DOCUMENT RESUME

ED 031 479 TE 001 443

By-Diederich, Paul B.

Grading and Measuring.

Pub Date 65

Note-13p.; In "Improving English Composition," ed. Arno Jewett and Charles E. Bish (Washington, D.C.: National Education Assn., 1965) pp. 81-91.

Available from-National Education Association, 1201 Sixteenth Street, N.W., Washington, D.C. 20036 (Cloth, $2.50, Stock No. 781-10508; Paper, $1.50, Stock No. 781-10510)

EDRS Price MF-$0.25 HC Not Available from EDRS.

Descriptors-Achievement Rating, *Composition (Literary), Composition Skills (Literary), English Education, *English Instruction, *Evaluation Criteria, Evaluation Methods, Grades (Scholastic), *Grading, Student Evaluation, Student Improvement, Testing, Writing Exercises, *Writing Skills

The low rate of agreement among readers of College Entrance Examination essays suggested the need to examine the qualities in student writing which caused wide variance in grading. To study this question, 300 homework papers by freshmen at three universities were graded by 60 distinguished readers in six fields. The following factors, by rank, seemed to influence readers: ideas expressed, grammar, punctuation, spelling, handwriting, organization, analysis, wording, phrasing, and "flavor." These factors reduced to "general merit" and "mechanics," in addition to three possible ratings of "high," "medium," or "low," were used to grade monthly test papers of English pupils in 17 high schools for 1 year. From this trial period, a means of measuring student growth in writing ability was developed. All students in a span of three grades would simultaneously write on the same topic several times a year. The unidentified papers would be graded, and the students' scores compared over a 3-year period, would indicate ther progress. (JM)

Improving English Composition

Edited by
Arno Jewett, Director of the Project, 1962-63
Charles E. Bish, Director of the Project, 1963-

NEA-Dean Langmuir Project on
Improving English Composition

Library of Congress Catalog Card Number 65-16639

Single copy: cloth, $2.50 (Stock No. 781-10508); paper, $1.50 (Stock No. 781-10510). Discounts on quantity orders: 2-9 copies, 10 percent; 10 or more copies, 20 percent. All orders not accompanied by payment will be billed with shipping and handling charges added. Orders amounting to $2 or less must be accompanied by payment. Order from and make checks payable to National Education Association, 1201 Sixteenth Street, N.W., Washington, D.C. 20036.

11

Grading and Measuring

Paul B. Diederich
Director of Research in English
*Educational Testing Service**

COLLEGE BOARD EXPERIENCE

The College Entrance Examination Board used nothing but essay examinations from 1900 to 1926, then used a mixture of essays and objective tests, and since 1941 has used chiefly objective tests. Although the latter yielded better predictions of academic success, and although their wide sampling of content gave teachers greater freedom, there was continual pressure to return to the essay in at least one examination. Several costly

* Paul B. Diederich took his B.A. and M.A. degrees at Harvard, and his Ph.D. at Columbia. He taught in private and public high schools from 1930 to 1940 and was associate professor of English and examiner in English for the United States Armed Forces Institute at the University of Chicago from 1940 to 1950. Dr. Diederich has been a member of the Research Division of Educational Testing Service in Princeton, New Jersey, since that time and is known chiefly for the experimental tryout of the use of readers to assist high school English teachers in grading and correcting compositions. With Osmond E. Palmer he is author of a book of instructional tests for college freshmen, *Critical Thinking in Reading and Writing* (Holt, 1955). His article "The Rutgers Plan for Cutting Class Size in Two" appeared in *The English Journal* in April 1960.

experiments were conducted using essays up to two hours in length, each graded by two or more College Board readers. But these readers did not agree very closely on the merit of the papers, and the students were even more erratic. The quality of their writing varied a great deal from one occasion or topic to another. As a result, final grades on two long essays agreed only 0.45 with one another, whereas scores on two objective tests of verbal ability, taken at the time of writing the essays, agreed 0.88.

It became obvious that further progress could be made only by finding out what qualities in student writing affect readers differently, causing a difference in their grading. It seemed unlikely that capable readers would disagree so wildly unless they were looking at different things or weighting them differently.

MATERIALS FOR A STUDY OF READER REACTIONS

To study this question, the writer and two colleagues [1] in the Research Division of Educational Testing Service secured 600 papers written as homework between one class meeting and the next by freshmen at Cornell, Middlebury, and the University of Pennsylvania. There were four topics, but only two were chosen by enough students: "Who Should Go to College?" and "When Should Teenagers Be Treated as Adults?" They were told that their papers would be read by 60 distinguished readers in six different fields: college English teachers, social science teachers, natural science teachers, writers and editors, lawyers, and business executives. The students were more stimulated than frightened by such an audience because they knew that their papers would be typed and reproduced without identification and that grades would not be reported to anyone.

We reduced the 600 papers to 300 (150 on each topic) without reading them: first, by dropping papers on the two less popular topics; second, by looking at the Scholastic Aptitude Test verbal scores of the writers. Since we wanted as wide a range as possible, we kept all papers written by students with either high or low SAT verbal scores and reduced the number

[1] Diederich, Paul B.; French, John W.; and Carlton, Sydell T. *Factors in Judgments of Writing Ability.* Research Bulletin 61-15. Princeton, N. J.: Educational Testing Service, 1961. (Out of print.)

with middle scores in such fashion that the distribution of verbal ability on one topic was parallel to that on the other. The remaining papers on both topics represented a wider range in verbal ability than any one teacher would be likely to encounter in a selective college. It may be said at once that we found no significant difference of any kind between one topic and the other. Hence our conclusions can be generalized at least to the types of short expository papers that are commonly assigned in both high schools and colleges.

HOW THE PAPERS WERE GRADED

The readers were told to sort the papers into nine piles in order of general merit. No instructions were given as to what to look for, since we wanted to find out what the readers looked for when they were free to grade as they liked. The only rules were that all nine piles must be used, and not less than six papers on each topic must appear in the smallest piles. The readers were also asked to comment on anything they liked or disliked in as many papers as possible.

The result was nearly chaos. Of the 300 papers, 101 received all nine grades, 111 received eight, 70 received seven, and no paper received less than five. The average agreement (correlation) among all readers was 0.31; among the college English teachers, 0.41. Readers in the other five fields agreed with the English teachers slightly better than they agreed with other readers in their own field.

This procedure has been criticized on the ground that we could have secured a higher level of agreement had we defined each topic more precisely, used only English teachers as readers, and spent some time in coming to agreements upon common standards. So we could, but then we would have found only the qualities we agreed to look for—possibly with a few surprises. We wanted each reader to go his own way so that differences in grading standards would come to light. We used readers in five fields in addition to English teachers because our colleagues also have opinions on the writing ability of our students, and so do representatives of the educated public.

THE FACTOR ANALYSIS

We correlated the grades of each reader with the grades of every other reader and put this large table of agreements and

disagreements through the mathematical procedure known as "factor analysis." This is too complicated to explain briefly, but the effect is as though the computer scanned all the correlations and picked out clusters of readers who agreed with one another and disagreed with other clusters to a greater degree than could come about by chance. There proved to be only five such clusters. They were clearly agreeing on something, and on something different in each cluster. What was it?

We found out by tabulating the comments of the three readers who stood highest on each factor (who came closest to the central tendency of each cluster) and only on papers graded either high (7-8-9) or low (1-2-3). We checked our conclusions by similarly tabulating the comments of the three readers who stood lowest on each factor. Comments were tabulated under 55 headings by a person who did not know the standing of any reader on any factor. In all, 11,018 comments on 3,557 papers were tabulated. They were reduced to percentages of total comments written by each reader so that readers who wrote the most comments would not unduly influence the interpretation.

It then became quite clear that the largest cluster (16 readers) was influenced primarily by the *ideas* expressed: their richness, soundness, clarity, development, and relevance. The next largest (13 readers) was most influenced by *mechanics*: the number of errors in grammar or usage, punctuation, and spelling. Seven of the ten English teachers stood high on this factor. The third (9 readers) showed the highest interest in *organization* and analysis. Four of the business executives stood high on this factor. (They were also especially sensitive to poor spelling but not to other elements of mechanics.) The fourth (9 readers) stood highest in specific comments on *wording* and phrasing: on verbal felicity or infelicity. The fifth (7 readers) emphasized style, individuality, interest, sincerity, the personal qualities of the writing, which we decided to call *flavor*. The four readers who stood highest on this factor were all writers or editors. They also had the lowest percentage of specific comments on mechanical errors.

Here, evidently, were some of the reasons why expert College Board readers had so long failed to agree. Like the distinguished readers assembled for this study, they were responding to different qualities in the papers, or they differed in the weights they attached to these qualities. One possible conclusion might be that papers in important tests of writing ability should be

rated by five different readers, each of whom was especially sensitive to one of these factors. Since this was hardly feasible, it was comforting to find no solid evidence that any reader was entirely blind to any of these qualities. There were only differences in emphasis, heightened by the absence of directives and amplified by the technique of factor analysis. If readers were asked for a rating on each factor or on some of its principal components, it seemed likely that all but a few readers would be able to follow these instructions.

This policy was tried out in three large high schools the following year. The principal new finding was that, under the pressure of time and the teaching tradition, these five factors collapsed into two: a general merit factor and a distinct mechanics factor. The ratings that had the highest "loadings" on the general merit factor were, however, four of our five original factors: ideas, organization, flavor, and wording. While we might have settled for a single rating on merit and another on mechanics, we decided to ask for a separate rating on the four main components of each in order to make the totals more reliable. Since we were now dealing with handwritten papers, the mechanics factor was broadened to include a rating on handwriting and neatness as well as on grammar and sentence structure, punctuation, and spelling.

DEFINITION OF POINTS ON THE RATING SCALE

During the past year, English departments in 17 high schools have rated monthly test papers written in class for these eight qualities, each on a scale of 1 (low) to 5 (high). For the benefit of students, high, middle, and low points on each quality were defined in very simple terms, as follows:

General Merit

1. Ideas

High. The student has given some thought to the topic and has written what he really thinks. He discusses each main point long enough to show clearly what he means. He supports each main point with arguments, examples, or details; he gives the reader some reason for believing it. His points are clearly related to the topic and to the main idea or impression he is trying to get across. No necessary points are overlooked and there is no padding.

Middle. The paper gives the impression that the student does not really believe what he is writing or does not fully realize what it means. He tries to guess what the teacher wants and writes what he thinks will get by. He does not explain his points very clearly or make them come alive to the reader. He writes what he thinks will sound good, not what he believes or knows.

Low. It is either hard to tell what points the student is trying to make or else they are so silly that he would have realized that they made no sense if he had only stopped to think. He is only trying to get something down on paper. He does not explain his points; he only writes them and then goes on to something else, or he repeats them in slightly different words. He does not bother to check his facts, and much of what he writes is obviously untrue. No one believes this sort of writing—not even the student who wrote it.

2. Organization

High. The paper starts at a good point, moves in a straight line, gets somewhere, and stops at a good point. The paper has a plan that the reader can follow; he is never in doubt as to where he is or where he is going. Sometimes there is a little twist near the end that makes the paper come out in a way that the reader does not expect, but it seems quite logical. Main points are treated at greatest length or with greatest emphasis; others, in proportion to their importance.

Middle. The organization of this paper is standardized and conventional. There is usually a one-paragraph introduction, then three main points each treated in one paragraph, and then a conclusion, which often seems tacked on or forced. Some trivial points may be treated in greater detail than important points, and there is usually some dead wood that might better be cut out.

Low. This paper starts anywhere and never gets anywhere. The main points are not clearly separated from one another, and they come in a random order—as though the student had not given any thought to what he intended to say before he sat down to write. The paper seems to start in one direction, then another, then another, until the reader is lost.

3. Flavor

High. The writing sounds like a person, not a committee. The writer seems quite sincere and candid, and he writes about something he knows—often from personal experience. You could never mistake this writing for the writing of anyone else. Although the writer may play different roles in different papers, he does not put on airs. He is brave enough to reveal himself just as he is.

Middle. The writer usually tries to appear better or wiser than he really is. He tends to write lofty sentiments and broad generalities. He does not put in the little, homely details that show that he knows what he is talking about. His writing tries to

sound impressive. Sometimes it is impersonal and correct but colorless, without personal feeling or imagination.

Low. The writer reveals himself well enough but without meaning to. His thoughts and feelings are those of an uneducated person who does not realize how bad they sound. His way of expressing himself differs from standard English, but it is not his personal style; it is the way uneducated people talk in his neighborhood.

4. Wording

High. The writer uses a sprinkling of uncommon words or of familiar words in an uncommon setting. He shows an interest in words and in putting them together in slightly unusual ways. Some of his experiments with words may not quite come off, but this is such a promising trait in a young writer that a few mistakes may be forgiven. For the most part he uses words correctly, but he also uses them with imagination.

Middle. The writer is addicted to tired old phrases and hackneyed expressions. If you left a blank in one of his sentences, almost anyone could guess what word he would use at that point. He does not stop to think how to say something; he just says it in the same way as everyone else. A writer may also get a middle rating on this quality if he overdoes his experiments with uncommon words: if he always uses a big word when a little word would serve his purpose better.

Low. The writer uses words so carelessly or inexactly that he gets far too many wrong. These are not intentional experiments with words in which failure may be forgiven; they represent groping for words and using them without regard to their fitness. A paper written entirely in a childish vocabulary may also get a low rating, even if no word is clearly wrong.

Mechanics

5. Grammar, Sentence Structure

High. There are no vulgar or "illiterate" errors in grammar or usage by present standards of informal written English, and there are very few errors in points that have been emphasized in class. The sentence structure is usually correct, even in varied and complicated sentence patterns.

Middle. There are a few serious errors in grammar and several in points that have been emphasized in class, but not enough to obscure meaning. The sentence structure is usually correct in the more familiar sentence patterns, but there are occasional errors in more complicated patterns such as parallelism, subordination, consistency of tenses, reference of pronouns, etc.

Low. There are so many serious errors in grammar and sentence structure that the paper is hard to understand.

6. Punctuation

High. There are no serious violations of rules that have been taught—except slips of the pen. Note, however, that modern editors do not require commas after short introductory phrases, around nonrestrictive clauses, or between short coordinate clauses unless their omission leads to ambiguity or makes the sentence hard to read.

Middle. There are several violations of rules that have been taught—as many as usually occur in the average paper.

Low. Basic punctuation is omitted or haphazard, resulting in fragments, run-on sentences, etc.

7. Spelling

High. Since this rating scale is most often used for test papers written in class, when there is insufficient time to use the dictionary, spelling standards should be more lenient than for papers written at home. The high paper usually has not more than five misspellings, and these occur in words that are hard to spell. The spelling is consistent: words are not spelled correctly in one sentence and misspelled in another, unless the misspelling appears to be a slip of the pen. If a poor paper has no misspellings, it gets a 5 in spelling.

Middle. There are several spelling errors in hard words and a few violations of basic spelling rules, but no more than one finds in the average paper.

Low. There are so many spelling errors that they interfere with comprehension.

8. Handwriting, Neatness

High. The handwriting is clear, attractive, and well spaced, and the rules of manuscript form have been observed.

Middle. The handwriting is average in legibility and attractiveness. There may be a few violations of rules for manuscript form if there is evidence of some care for the appearance of the page.

Low. The paper is sloppy in appearance and difficult to read.

THE MEASUREMENT OF GROWTH IN WRITING ABILITY

The only scientific way known to the writer to measure growth in writing ability by means of essays is to have all students in a span of three grades write a paper on the same topic and on the same day, at least four times a year and preferably six or eight. To keep nervous teachers from coaching students on the topic set for each date, the department may first agree on a long list of topics as suitable for short, impromptu

compositions to be written in class. Then, at the beginning of each testing day, the department head may simply announce, "Today we'll use Topic 7," or "Today we'll use Topic 18." All English teachers write this topic on their blackboards, read aloud any explanatory material that accompanies it, and devote that day to the writing of test essays. Students number their own papers with any number of six digits that pops into their heads, such as 924,332 or 001,644, and they write no other identification on their papers. They copy this number on a 3 x 5 index card and add their name, grade, curriculum, other designations such as "regular" or "honors," and their teacher's name. These cards are locked up by the principal until the grading is finished.

The papers are distributed in a random fashion to all members of the department and rated on the scale previously discussed, without knowledge of the identity of the writers or their grade, curriculum, or teacher. In experimental studies, these ratings are usually recorded on separate 3 x 5 cards and no comments or corrections are written on the papers, so as not to influence the ratings of a second reader. For ordinary school use, however, each student may be asked to write a column of numbers from 1 to 8 in the upper left-hand corner of his first page. These numbers refer to the eight qualities defined in the rating scale, and the teacher who first gets the paper records his ratings on a scale of 1 (low) to 5 (high) opposite each of these eight numbers. When the paper is returned to this student's English teacher, he rates the paper again and records his ratings to the right of those already recorded by the first reader. He then adds together both sets of ratings to get a total rating for that paper, which may range from 16 (low) through 48 (average) to 80 (high). After four test papers, the cumulative total ratings may range from 64 to 320.

At the end of each period on testing days, when students hand in their papers, their teacher sorts the papers into as many piles as there are teachers and/or readers to read them. If there are eight, he sorts the papers into eight piles. At the end of the testing day, he cross-stacks these piles and takes them to the room of the department head, who has eight chairs lined up to receive them. Each teacher drops one pile of his papers on each chair until each chair holds a random eighth of the papers written in each English class that day. Each teacher or reader picks up his eighth and rates the papers at home. After a little

practice, most teachers learn to rate these short test papers in about two minutes per paper if they do not write in corrections. They may, however, write a brief comment on anything they like or dislike.

Teachers often complain that they do not know how to rate a paper if they do not know whether it comes from the tenth or twelfth grade or from regular or honors classes. They hold up a paper and say that it should get a 4 in some quality if it comes from a regular class but only 2 or 3 if it comes from an honors class. There are many replies to this objection, but the most devastating is that, if they had this knowledge, the effect would be precisely the opposite. Benjamin Rosner of Brooklyn College added such bits of information to otherwise anonymous papers to see what the effect would be; what the readers did not know was that half of his information was true and half was false. Papers labeled "boy" received the same average grades as when they were labeled "girl," but papers labeled "honors" received average grades that were significantly higher than when these same papers were labeled "regular." This deception was tried out on so many teachers in different schools that there is no doubt that this tendency is general. We find what we expect to find. If we think a paper was written by an honors student, it looks better than if we think it was written by a regular student.

Anyway, all that the rating yields is a series of numbers representing total ratings on each paper. These numbers can then be adjusted for grade and curriculum before being translated into grades that will stand in the record. One simply makes a distribution of these totals for each curriculum within each grade. Then, if it has been decided that the tenth-grade regular students include (let us say) 20 students who ought to get A's, one counts 20 ratings down from the top for that group, draws a line, and calls everything above it an A. The 20 students who stand above this line may not be the same 20 who "ought" to get A's (and who *will* get A's on the other bases that were used in coming to this decision), but at least this procedure assures the desired proportions of letter grades for each group. No one gets a D simply because he is a tenth-grade vocational student who cannot yet meet the competition of higher grades and harder curriculums. If he stands high among his own group, he gets a high grade, no matter where his total rating falls in the distribution for the entire school.

This latter distribution, however, will show him where he stands in relation to the entire student body and how his standing changes from one year to the next. In grade 10, the average student stands in the lowest third of this distribution; in grade 11, in the middle third; in grade 12, in the top third; and in all grades the academic students tend to stand far above the non-academic. This is a realistic view of one's competition, and it is the only scientific way thus far developed to measure the amount of improvement in writing from one year to the next. The idea that teachers can judge the amount of growth by the old process of marking papers severely at the beginning of a year and leniently at the end is utter nonsense that ought not to deceive a child. It does, but it is a deception that should not be practiced on the young. Growth can be plotted only when each test paper is judged against a background of a representative sample of papers from the entire school, and only when the teachers do not know which papers are which. Then, if a student accumulates 128 points in his first year, 192 in his second, and 256 in his third, the rise in his standing is meaningful.

THEME GRADING

Ordinary grading of homework assignments in composition cannot make use of the rigorous departmental procedures we have recommended for test essays. On the whole, it is better not to attempt anything of the sort, since anonymity works better in testing than in instruction. One of John McNulty's sketches is charmingly entitled "A Man Like Grady, You Got To Know Him First." To help a student, you also have to know him first. It remains to be seen, however, whether it is wise or appropriate to grade these homework assignments at all, so long as the test essays are there to give the student his bearings. Many teachers prefer to give their reactions and suggestions entirely by written or spoken comments. Others like to use the rating scale, but only as an estimate of probable ratings had this been a test essay, not as marks that stand in the record. This appears to be a matter of preference. One must only remember that the homework assignments reveal problems that have no numerical solutions. It would be unfortunate if ratings on these papers were mistaken for answers and thereby headed off any real effort to find answers.