To determine what influences a teacher's rating of student writing, one must determine from the aspects of that writing what seems to affect the rating. One runs the risk that correlation is not causation; a careful multivariate design does not minimize that risk as this article shows but does help warn the unwary. Reviewed by R. M. W.

An Analysis of Readers' Responses to Essays

ELLEN W. NOLD
Stanford University

SARAH W. FREEDMAN
Stanford University

Judging student writing is a difficult, time consuming task for both the classroom teacher and the researcher. Criteria of excellence for compositions remain ill-defined. This study is designed to define some easily measurable syntactic and semantic cues within student essays that predict and perhaps determine readers' responses, thereby facilitating the evaluators' task. In searching for the covert, quantifiable written cues that correlate with reader response, we rely on frequency counts to, as Braddock advises in *Research in Written Composition*, "*discover certain key situations which are indices of larger areas of concern*" (21). Our larger area of concern consists of defining the basis for, or at least some predictors of, reader response.

PREVIOUS RESEARCH

Past researchers in composition evaluation have discovered several factors, such as handwriting and spelling, tend to influence readers. The most significant study on reader response, Diederich, French, and Carlton's *Factors in Judgement of Writing Ability* (1961), focused on the major differences between readers. Diederich found that readers' written comments on papers cluster into five groups, each group responding to something basically different in an essay. While some of his readers' comments were most concerned with ideas, others stressed mechanics, organization, wording

The research reported here was sponsored by the Office of the Dean of Undergraduate Studies. The authors wish to thank Dean James L. Gibbs for his cheerful encouragement, Frederick C. Nold for his individual midnight statistical guidance, May L. Dageforde for her research assistant, and Heather Holmes for administering the tests

or flavor. In a subsequent analysis, Diederich found that elements in these clusters explained 43% of the variance in grades. In effect, Diederich, by asking readers to explain their response in the form of written comments, could measure only gross factors in readers' responses. In contrast, instead of asking readers to tell us about their responses, we examined the papers that the readers judged in order to isolate predictors of their response and the subtle features to which readers respond when judging compositions. Such a technique complements Diederich's survey of readers' comments by attempting to account for the influences that lie within the writing they judge but are too subtle for mention in their comments.

THE DESIGN

For this experiment, 22 Stanford freshmen, whose English Composition SAT scores were typical of all freshmen required to take writing at Stanford, wrote four in-class essays in two one hour sessions separated by a span of several months. Our total sample consists of 88 essays written under test conditions. During each session the administrator first gave out one question, collected the responses ½ hour later then gave the second question, collecting those in ½ hour. The questions were always distributed in the same order, and the same directions were read to all students. The two sets of two questions each, parallel in design and difficulty, follow:

Test Questions

Test A

1) President Ford gave Nixon an "unconditional pardon." Do you agree or disagree with Ford's decision? Give reasons for taking your position. Think through your reasons carefully before you write.

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

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

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

Test B

1) 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. Think through your reasons before you write.

2) "If a society is to strive with any hope of success toward peace and prosperity in a commonwealth, the authority governing that society must not only be able to pass laws and to reassess those laws constantly as circumstances change . . ., it must also be enabled to enforce those laws and to exact penalties for their violation."

"Under a government that imprisons unjustly, the true place for a just man is also in prison."

Assignment: Write an essay on the two passages above in which you answer the

do they differ? What strong or weak points does each position have? To what extent might a person accept both positions?

Our four different topics, all in the argumentative mode of discourse, insured that our findings would not be topic specific and could be generalized across more than one topic in the argumentative mode of discourse. We decided to choose topics in only one mode of discourse because Braddock warns, "variations in mode of discourse may have more effect than variations in topic on the quality of writing" (8). In addition, he cites past research indicating that mode of discourse affects sentence structure. Since the influence of sentence structure was one of the factors we were examining, we had a second reason to control the mode of discourse.

Reading We were careful to insure that the evaluation of the essays be as reliable as possible. We hired six experienced teachers, each to read all 88 essays. All had had at least one year's experience teaching Stanford freshmen. *Readers read and rated the 22 responses to each of the four questions separately.* The readers read on two consecutive Saturdays in adjoining rooms. On the first day of reading, three readers evaluated question 1, Test A in the morning and question of Test B in the afternoon while the other three readers evaluated question 2 of Test A in the morning and question 2 of Test B in the afternoon. On the second Saturday, the groups of readers switched rooms and tasks, evaluating the questions that they had not evaluated before in the same order as on the previous Saturday. On both reading days one investigator supervised and trained for question 1's while the other investigator took the responsibility for question 2's. In this way, readers always read each set of essays in the same order, in the same room, and under the supervision of the same investigator.

Every essay was typed both to avoid bias that might be caused by handwriting and to make the rating task less fatiguing. We coded each essay to conceal the identity of the writer. Braddock warns, "the anonymity of the writer should be preserved to prevent the *personal feelings* of the rater from coloring his evaluation" (10). Raters also received regular rest periods to avoid the erratic evaluations that Braddock warns may result from fatigue.

Readers were instructed to read quickly and to rate each paper on a four point scale, one being the best score and four the worst. We chose this holistic method of rating over an analytic method because we did not want to bias the reading by providing categories that might correlate with our later frequency counts. We were careful during the traning not to discuss any of the independent variables that we would later measure as possible predictors of reader response, but questions about incomplete essays did arise. Our major purposes in training were to insure that the raters understand the topics and develop similar expectations for responses on the topics. We encouraged them to agree with one another, to develop a common set of criteria on which to judge a given question.

---

of Test A and B were developed by the California State Universities and ... Freshman English.

Before the reading session on each question, every reader received a training packet, which contained a copy of the essay question, a sample rating form, and eight training essays arranged in a random order. Students who were not part of the experimental group wrote these training essays at the same time and under the same conditions as those written by the experimental group. We chose training essays to represent the range of quality within the experimental group. Each training session consisted of a brief discussion of what an adequate response to the question should contain. Next we explained the "general impression" method of reading. After the readers rated the 8 training essays, the discussion focused on essays about which the raters disagreed, the investigator attempting to guide raters to reconcile their differences.

After the training for a particular question, the raters read and rated the 22 experimental essays in response to that question. Every reader received each set of 22 essays in the same random order. An identical sequence of training then reading followed for each of the four questions.

DEPENDENT VARIABLE To determine a quality rating for each essay, we summed the scores given by the six readers. Hence, our dependent variable, quality, could take a value from 6 (for the best essay) to 24 (for the worst). Because our dependent variable is a summed score, interval reliability need not be assured. Reading reliability determined by Diederich's technique in *Measuring Growth in English* is over .85.

INDEPENDENT VARIABLES Selecting the independent variables to predict quality was much more complex and less certain because research in composition lacks a widely accepted theoretical basis. We made *ad hoc* decisions about the features of the compositions we wished to study, decisions informed in part by other empirical studies such as Hunt's (1965), Christensen's (1967) and Diederich's (1961) and in part by our own judgements. Our composite, working hypothesis, then, was that readers responded to four general features of the essays: (1) the number, development and logic of ideas (semantic content), (2) the existence and appropriateness of the organization of those ideas, (3) the complexity, variation and appropriateness of the syntax and (4) the richness and appropriateness of the vocabulary. We omitted variables concerned with mechanics, spelling and handwriting: the latter variable controlled by typing the essays. Stanford students compared to other college students make very few mechanical and spelling errors; the few committed, we judged, would be largely ignored by our readers, who would attribute them to haste.

Because of the difficulties inherent in defining "objective" ways to determine the level of some of the variables (such as appropriateness or logic) specified by our working hypothesis, we preferred to focus upon more easily counted and agreed-upon features of the essays, mainly syntactical. As a result, we most probably slighted some very important determinants of quality: for this ... findings must be viewed ...

To begin, we adopted the variables which Lester Golub used to compute his yntactic Density Score (SDS), reported in *Measures for Research and Evaluation n the English Language Arts* (NCTE, 1975). Golub, influenced by Hunt, se- ected his variables in an effort to define a measure of reading difficulty as well as riting maturity. To compute the SDS, one sums positively weighted occurrences f the following ten variables and divides the sum by the number of T-units Hunt's minimally terminal unit) in the sample. The variables are:

(1) words per T-unit
(2) subordinate clauses per T-unit
(3) mean main clause length
(4) mean subordinate clause length
(5) number of prepositional phrases
(6) number of possessive nouns and pronouns
(7) number of adverbs of time
(8) number of modals
(9) number of be and have forms in the auxiliary
(10) number of gerunds, participles and absolute phrases

Golub suggests limiting the sample by counting only to the end of the T-unit after the two hundredth word. We followed this convention for the 56 essays over two hundred words, but not wanting to exclude or handicap the 24 essays under two hundred words, we counted all their syntactical features, too. We prorated the occurrence of prepositions, possessives and unbound modifiers over the number of words in the syntactical sample for essays both over and under the two hundred word limit.

We redefined the ten original syntactical variables, expanding the last two into six separate variables, ending up with these fourteen:

(1) words per T-unit
(2) subordinate clause per T-unit
(3) mean main clause length
(4) mean subordinate clause length
(5) % prepositions in syntactical sample
(6) % possessive nouns and pronouns in syntactical sample
(7) % adverbs of time in syntactical sample
(8) modals per finite verb in syntactical sample
(9) *bes* and *haves* in auxiliaries per finite verb in syntactical sample
(10) passives per finite verb in syntactical samples
(11) progressives per finite verb in syntactical sample
(12) % gerunds, participles and absolutes in syntactical sample
(13) % words in final free modifiers in syntactical samples
(14) % words in medial free modifiers in syntactical sample

To these fourteen we added three others:

(15) common verbs per finite verb in syntactical sample
(16) dummy variable for long essays

We added passives and progressives to Golub's original (9) in order to determine any differential effects of the two separate uses of *to be* in the auxiliary. Following Dixon's Indexes of Syntactic Maturity, developed in 1970 using data from research done by Hunt (1965) and Christensen (1967) and reported in *Measures for Research and Evaluation in the Language Arts* (1975), we counted words in free final and free medial modifiers. (We also counted words in initial modifiers, but they were found to be, on face value, unrelated to quality ratings; thus, we deleted them from consideration in our list of possible predictors.) We felt it necessary to distinguish these more complex types of modifiers because in our original counting we noted a heavy occurrence of gerunds and participles in all papers, not respective of quality.

To our list of syntactic independent variables, we added three others, designed in part to measure vocabulary and ideas. The vocabulary component was defined as the percentage of finite verbs that were "common verbs."[2] The common verbs were chosen arbitrarily by the investigators, taking into account their experience writing curriculum material for the lower grades (3 through 8) and their experience with the vocabularies of Stanford freshmen. Obviously, much more research on vocabulary must be done before this list could be based on anything but intuition. The choice of common *verbs* rather than, say, nouns, adjectives or adverbs was also pragmatic and arbitrary. Since every T-unit contains at least one finite verb and since we had already determined the number of finite verbs by counting the number of T-units and subordinate clauses, it was an easy step to look at each finite verb and determine whether it was common or not.

We used length of the essay as an approximate indicator of the number and development of the ideas expressed therein. Rather than rank the essays by length or use the length of the essay as a raw variable, we determined the mean and standard deviation of the lengths of the essays written on each topic. We then created a variable, LONG, and assigned the number *1* to an essay if it was one standard deviation above the mean and *0* if it was not. We also created a variable, SHORT, and assigned the number *1* to an essay if it was one standard deviation below the mean and *0* if it was not. These kinds of variables are called dummy variables because they are categorical abstractions from the data. They are useful when an investigator hypothesizes that variations in a particular variable are correlated with changes in another variable but does not wish to posit a particular functional form.

RESULTS AND INTERPRETATIONS

Before giving the correlational results of the research, we wish to familiarize the reader with the descriptive results. The kind of essay affected the amount of writing produced, as Table 1 shows.

---

[2] The following were counted as common verbs: be, break, do, feel, follow, find, get, give, go, have, hold, keep, know, look, live, make, mean, need, put, say, see, seem, stop, take, tell,

TABLE 1
Amount Written on Different Topics

	Mean	Standard Deviation
Test A, topic 1 (Ford)	262	113
Test A, topic 2 (Compare personal philosophies)	235	92
Test B, topic 1 (Abortion)	264	83
Test B, topic 2 (Compare government philosophies)	236	63

The matched personal opinion topics (Topics 1), the first topics of the day, are characterized by more writing and a greater variation in amount of response. In the Topic 2 essays, in which they were to compare quotations, the students wrote less, and there was less variation. A t test on the differences between the means of the two topics indicated that they were not significant.

Table 2 gives the mean and standard deviations for the eighteen variables in the study (one dependent variable, quality, to be predicted and seventeen independent variables to predict).

TABLE 2
Means and Standard Deviations for the Variables

	Mean	S.D.
Quality (Qual)	16.98	3.48
Words per T-unit (Wdprt)	17.27	4.10
Subordinate clauses per T-unit (Subprt)	.77	.45
Mean main clause length (Main)	10.68	5.16
Mean subordinate clause length (Sub)	8.40	2.19
% prepositions in sample (Prep)	.10	.03
% possessives in sample (Poss)	.02	.01
% adverbs in sample (Adv)	.23	.12
% finite verbs which have modal auxiliaries (Modal)	.19	.13
% finite verbs which show *be* or *have* as auxiliaries (Be-have)	.10	.08
% finite verbs in the passive voice (Pass)	.03	.08
% finite verbs in progressive mood (Prog)	.03	.02
% gerunds, participles, absolutes (Ger)	.03	.02
% words in final free modifiers (Final)	.04	.05
% words in medial free modifiers (Medial)	.02	.03
% finite verbs which are common (Common)	.54	.17
% essays assigned LONG (Long)	.16	
% essays assigned SHORT (Short)	.19	

The mean quality rating is near 17, two points lower than the arithmetic mean (15) of the scores between 6 and 24. In fact, the most prevalent score was 3:

TABLE 3
Simple Correlations for Variables, N = 88 ($p < .05$ when $r \pm .214$)

	qual	wd-prt	sub-prt	main	sub	prep	poss	adv	modal	be-have	pass	prog	ger	final	me-dial	com-mon	long
wdprt	−.08																
subprt	−.03	.51															
main	−.09	.36	−.41														
sub	−.06	.33	.16	−.13													
prep	−.12	.19	−.33	.50	−.12												
poss	−.11	.10	.01	.09	.16	.05											
adv	−.02	.03	.19	−.11	.07	−.11	.12										
modal	.38	−.08	−.16	.02	.15	−.13	.14	−.13									
be-have	.32	.06	−.03	.19	−.11	.07	.08	.05	.19								
pass	.07	−.10	−.06	−.07	−.11	.00	−.07	−.02	.00	.53							
prog	.24	−.12	.07	.02	−.05	−.13	.12	.15	.19	.41	.01						
ger	−.18	.02	−.21	.32	−.05	.30	−.08	−.02	−.11	.02	.07	−.01					
final	−.42	.14	−.05	.25	−.02	.18	.16	.02	−.24	−.14	−.15	−.09	.23				
medial	−.03	−.06	.08	−.08	−.12	−.02	.04	−.12	.03	−.16	−.02	−.09	.04	−.03			
common	.18	−.22	.08	−.25	.02	−.25	−.08	.05	−.10	−.18	−.20	−.08	−.28	−.17	−.05		
long	−.25	−.04	−.16	.08	−.06	.21	.13	.10	−.19	.08	.07	.13	−.06	.20	.01	.01	
short	.44	−.14	−.23	.11	.00	.16	−.08	−.19	.26	−.21	−.06	−.10	.03	−.22	.16	−.01	−.21

55 of the 88 essays were assigned a 3 by three or more of the six readers. Only one essay received a 1 from three or more readers. The distribution of scores in this sample, then, is skewed heavily towards scores below the arithmetic mean. The readers' informal comments about the papers to the investigators indicated that they expected better papers from Stanford students. Still, Stanford students are an unusual group, as shown by the mean number of words per T-unit, a remarkable 17.27. Dixon (1970) reports 12.25 words per T at grade 12 and 13.33 at grade 16. Hunt (1965) reports 14.4 words per T at grade 12 and 20.3 for superior adults. The average words per T was above 20.3 in 19 of the 88 essays (22%).

Table 3 gives the simple correlations among the eighteen variables. *Since we used a low score (6) when we meant high quality, the signs in the quality column are reversed.* That is, the correlation between final modifiers and quality is −.42, but this indicates that final modifiers are *positively* associated with quality. The signs in all other rows and columns express the usual, expected relationship.

The results of a stepwise multiple regression are presented in Table 4. The stepwise approach selects the most powerful explanatory variable first, then selects the second most powerful, and then the third—until the researcher determines from the F statistic that the variables being added are statistically insignificant. Of our seventeen independent variables, five are statistically significant (Table 4).

TABLE 4

Stepwise Regression on 17 Variables as Possible Predictors of Quality, 82 degrees of freedom

Step Number	Variable	B	Standard Error	F	R^2
1	Short	2.53	0.78	10.45	.196
2	Final	−16.27	5.98	7.41	.312
3	Modal	6.22	2.58	5.82	.353
4	Be-have	5.72	2.31	6.15	.384
5	Common	4.19	1.77	5.58	.423

The multiple regression reveals that, in combination, our five variables explain 42 per cent of the variance in quality scores. The dummy variable, Short, explains 20 percent; the percentage of words in final modifiers explains another 12 per cent; the percentage of finite verbs with modals explains another 4 per cent; the percentage of verbs with be or have as auxiliary explains yet another 3 per cent; the percentage of common verbs explains the final 4 per cent. The F statistics for the entire equation is 12.02 ($P < .001$). Only one variable, Final, is associated positively with quality; the other four variables—Short, Modal, Be-have and Common—are negatively associated.

CONCLUSIONS

Our study lends support to Christensen's hypothesis that sophistication in modification, especially free final modification producing a cumulative sentence, is indicative of good writing in expository and argumentative discourse—as well as the narrative and descriptive discourse on which Christensen based most of his work. Free final modifiers appear in only 41 of our 88 essays; when they appear, there was an average of 1.5 incidences in an essay. Each modifier averages 10 words. Fewer instances of free final modification appear in our sample than Dixon reports in *Measures for Research and Evaluation in the English Language Arts,* (2.45 instances per essay in grade 12 and 2.0 in grade 16), but the modifiers in our sample are longer on the average than Dixon's 7 words per final modifier in grade 12, and 6.7 in grade 16. Mode of discourse may affect the number of instances of free final modification, the argumentative mode eliciting fewer.

That length is a predictor of quality corroborates the findings of Diederich (1961), Page (1966) and Bracht and Hopkins (1968). The simple correlation between length and quality in our sample was −.57, indicating a high positive correlation. Our study shows that it is more damning to write a short essay than elevating to write a long one. Besides lack of ideas, short essays have two faults: lack of sophistication in modification and a greater percentage of sentences in which the finite verbs are tempered by modals (should, could, would); when sentences containing modals are not properly supported by others containing examples and justifications of the judgements and hypotheses expressed by the modals, the essay is weakened.

Another way to weaken an essay is to use a high percentage of *bes* and *haves* as auxiliaries. While progressives are significantly more indicative of low quality than passives, summing all forms of *be* and *have* in the auxiliary has by far the most predictive power. One reason for this finding is that *have* in the auxiliary position in our argumentative essays is often employed in the fallacy of hypothesis contrary to fact ("If he *hadn't* done it then, . . ." or "He *could have* thought of a better way") which weaken the credibility of an argument.

The vocabulary measure comes in significantly, but last, as a predictor of quality. It is associated negatively with long sentences ($r = -.22$) and long main clauses ($r = -.25$), indicating that students with weak vocabularies are likely also to be writing less complex T-units than their peers. Sophistication in syntax and diction appear to develop simultaneously.

The strength of the predictive quality of our regression equation may be judged by looking at the residuals, the difference between the composite score given by the readers and the composite score computed by the equation. We count as "mispredicted" those computed scores whose residuals differ by more than 3.5 points from their actual composite ratings. For example, if the readers' composite score for a particular essay were 14 and the computed composite score were 15.5, the residual is −1.5, and we would count the score correctly predicted. But if the computed composite score were 16.7, making the residual −3.7, then we would count the score as mispredicted. Using this rule of thumb, we find that our regression equation correctly predicts the scores of 76 of our 88 essays (86%).

SUGGESTIONS FOR FUTURE RESEARCH

Though we have found some significant predictors of quality, we do not claim to have identified all the important variables which will distinguish between higher quality and lower quality essays. We do know that words per T and other standard developmental measures are not useful in predicting perceptions of quality on the college level. Preliminary results of a validation study associated with this research indicate that the significance of variables may be sensitive to the expectations of readers about the level of ability of the writers and/or the range in number and quality of the essays read. Care must be taken that both readers and tasks remain as consistent as possible across studies to ensure the reliability and validity of readings, even when the researcher sums the scores of as many as six readers.

Careful design of studies and creative definition of variables will most probably lead the inquisitive researcher onto other important variables. What we may be missing in part are measures of the sophistication of inter T-unit and inter-paragraph coordination, the conciseness of the thesis paragraph, and the recognition of points of view other than those presented in the thesis. To quote e.e. cummings, our intuition tells us that "since feeling is first/who pays any attention/to the syntax of things" explains at most one third of the variance of the scores. For essays in the expository and argumentative modes, the other two thirds or more of the variance most probably is explained by the depth and elaboration of the writer's arguments and the writer's stance towards the reader, variables difficult to measure by simple frequency counts. Though the discovery of what affects the judgments of readers is an incredibly complex business, we hope that it will have rewards for the researcher, the test maker, and the composition teacher.

REFERENCES

Bracht, G. H., and K. D. Hopkins. "Objective and Essay Tests: Do They Measure Different Abilities?" Paper presented at the annual meeting of the American Education Research Association, Chicago, February 2, 1968.

Braddock, Richard, Richard Lloyd-Jones, and Lowell Schoer. *Research in Written Composition.* Champaign, Ill: NCTE, 1963.

Christensen, Francis. *Notes Towards a New Rhetoric: Six Essays for Teachers.* New York: Harper and Row, 1967.

Diederich, P. E., S. W. French, and S. T. Carlton. *Factors in Judgements of Writing Ability* Research Bulletin RB-61-15, Princeton: Educational Testing Service, 1961, ED002 172.

Dixon, E. *Indexes of Syntactic Maturity.* 1970, ED 091 748.

Fagan, William T., Charles R. Cooper, and Julie M. Jensen. *Measures for Research and Evaluation in the English Language Arts.* Urbana, Ill.: NCTE, 1975.

Hunt, Kellogg W. *Grammatical Structures Written at Three Grade Levels:* Research Report No. 3. Urbana, Ill.: NCTE, 1965.

Kinneavy, James L. *A Theory of Discourse.* Englewood Cliffs, N.J.: Prentice Hall, 1971.

Page, E. "Grading Essays by Computer," in *Proceedings of the 1966 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1966.

Grading papers gives one sense of a teacher's criteria. Marking may give another. This study seeks to disentangle the criteria used in marking, grading, ranking, and commenting upon student papers. Reviewed by C. R. C.

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

WINIFRED HALL HARRIS
Troy State University

Of the three segments of the English curriculum, language, literature, and composition, the stepchild seems to be composition. Few English teachers are likely to prefer teaching composition to literature, and composition seems to be most often neglected (Squire and Applebee, 1968). Of the thirty-six English teachers who participated in the study reported here, only four preferred to teach composition. Since both the teaching and the evaluation of writing are so often frustrating experiences and the results of hours and even years of instruction so often unrewarding when the end product is considered, it is not difficult to sympathize with English teachers' preference for teaching literature instead of composition. At the same time, English teachers have complained of the general lack of research in the area of composition, such insufficiency making their task even more difficult and frustrating because of their need for specific evidence that might corroborate their practices, provide new insights, and give them direction for new or different approaches to the teaching and evaluation of writing.

Attempts to measure the effectiveness of instruction in composition or the quality of the writing produced thereby are more often discouraging than rewarding because of the subjective nature of the task, the many variables