. Analytic ratings were collected during the afternoons.[2] On the first afternoon, sessions I and II of the morning were repeated; on the second afternoon, sessions III and IV of the first morning were repeated.

TABLE 1

Plan for Collecting Ratings

Day 1

	X	Y
I	RS-3	TU-8
II	RU-4	TS-2
III	TU-6	RS-5
IV	TS-7	RU-1

Day 2

	X	Y
I	TU-5	RS-6
II	TS-1	RU-7
III	RS-8	TU-3
IV	RU-2	TS-4

Key to abbreviations: R, S, T, and U are the four readers; X and Y are the two trainers; the quotation paper topics are 1, 2, 3, and 6; the opinion topics are 4, 5, 7, and 8. I, II, III, and IV refer to the different rating sessions.

The training packets for each topic contained two training essays. Holistic scoring forms were included for training before holistic evaluations; an analytic scale and scoring forms were included for training for the analytic evaluations. In the reading packets several additional training essays preceded the eight experimental student essays. The reading packet always consisted of eleven essays with the potential training essays first and the eight student essays next. The eight student essays for the study were randomized twice for each reader, once for the holistic ratings and once for the analytic ratings.

[2] The analytic rating scale was adapted from scales developed by Diederich (1974) and Adler (1971) and can be found in Freedman (1977).

PROCEDURE FOR EVALUATING ESSAYS

The evaluators judged the original essays on two Saturdays, two weeks apart. Before the evaluations, the readers were informed only that all essays were produced mostly by college students during a 45-minute, in-class session.

On the first morning every reader evaluated four of the eight topics according to the four-point holistic rating scale. The holistic rating sessions lasted approximately one hour each. The analytic scale demanded extra time for rating; the afternoon analytic evaluation sessions lasted about two hours each.

Pairs of readers were trained for each topic before every holistic and analytic rating. Training aimed to establish a realistic context for the rating and to provide opportunity to practice applying the rating scales.

In the end, all four readers judged all eight topics with the holistic scale. Every reader judged the papers on four topics with the analytic scale, two readers judging each topic analytically. So each paper received four holistic and two analytic ratings.

RESULTS

Evaluator Reliability

To establish the readers' reliability, Cronbach's alpha (Cronbach, 1970) was computed. Even with only two readers giving a score to each essay on day one and on day two, the consistency of the differences in the papers proved quite high ($\alpha = .58$ and $\alpha = .62$) while there was little consistent difference between the readers ($\alpha = .0$). After combining the readers' scores across both days so that each paper received four scores, the reliability of the differences in papers was even higher ($\alpha = .84$) and the consistency of differences between readers was extremely low ($\alpha = .20$). Thus, it was concluded that all four of the readers consistently agreed with each other on the scores they gave the student papers, that indeed the group rated homogeneously.

Influences of Essays, Topics, and Evaluation Procedure on Holistic Evaluations

An analysis of variance was performed to determine how parts of the readers' environment during the rating affected differences in scores (Table 2). In comparing the effects of the reader, the essays and parts of the environment on the essays' scores, it was found that characteristics of the essays themselves contributed most to the scores ($p<.001$). The only other contributor to the scores proved to be one part of the environment, the trainers. The readers did not contribute significantly to the differences in the scores they gave, an indication of their homogeneity.

Although the different topics and different types of topics did not affect the scores significantly, readers gave higher scores to one of the opinion topics. A further analysis of how the two separate types of topics contributed to the variance between the holistic scores revealed a significant difference between the scores given the different opinion type topics (Table 3). Although these similar topics did not call for different scores generally, topics can affect raters' scores.

TABLE 2

Analysis of Variance for Holistic Scores: Contributions of Parts of the Environment for the Reading

Source	df	MS	F
Reader (R)	3	.45	1.33
Topic (T)	7	2.42	1.81
Type (Ty)	1	3.75	5.95
T (Ty)	6	2.20	1.50
Essays (T(Ty))	56	2.27	6.72***
Sessions	3	.75	2.28
Day	1	.09	.27
Trainer	1	3.28	9.72**
R × T	21	1.34	
R × Ty	3	.63	
R × (T(Ty))	18	1.46	
R × Essays (T(Ty))	163	.37	

F ratio for the sources T, Ty, and (T(Ty)) are based on the corresponding R interactions (i.e., R × T, R × Ty, and R × (T(Ty))).
All other F ratios are based on the main residual, R × Essays (T(Ty)).
** $p < .01$
*** $p < .001$

Relationship Between Holistic And Analytic Rating Scales

Table 4 details the Pearson product moment correlations between the holistic score, six categories of the analytic rating scale, and the sum of the scores on the categories of the analytic scale. The correlation matrix reveals high correlations between all scores except the ones for usage.

Because of the extent of the correlations, a factor analysis was carried out to determine a simpler structure for the rating scales. To examine the categories of the analytic scale with the factor analysis, I created three orthogonal, linear contrast scores from the categories of the scale. Table 5 shows how the numerator for each contrast was computed. In forming the first contrast, *content/style*, I hypothesized that three of the categories, voice, development, and organization, should measure most directly the overall content of a composition and that sentence structure, word choice, and usage most likely measure more micro-stylistic aspects of a composition. The first contrast opposed these two subcategories of the analytic scale, content and style. Next I hypothesized that

voice was somewhat discrete from development and organization and that usage was somewhat discrete from word choice and sentence structure. Voice should measure the personality behind the prose; development and organization should

TABLE 3

Holistic Evaluations by Topics

Summed Holistic Scores for Each Topic

Quotation				Opinion			
63	66	72	70	76	73	90	63

Analysis of Variance for Holistic Scores: Contributions of Topic Types

	df	MS	F1
Readers	3	.45	
Topic Type	1	3.75	
Topic (Q)	3	.51	1.21
Topic (O)	3	3.89	7.66***
Essays (Topic (Q))	28	2.64	6.33***
Essays (Topic (O))	28	1.90	3.75***
Readers × Essays (Topic (Q))	84	.42	
Readers × Essays (Topic (O))	84	.51	

*** $p < .001$

TABLE 4

Pearson Correlations of Different Ratings

	HSC	ASC	V	D	O	SS	WC	US
Holistic Score (HSC)	1.00	.76	.67	.76	.74	.63	.61	.36
Analytic Score (ASC)		1.00	.86	.93	.91	.83	.84	.63
Voice (V)			1.00	.83	.73	.64	.66	.41
Development (D)				1.00	.84	.73	.74	.46
Organization (O)					1.00	.75	.69	.45
Sentence Structure (SS)						1.00	.75	.57
Word Choice (WC)							1.00	.58
Usage (US)								1.00

measure how the writer treated the content in terms of expanding and ordering it. This contrast was labeled *voice/content*. Usage should measure purely mechanical aspects of standard edited English like punctuation and capitalization. Sentence structure and word choice, on the other hand, should not be dictated completely by matters of convention and thus should have more to do with the writer's actual style. This third contrast was labelled *usage/style*.

TABLE 5

Factors from Rating Scales

Linear Contrast Scores from Analytic Scale	
Content/Style =	V + D + O − SS − WC − US
Voice/Content =	2V − D − O
Usage/Style =	SS + WC − 2US

Factor Loadings

	Factor 1	Factor 2
HSCORE	.837	.326
ASCORE	.845	.148
CONTENT/STYLE	.259	.782
VOICE/CONTENT	−.255	.011
USAGE/STYLE	−.003	.583

Abbreviations: V = voice; D = development; O = organization; SS = sentence structure; WC = word choice; US = usage; HSCORE = holistic score; ASCORE = analytic score

A principal component factor analysis with varimax rotation was performed on the correlation matrix of the three linear contrast scores, the holistic score, and the sum of the categories on the analytic scale. The results of the factor analysis, also in Table 5, reveal that the five contrast scores represented two discrete, independent qualities of the papers, two factors. The two variables loading highest on Factor 1 after rotation were the holistic and summed analytic scores. The variables loading highest on Factor 2 proved to be the content versus stylistic categories and within the stylistic categories usage versus the others. In short, these two factors suggested that holistic and analytic scales measured one trait of compositions, and that a stylistic category can be separated from a content category with usage probably dominating the separation.

DISCUSSION

The unexpected significant effect on the holistic scores caused by the different trainers was disturbing. The investigator and a colleague, who is a known expert on evaluation, conducted the training sessions. Both trainers worked together before the rating sessions to plan and standardize the training. We purposefully aimed to avoid the training effect by selecting training essays on each topic that represented the range of the responses and by agreeing on the scores we thought these training essays deserved. For the training, we planned first to discuss the topic and then to present these essays to the readers.

After finding differences in scoring due to trainers, I reviewed tape recordings of the training sessions, looking for differences in the *actual* training that might have contributed to the effect. I found first that the two trainers approached the discussion of the topics differently. The first trainer simply presented the topic, and the readers discussed its meaning. The second trainer asked the readers to discuss what they would expect from a good essay on the topic. During the discussion, this trainer reminded the readers that the students only had 45 minutes to complete the task. Such a discussion could have lowered the readers' expectations for a good essay and encouraged them to award higher scores. The readers did award higher scores under this trainer.

Interestingly, during the discussion of the training essays, the trainers stuck to their agreed upon scores, but sometimes they differed in how they handled the discussion about the scores. For example, in one case in which the trainers had previously agreed on a 2 for a particular essay, the two readers in each session each gave a 2 and a 1. Both trainers revealed that they had agreed on a 2 for the essay, but the first trainer then continued by saying that the trainers thought the essay represented the low range of the 2's and that they could understand a 1 score. Given the same situation, the second trainer continued by saying that 1's should be saved for worse essays. Again, it was this second trainer who elicited higher scores from the readers.

Although I cannot say why the trainers caused scoring differences, in retrospect I hypothesize that subtle and not such subtle remarks can push readers to score higher or lower, regardless of the training essays. It stands to reason that if readers can be trained to agree, they can also be trained, advertently or inadvertently, to score higher or lower. Since readers have a natural tendency to rate low, to shy away from using the upper end of the scoring range, it is traditional in training sessions for holistic scoring to push readers to raise their scores (see "How the Essay in the CEEB English Composition Test is Scored: An Introduction to the Reading for Readers," ETS mimeo, 1976, p. 4). But no one has investigated exactly what it takes for a trainer to effect that push, and no one has explored whether or not readers can be pushed too much.

This study establishes the fact that training changes readers' rating behavior. To establish the cause of the trainer effect, I recommend experiments in which carefully described training conditions are varied. I know of no cases in which the differential effect of trainers has been investigated. Usually small research projects employ only one trainer; large scale readings, as for College Board or Educational Testing Service employ many trainers, but all are under the direction of one chief trainer, the question leader. In small scale projects the investigator probably should not train the evaluators. Since trainers affect evaluators and since the investigator knows the results he or she expects, the investigator may help to produce, either intentionally or unintentionally, the hypothesized results with the training. In all projects, the effect of trainers needs to be monitored more precisely; what is it that trainers do and what causes their influence?

Although this study shows that topics generally do not affect readers' scores, it shows that topics can have significant effects. Because in most testing situations the topic must change from one administration to the next in order to insure the security of the test, topics should be pretested carefully enough to establish their equivalence with past topics. Indeed because both topics and trainers usually change with the test administration and because both of these variables can significantly affect essay scores, it becomes impossible to compare

test results from one administration to the next; for that matter, it becomes difficult to compare research results which do not account for the effects of topics and trainers.

The high correlation between the holistic scale and all categories of the analytic scale except usage indicates that researchers and testers should not use the more time-consuming analytic scale. They will gain little more from it than from a holistic rating. Techniques for gaining additional information beyond a usage score remain to be discovered. It is possible that the scales showed such high correlations in this study because the design allowed the evaluators to remember the holistic scores they gave, and the raters adjusted their analytic score to match the holistic scores. This seems unlikely, however, because there was a two-week lapse between the holistic and analytic ratings for half of the papers. Also, the raters were instructed that if they remembered their earlier holistic scores, they were not to be concerned about giving different ratings to the same paper because the analytic scoring allowed them extra time to ponder. It was expected that this extra thought would allow them a chance to "correct" any "mistakes" they made during the snap judgment required by the holistic scoring.

In general, the strong relationship between the two scales could be due to the fact that the raters were unable to separate the categories. Descriptions of the categories on analytic scales must be written to remove any inevitable relationship between them. I revised the Diederich and Adler analytic scales so that the categories would be described more discretely. However, my rewriting for some of the categories was probably not extensive enough. For example, the low range under "voice" is described as follows: "The reader gets the sense that the writer has nothing to say on the topic. It is either difficult to tell what this writer is trying to get across, or the thoughts are so silly as to be better left unsaid." It is difficult to imagine a paper fitting this description also to fit under the high range of "development": "A thesis guides this paper; all points clearly relate to it. Each main point proving the thesis is supported concretely, with sound arguments or convincing examples." Such descriptions insure the correlation of the scores for voice and development. However, descriptions of other categories (e.g., organization and sentence structure) appear to allow for discrete ratings. In such cases, the correlation that occurred might have been lessened by training the readers more intensely in the application of the scale. It is still possible that the relationships between these categories stem from the way writers write; a writer who performs well or poorly on one category will do similarly on most of the others and on a holistic scale. In any event, the reasons for these high correlations deserve further investigation.

In spite of the significant effects of trainer and topic on essay scores and in spite of the high correlation between the holistic and analytic rating scale, the essays themselves still contributed most to the different scores ($p<.001$). In another study, I examined the relative effects on readers' scores of four qualities within the essays: development, organization, sentence structure and mechanics. I found the larger discourse categories of development and organization to be most influential (Freedman, 1979 a, b). Finally, the readers themselves and two parts of the rating environment, the session of the rating and the rating day did not affect the scores.

In conclusion, the influences on readers' scores are numerous. Although the essay itself contributes most to the score, other influences beyond the text, such as the trainer and the topic, also have their effects.

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.

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

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.

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

Freedman, S. How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, 1979, *71*, 328–338. (a)

Freedman, S. Why teachers give the grades they do. *College Composition and Communication*, 1979, *30*, 161–164. (b)

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.

Markham, L. Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 1976, *13*, 277–283.

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

Nold, E., & Freedman, S. An analysis of readers' responses to essays. *Research in the Teaching of English*, 1977, *11*, 164–174.

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

Slotnick, H., & Knapp, J. Essay grading by computer: A laboratory pheomenon? *Educational Measurement*, 1971, *9*, 253–263.

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).