DOCUMENT RESUME

ED 091 750 CS 201 355

AUTHOR Diederich, Paul B.
TITLE Cooperative Preparation and Rating of Essay Tests.
PUB DATE 66
NOTE 17p.; Reprinted from "English Journal," April 1967. Paper presented at the Houston meeting of the National Council of Teachers of English. See related documents CS 201 320-375

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Composition (Literary); *Educational Research; *Evaluation Methods; Intermediate Grades; Language Arts; *Measurement Instruments; Post Secondary Education; Research Tools; Resource Materials; Written Language
IDENTIFIERS *The Research Instruments Project; TRIP

ABSTRACT

To evaluate the quality of written compositions, researchers at Educational Testing Service developed the Composition Evaluation Scales (CES), after factor-analytic studies of the reasons teachers gave for their judgments of compositions. This is a set of eight scales: ideas, organization, wording, flavor, usage, punctuation, spelling, and handwriting. Each scale is marked on a five-point line--with the scales of ideas and organization receiving double weight--yielding a total score of 50. The CES is most appropriately used with expository papers on a set topic. [This document is one of those reviewed in The Research Instruments Project (TRIP) monograph "Measures for Research and Evaluation in the English Language Arts" to be published by the Committee on Research of the National Council of Teachers of English in cooperation with the ERIC Clearinghouse on Reading and Communication Skills. A TRIP review which precedes the document lists its category (Writing), title, authors, date, and age range (intermediate--postsecondary), and describes the instrument's purpose and physical characteristics.] (RB)

NCTE Committee on Research

The Research Instruments Project (TRIP)

The attached document contains one of the measures reviewed in the TRIP committee monograph titled:

Measures for Research and Evaluation in the English Language Arts

TRIP is an acronym which signifies an effort to abstract and make readily available measures for research and evaluation in the English language arts. These measures relate to language development, listening, literature, reading, standard English as a second language or dialect, teacher competencies, or writing. In order to make these instruments more readily available, the ERIC Clearinghouse on Reading and Communication Skills has supported the TRIP committee sponsored by the Committee on Research of the National Council of Teachers of English and has processed the material into the ERIC system. The ERIC Clearinghouse accession numbers that encompass most of these documents are CS201320 -CS201375.

TRIP Committee:

W.T. Fagan, Chairman
University of Alberta, Edmonton

Charles R. Cooper
State University of New York
at Buffalo

Julie M. Jensen
The University of Texas at Austin

Bernard O'Donnell
Director, ERIC/RCS

Roy C. O'Donnell
The University of Georgia
Liaison to NCTE Committee
on Research

NATIONAL COUNCIL OF TEACHERS OF ENGLISH
1111 KENYON ROAD
URBANA, ILLINOIS 61801

Category: Writing

Title: E.T.S. Composition Evaluation Scales

Authors: Paul Diederich, John French, Sydell Carlton

Age Range: Intermediate--Post-Secondary

Description of Instrument:

Purpose: To evaluate the quality of written compositions.

Date of Construction: 1961

Physical Description: The CES was developed by researchers at Educational Testing Service after factor-analytic studies of the reasons teachers gave for their judgments of compositions. It is a set of eight scales: ideas, organization, wording, flavor, usage, punctuation, spelling, and handwriting. Each scale is marked on a five-point line--with the scales of ideas and organization receiving double weight--yielding a total score of 50. In the full report where CES appears, the high, middle, and low points on each scale are described in detail.

The CES is most appropriately used with expository papers on a set topic. It can be compared with the London Scales for creative or imaginative writing reviewed in this monograph.

Validity, Reliability, and Normative Data:

The validity of CES resides in its basis in a study of teachers' reasons for their judgments of compositions. Like all rating scales it has high face and "content" validity since it is used with whole pieces of written discourse.

Diederich claims that with practice teacher-raters can achieve a reliability of .90 for a cumulative total of eight ratings, two each on

four different papers by the same writer. In the reports noted below Diederich outlines a school-wide cooperative rating scheme based on the CES.

Ordering Information:

EDRS

Related Documents:

Diederich, Paul B. "Cooperative Preparation and Rating of Essay Tests," English Journal, 56 (April 1967), 573-584, 590.

Diederich, Paul B. "How to Measure Growth in Writing Ability," English Journal, 55 (April 1966), 435-449.

ENGLISH JOURNAL

The Official Journal of the Secondary Section of
National Council of Teachers of English

Editor: RICHARD S. ALM
University of Hawaii

Volume 56 April 1967 Number 4

Cooperative Preparation and Rating of Essay Tests

Paul B. Diederich

Senior Research Associate
Educational Testing Service
Princeton, New Jersey

MY real topic is the improvement of measurement in education by cooperative action of departments or teaching teams, but I shall give particular attention to the cooperative preparation and rating of essay tests and examinations. I believe that neither the grading of essays nor any other measurement performed by teachers is likely to improve until responsibility for measurement of the most important objectives in each field has been transferred from individual teachers to the department or team.

For the past 32 years I have been visiting schools in many parts of the country, trying to help teachers with their problems of testing, grading, recordkeeping, and reporting. Although this is not a common occupation, I have not been alone in this endeavor. Hundreds of courses in tests and measurements have been offered, summer institutes and workshops on evaluation have multiplied, dozens of books and thousands of articles have been written, and hundreds of communities have brought in consultants on measurement to provide in-service training. In addition to these outside influences, almost every school district tries to improve its report cards about once every ten years. What has been the result?

As I visit schools now and examine everything that teachers are doing to appraise what their students have learned, I cannot point to a single important change from the measurement practices of 1935, when I first began visiting the 30 school systems that were involved in the Eight-Year Study. Each teacher still makes up his own tests and examinations without the help or criticism of his colleagues, and it is still uncommon for two or more teachers to grade them independently. These tests and examinations rarely get at anything more than knowledge and skills. When they try to get at anything like creativity, imagination, appreciation, critical thinking, or attitudes, they usually come acropper. The reliability of these tests is rarely computed, and item analysis is almost unknown. Students are still marked on almost everything they do, and teachers

Editor's Note: This paper was presented at the Houston meeting of NCTE, November 1966.

have no idea that there is any way to find out how much measurement of a given objective is enough. At the end of each marking period, they add together different kinds of measures of different objectives and translate the "average" into a single course grade. Although this process seldom has any rational or mathematical foundation, it probably causes teachers, students, and parents more trouble, worry, and heartache than any other aspect of school work. No competent investigator would use these grades as evidence of what the students had learned, but they pass as coin of the realm even though we all know that many of them are worthless. This was the situation in 1935, and that is the situation today. Why have the time, effort, and ingenuity devoted to improving measurement by teachers produced not one single change of any consequence in 32 years?

The only common characteristic of all these efforts at improvement that I can think of is that they left the problem where it had been to start with: namely, in the laps of individual teachers. My own conclusion is that the individual approach has clearly failed, and there is no reason to suppose that it will succeed any better in the next 32 years. I see no hope for any significant improvement until the individual approach is abandoned, and measures of the four, five, or six most important objectives in each field are prepared, reviewed, revised, administered, scored, reported, and analyzed by cooperative action of departments or teaching teams.

This leaves room for two other types of evaluation that should continue to be handled by individuals: what I call "instructional evaluation" by teachers and "self-evaluation" by students. Instructional evaluation includes everything that a teacher does in class, in conference, and in grading his own tests and assignments to keep an eye on how things are going. Its principal function is the guidance and reinforcement of his teaching. I have some doubt that it should ever enter or affect the permanent records of students, but this may be going too far. Self-evaluation by students is not as well understood or used as instructional evaluation, but it can be argued that students should have a recognized part in evaluating their own development until they are almost independent of external evaluation—as they must be most of the time in adult life. In other words, one of the objectives of each field must be to help them become more competent and more responsible evaluators of themselves. Both instructional evaluation by teachers and self-evaluation by students are outside the jurisdiction of the department or team. But there remain usually four, five, or six major, continuing objectives of instruction in each field that are better regarded as the collective responsibility of the department than as the individual responsibility of each teacher. It is my contention that responsibility for the measurement of these objectives should be transferred from individual teachers to the department or team, and that immediate and striking improvements over the measurement practices of individual teachers will result.

BEFORE I tell you how to do it, I should say something about five objections or questions that occur to everyone immediately, and that may lead you to reject this scheme out-of-hand and to stop reading at this point. The first is that departmental examinations will drastically curtail freedom of teaching. Suppose one teacher prefers to introduce the detective story through a short story by Conan Doyle while another prefers Edgar Allan Poe. Will both have to use the same story to prepare for the departmental examination?

Not at all. Our practice is to set for the examination an entirely different story—for example, one by Carter Dick-

son or Rex Stout—that all teachers are forbidden to discuss and that students have to read on their own before the examination. They learn how to interpret and analyze such a story through any example of the detective story that each teacher likes to teach, but they show that they know how to do it through a somewhat easier story that they have to figure out for themselves. This practice avoids the danger that literary competence can sometimes be counterfeited by ability to recall and reproduce—to "parrot back"—the teacher's analysis and interpretation. Since questions are written by several teachers and reviewed by others, it also avoids the danger that the examination may be dominated by the point of view of a single teacher. This requires an ability to diagnose the teacher that is hard on students with independent minds. Occasionally one hears them saying that they answered as they did, not because they believe it but because that is the sort of thing their teacher likes. In the situation I have described, they do not know who wrote the question or who will grade their answers, and so the safest thing to do is to write what they really think.

A second objection to departmental examinations that may be in your minds runs something like this: "If I had to work on an examination with those feebleminded buzzards who make up my department, I'd resign." I know that feeling, but in my experience it makes no difference. You don't have to like or trust your colleagues to prepare a good examination or other measure by a division of labor. My hunch is that the result is somewhat better if not too much brotherly love prevails in a department, for then the parts of the examination will be prepared with greater care and criticized more rigorously. The worst measures I have seen were produced by a team in which there was so much togetherness that they thought they had to meet every day and gabble until the questions somehow got written. Very little talk is needed. There is one meeting to agree on an outline (under the firm guidance of the department head) and to divide up the work of preparing the questions. These are circulated in photocopies, and all members of the committee write in their objections and suggestions for improvement. After some time for revision, the department head confers with each author to see whether all reasonable objections have been satisfied. It is seldom necessary or wise to call a second meeting to consider the revised examination. All proposed changes are settled in these conferences. If any part of the examination has to be graded, no one need fear that his enemies will stick a knife into his students, since the papers are identified only by code numbers, usually chosen at random by the students themselves. The papers are also graded independently by two different teachers, and whenever there is a substantial difference between their grades, the disputed papers are referred to a small committee of the most trusted readers.

A third objection is sure to arise at this point: "We are too busy already, and work on these departmental exams will impose an additional burden on us that we simply cannot accept." You may not believe me until you get involved in cooperative measurement, but I will stake my reputation on the promise that after you learn how to do it and cut out the busywork that it makes unnecessary, it will reduce the whole task of measurement, grading, and record-keeping to a properly subordinate role. The principle of a division of labor was discovered a long time ago, and its uniform effect is to reduce the time it takes to get a job done—as well as doing a better job.

For example, take the job of grading papers for writing ability. In the school district in which I have done the most work on this problem, many devoted teachers used to think they had to grade a paper a week; others were willing to

settle for a paper every two weeks; and no one thought he could get away with less than a paper a month. But when we studied this problem scientifically to determine how many papers were necessary to get reliable scores on eight components of writing ability, the answer turned out to be four papers a year, each rated independently by two different teachers. Every student in the three junior high schools of this district writes a test paper in his English class on the same topic and on the same day during November, January, March, and May. Each student numbers his own paper with any number of six digits that pops into his head and writes no other identification on his paper, but he copies this number on a separate slip and adds his name, grade, teacher, section, and the date. These name-slips are locked up until the rating is completed. The papers are arranged in the numerical order of these self-chosen numbers, which puts them in an obviously random order, and are then divided into as many piles as there are teachers to rate them. Each pile has about the same number of papers that each teacher gets on an ordinary homework assignment (between 120 and 150). But these test papers, planned and written within one class period, are much shorter than homework papers, and they represent a much wider range in ability, because each teacher gets papers all the way from the top class in Grade 9 to the bottom class in Grade 7. Since the differences among these test papers are much more obvious than among those that a teacher gets from any one class they are quicker and easier to rate. Moreover, teachers are forbidden to write any comments or corrections on these papers, because that would influence the judgment of the second reader. They encircle one number for each quality on rating-slips like the one below:

Topic_____ Reader_____ Paper_____

	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	
						———
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	
						———
					Sum	———

How long does this take? Now that we have developed a systematic way of doing it, and teachers have had a good deal of practice, the answer is an average of two minutes a paper. We find that it actually increases accuracy to work rapidly, and so we encourage teachers to trust their first impressions and rate boldly—with confidence that any serious error in judgment will be caught by the second reader. They do not have to believe that the second reader will be a better judge but only that he is unlikely to misjudge any given paper in the same

direction. Hence, when one rating is far off the beam, the other rating is likely to differ in its total by more than ten points, and all such papers are referred to a small committee of the most experienced teachers for a third reading. At present, we find that only about one paper in 12 requires a third reading. Since most papers get two readings at two minutes apiece and a relatively small number get three, the total time for rating one of these tests of writing ability now averages about five minutes per student. How much time does it take to grade, correct, and comment on the average homework paper? Our figure is eight minutes per paper, and this was confirmed by a careful study under different auspices in California. Hence, even with the double reading, the time spent by teachers in rating one of these writing tests is less than they spend on a homework assignment—and of course there is no homework assignment during the four weeks per year in which these papers are written and rated.

The reliability of the cumulative total of eight ratings on four test papers per year normally reaches or exceeds .80. This is lower than one wants in a controlled experiment but high enough for a practical judgment in the ordinary course of schoolwork—and much higher than one ordinarily gets. This means that if we added a fifth test paper, it would not change the relative position of enough students to justify the additional time. You can see why if you consider the large number of rating-points that students accumulate. The lowest possible total for the year is 80 points; the average is 240 points; and the highest possible total is 400 points. This spreads the students out so widely that an additional rating would not change the picture very much or in very many cases. That is what I meant when I said that most teachers have no idea that there is any way to find out how much measurement of a given objective is enough. A department soon finds out. We stop when the reliability of our cumulative total reaches or exceeds .80.

Of course, we do not reduce practice in writing to these four test papers per year. Before each test there are at least four homework papers that receive careful comments—but why grade them? I know your answer: "Because students raise Cain if we don't." That is true, but you should add, "under present conditions." When the grade that enters the record depends on the average of these homework papers, naturally they want to know how well they did on each one. But when the grade depends entirely on how well they write in four tests, they soon regard the homework papers as training for the tests, and then they value tips on what they did well or badly more highly than grades. Cutting out grades on homework papers saves time, worry, and arguments. Hence, even if rating the tests took more time than a homework assignment, it would save time in the more difficult task of dealing with 16 to 30 homework papers per year.

You may be thinking, "But grades based on these tests are obviously unfair. Since papers from Grades 7, 8, and 9 are mixed together without identification, seventh-graders are bound to get the lowest ratings and ninth-graders the highest." So they do. That is why students receive at least two general indications of their position after each test. The first is their position up to this point in Grades 7-8-9 combined. That is a very important figure, because that is the one that moves. Since there is a great deal of natural growth in writing ability during these grades, the average student stands in the lowest third of this distribution in Grade 7, the middle third in Grade 8, and the highest third in Grade 9. Hence, we can measure growth much more accurately and convincingly than by our usual practice of grading severely at the beginning of each year and more leniently toward the end.

The second general indication of position is where each student stands among other students in the same grade with whom he may reasonably be compared: for example, remedial, regular, or honor students. This is more nearly like present grades, but note that remedial students are not forever condemned to the equivalent of D's and F's, nor are honor students guaranteed the equivalent of A's or B's. We show them where they stand in their own league, but we also know where they stand in the total population of the school.

An incidental benefit of this double grading of unidentified papers is that it puts the teacher and his students on the same side of the fence. He wants all of them to make the best possible showing on each test, but he cannot *give* them a high grade; their papers will be rated anonymously by all members of the department. If a student gets a lower rating than he expects, his teacher can say quite honestly, "I have no idea who gave you that rating, and I have no power to change it. But let me get your paper and show you what you need to work on. I'll help you, and if you work hard, you can improve your position in the next test." That is a refreshing change from the present situation in which we have to argue with some students over the grades we "gave" them. With departmental measures, these arguments vanish.

Now let me turn to a fourth question about these examinations: "Will our teaching be judged by the results?" Certainly not. Everyone knows that some classes are brighter, better prepared, and more highly motivated than other classes, and their high scores or ratings can lead to no defensible conclusions about their present teacher. When teachers analyze the results of these examinations, they first look at the kinds of questions or tasks on which each meaningful subgroup of the population did well or badly. For example, in the writing tests we usually find that students in these grades show greater improvement in ideas, organization, and wording than in mechanics, and that students from disadvantaged areas quite naturally have the most serious difficulties with mechanics. But occasionally we find that some of the disadvantaged students have improved much more in mechanics than we usually expect—even though their scores are still low. How in the world did they do it? Sometimes their teachers can offer a pretty shrewd guess. As other teachers of these students try similar procedures, they may find a similar improvement. Our policy is always to look for some favorable result and to try to discover what accounts for it. As these successful practices are more widely adopted, they will automatically replace the less successful. In any case, we do not want teachers to think of their measurement program primarily as a way of finding out what they are doing wrong. We prefer to look for things that work.

A host of other objections to cooperative measurement are summed up in the statement, "I don't think I'd like it." That is quite natural, for teachers tend to be the most conservative element in the community, and they can be counted on to oppose any procedure that is unfamiliar to them; but after they get used to it, they will defend it to the death against any further change. One advantage of cooperative measurement is that it makes very little difference whether one likes it or not. It gets sold to the superintendent, the Board of Education, and the principals on the ground that no significant change has come about in the measurement practices of individual teachers as far back as anyone can remember, and it is high time to adopt a departmental approach that has power to initiate change. The administrators then bring together the department heads or team leaders in an Evaluation Committee, and in that public setting—one after another answers the questions, "What objectives will your department

try to measure? On what dates? By what means?" When these people agree that a certain objective will be measured by a certain procedure within certain dates, no individual teacher can ignore it. The measure is prepared by a division of labor and administered on the scheduled dates, and all students to whom it applies take it. Then the papers are scored or graded and the results analyzed by the teachers who are given this responsibility, and a report on these results is given at the next meeting of the Evaluation Committee. At no point is there an opportunity to say, "I don't think I'd like it." It is simply assumed that if a department professes to be teaching something, it has an obligation to present some sort of evidence that that thing is being learned.

It is noteworthy that, whenever and wherever I have initiated a departmental measurement program, I have never known a department head to report failure to prepare or administer a promised measure within the scheduled dates. These are public commitments, motivated in part by rivalry with other departments, and it would seriously embarrass a department head to report in a meeting of his peers that his group had failed to meet its obligations. Compare this record with the usual result of exhorting individual teachers to go home and improve their measurement procedures. They may try something once, although even that is unusual, but thereafter they go on doing what they have always done, and what teachers before them have done for generations. That might be all right if these traditional practices were satisfactory to teachers, students, and parents, but we hear complaints about them on all sides. They are maintained only by inertia and custom. On the other hand, in a departmental measurement program, changes can be initiated and maintained by the binding force of public commitments, deadlines, and reports. The control is democratic, but things get done.

I know that this sounds hardboiled, and it is intended to be hardboiled, for I am fed up with exhorting teachers to do something intelligent about measurement and getting nowhere. As a matter of fact, however, as soon as teachers get involved in cooperative measurement, they like it. It makes the job easier, quicker, and more interesting by a division of labor; it puts teachers and students on the same side of the fence; it reveals answers to many teaching problems; it provides ammunition against our critics; and it adds fun and excitement to both teaching and learning. Incidentally, it brightens up the usual meetings of departments or teams because the teachers have something of real substance to work on together, and it yields results that they all want to discuss.

I have now dealt with five objections to cooperative measurement:

1. It will interfere with freedom of teaching.
2. It is disagreeable to work on examinations with other teachers.
3. It will take too much time.
4. Teachers will be judged unfairly by the results.
5. "I don't think I'd like it."

I PROMISED to clear these objections out of the way before telling you how to do it, but I was not quite honest. In the course of dealing with these objections, I think I have given you a pretty clear idea of how this plan works. The first step is to appoint an Evaluation Committee, consisting of heads of departments and special services, such as library and guidance. In the school district in which I have done the most work on this program, we built up this committee gradually. In the first year it represented guidance (including the assistant principals with special responsibility for discipline), English (together with the library), and social studies. In the second year we added mathematics, science, and foreign languages. In the third year we took on the fine and

practical arts, vocational education, and physical education. This kept us from having to develop measures of too many different objectives in any one year. We were also content to start with cooperative measures of even one or two objectives in each field, knowing that if we broke the ice, other objectives would gradually be added. The Evaluation Committee met only four times a year but each time for a full morning, with substitutes hired to cover classes. A clean break with the individualistic tradition of school evaluation cannot be made by tired people who always have to meet after school. The real work went on behind the scenes as committee members met with their departments or teams in their own schools to prepare, review, revise, administer, score, report, and analyze the results of the measures for which they were responsible.

I hope you will not go away with the impression that all of these cooperative measures have to be something unusual, like the tests on literary works that all teachers were forbidden to discuss, or the four writing tests per year. The backbone of every school measurement program is the "ordinary" subject-matter examination that is given four, five, or six times a year. I have put "ordinary" in quotation marks because, when teachers work on these examinations together and expect them to provide defensible measures of the most important objectives of their program, they turn out to be anything but "ordinary." There may be nothing unusual about the format, but the questions are prepared and criticized and the answers are scored or rated with a very clear idea of the objectives that are to be measured.

Some of the other measures that we have developed are extremely simple but helpful to both students and teachers, like our Record of Independent Reading, which is kept on 3 x 5 index cards. As soon as a student finishes a book or decides to give it up, he fills out one of these cards with his name, grade, and the date; author and title; a number indicating the type of book; a rating of how much he liked it; and an indication of its difficulty (easy, medium, hard). Then he writes a candid comment about the book for the benefit of other students who are looking for something to read. In the periods reserved for independent reading, he sees other students using these cards, looking for a book that their friends have recommended with apparently genuine enthusiasm. Hence the comments are extremely candid, and some of them curl the teacher's hair, but there must be no reprisals or these cards would lose their value for other students. Teachers also find them useful. In preparing for a conference on reading, they leaf through these cards and get a pretty clear idea of what the student likes and dislikes and the types of books that he has not yet explored. One of our most important and most disturbing findings also grew out of this record. We found that there is a serious and widespread decline in the number of books read independently beginning in Grade 9. The ninth-graders turned in just two-thirds as many book cards as the eighth-graders in all three schools. We ran through many glib explanations of this decline and finally came to one that concerned us deeply: that this is the point at which most students have to make the transition from juvenile to adult reading, and a surprising number can't do it. As a result, our tests on literary works focus on the difficulties in adult books that the average student cannot cope with, and we are making a concerted effort in our classes to find out how these difficulties can be overcome.

WITH this general background in mind, let me turn to the preparation and grading of essay tests and examinations. Although this is the happy hunting ground of the individual teacher, it is in this field that I see no possibility what-

ever that an individual can improve his measures all by himself. As for the preparation of these tests, if I had to pick out a single fault of a typical essay examination that I would condemn above all others, it is that it depends altogether too much on the slant—the opinions and preferences, and sometimes the ignorance and dogmatism—of a solitary individual, unchecked by the criticism of his colleagues. Although teachers try to be tolerant of divergent opinions and undoubtedly welcome them when they are expressed cogently by superior students, the safest course for the average student is to give the teacher what he wants. The only antidote I know to this dominance of a single point of view is the cooperative preparation, review, and revision of examination topics or questions, and of the guidelines that are to be used in grading the answers.

As for grading the essays, an individual teacher never finds out when he is wrong—or may be wrong—because other teachers never disagree with him. He sees the students every day in class and quickly forms an impression of their ability, attention, industry, thoroughness, and the like. Then, when he reads their papers, knowing who wrote them, he unconsciously reads into the papers either more or less than is actually there.

This effect was prettily illustrated in a study conducted by Dr. Benjamin Rosner a few years ago in which test papers from 12 school districts were sent to a central office where all identification except a code number was removed, and the papers were sent back in a random order to be graded. The teachers protested that they could not grade them fairly unless they knew at least whether they came from regular or honors classes because honor students should be graded by higher standards. Dr. Rosner said that this presented an opportunity to find out what information about the writers was essential to accurate grading, and he promised to supply the information they wanted one bit at a time on subsequent papers. Hence the papers were stamped either "Boy" or "Girl," "Grade 9" or "Grade 10," "Regular" or "Honors," and so on. What the teachers did not know was that half of this information was true and half was false. The papers had been written on carbon-backed forms, so that Dr. Rosner had three identical copies of each paper. One of these was stamped "Regular" while another copy of the very same paper was stamped "Honors." He made sure that no school received both copies of the same paper, but otherwise the papers were sent back in a random order.

"Regular" vs. "Honors" proved to be the only bit of information that made any difference, and the effect was the opposite of what the teachers expected. The papers stamped "Honors" received significantly *higher* grades than the other copies of the very same papers that were stamped "Regular." The explanation seems to be that we find what we expect to find. If we think a paper came from an honors class, we expect it to be pretty good—and that is what we find. But if we think it came from a regular class, we expect it to be only so-so—and that is what we find. If a single word stamped on a paper can have that much effect on grades, consider how much effect the full personality of the student must have. That is why papers so rarely surprise us. We keep on reading into them our impression of the student that we gathered during his first month in class. And even when the paper does surprise or disappoint us, we may change too little. I often used to think, "Too bad; he had an off day. I'm afraid I'll have to reduce his grade to a B." But the same paper written by a poor student might easily have received a D or an F.

HENCE, I believe that the first step toward the improvement of essay grading is to find out how widely teachers disagree when they all grade photocopies of the same paper and do not

know whose paper they are grading. In college I used to reproduce about one paper a month and have it graded and commented on by all members of the department. At the beginning of each year I never failed to get every grade from A to F. In our next meeting I would write on the blackboard how many gave it an A, how many gave it a B, and so on. Although the teachers were always shocked, I tried not to be. I would just say that this always happened, and the only thing we could do about it was to argue out our differences. Then I would turn to some highly respected teacher and ask him why he gave it an A. After listening to his explanation, I would turn to some friend of his and ask him why he gave it an F. The curious thing was that both teachers often saw the same things in a paper but weighted them differently. One might say that there were a great many careless errors, but what counted was that the boy had something to say and said it rather forcibly. The other might reply that this was true, but when a student with so much natural talent had gone this far in school without bothering to learn the ordinary conventions of writing, he gave the paper an F to show him that he could not get away with it any longer.

That brought up a policy question: how should we grade a paper that had ideas and managed to get them across but contained this many mechanical errors? On the other hand, how should we grade a paper that was impeccable in mechanics but said practically nothing? As we argued over questions like these—not in the abstract, but in the presence of examples of what we were talking about—we gradually came closer together. In judging anything as complex as writing ability, however, I think it is unrealistic to expect a higher average agreement in a department than is represented by a correlation of .5. This is the usual correlation between height and weight. It is by no means hopeless. As I have previously illustrated, all that is necessary to get it up to a reliability of .8 is four samples of each student's work, each rated independently by two readers, with a third rating for papers on which there is substantial disagreement.

Some teachers profess astonishment at the low level of agreement that I expect and tell me that in their department they hardly ever disagree on an essay grade by more than a plus or minus. I know how to do that, too. One way is to put the grade at the top of each paper and back it up with a number of corrections and comments in red ink; then hand the papers to some other teacher to see whether he agrees. Of course he will. Grading is such a suggestible process that a paper with a B on it already begins to look like a B paper. But, you may say, I put my grade on the back and ask him not to look at it until he puts his grade on the front. I am sorry, but I cannot believe that this way of concealing the prior grade is very effective, because I get nothing like this agreement when there is no grade or comment written on the papers by any teacher and when the readers do not even know who graded them previously.

Another way to reach high agreement may be illustrated by an essay question I remember from an examination on Homer's *Odyssey*: "Write a unified essay on the women in the *Odyssey*." This is the "unstructured" type of question that literature teachers love. It is supposed to get at ability to organize material, independent thinking, critical insight, originality, imagination, and other fine qualities. But the specifications used in grading the answers were quite different. First, the staff made a list of about 12 women in the *Odyssey* that they thought students should remember and gave five points for each one that a student mentioned. But they subtracted one point for misspelling the name, another for omitting or mistaking the place where she lived, and a third for mentioning her out

of order. Then they put down three things about each woman that they thought students should remember and gave either one, two, or three points for each one, depending on the accuracy of the statement. At the end, they allowed each reader to give from one to five points for what they called "good writing." Each paper was graded quite independently by two readers, and they boasted that the average agreement or correlation between pairs of readers was .80. I did not doubt it, but what about all those fine objectives? All that they really measured was total recall of what happened plus ability to spell some rather difficult Greek names.

SINCE this is obviously not the best way to grade an essay question that has some factual content, what is a better way? First, in the way the question is stated, I believe that there ought to be more "structure," for in my experience the "unstructured" question gives more weight to memory than we want. Even if we are not so obvious about it as the staff I mentioned, we are unconsciously influenced by such details as getting Circe on the wrong island, misspelling Nausicaa, or forgetting about the slave-girl Melantho. I would give students most of the details that this staff expected them to remember: a list of women in the order of their appearance, the place where each lived, and one fact about each one that would recall her to their minds, such as "Circe, Aeaea, changed men into pigs." Then I would indicate that they were not expected to comment on all of them but on not more than five or six that would illustrate the points they intended to make. I would even go so far as to suggest some of the kinds of points they might make, such as the traits of character in Odysseus that these encounters brought out, the ways in which these women resembled or differed from modern women, or the speculation of Samuel Butler that the prominence of women in the *Odyssey* suggests that it was composed by a woman. Of course I would indicate that these points were intended only to illustrate the kinds of points they might make, and that they should feel free to comment on anything they noticed about the women in the *Odyssey* that struck them as interesting.

The next thing I would like would be for two or three members of the committee to write a short paper on this question within the time limit to be observed by students. Such papers may bring to light unusual treatments of the topic that might not occur to most staff members and that might be rejected as unsound if they were first encountered in student papers. In other words, the staff papers may break up preconceived ideas of the sort of essay that students ought to write. They may also suggest some of the qualities that should be looked for in superior papers, and they may keep the younger staff members from expecting more of students than teachers can do.

I would usually insist that students bring their copies of the *Odyssey* to such an examination, and I would have some extra copies for those who forgot. This open-book policy reduces fear of the examination and our own reliance upon accuracy of recall in setting the questions and grading the answers. It also enables us to encourage students to support their points by relevant short quotations. I myself believe that even examinations on some portion of a textbook in history, science, and the like should usually be open-book examinations, but I can imagine some situations in which this would be unnecessary or inappropriate.

In preparation for grading the answers, I like two or three staff members to take home a number of papers and bring back sample papers to illustrate one or more types of good, average, and poor answers, possibly with a few comments pointing out the distinctive characteristic of these papers. These may be duplicated in

photocopies and discussed in a short meeting before the papers are distributed for rating. If any staff member finds some papers hard to rate because they bear no resemblance to any of the sample papers, he should be encouraged to discuss them with a staff member who worked on the selection of these samples. Usually we insist that nothing be written by teachers on any test paper, but that ratings (and comments when necessary) be recorded on a separate sheet, or sometimes on a small rating-slip for each paper. These are handed to the department head along with the papers as soon as each reader finishes his set. The department head locks up the ratings in a safe place but hands on the papers to some other reader, usually selected at random, for a second rating.

After this second rating, both the papers and the ratings are usually arranged in the numerical order of the code numbers written on the papers, and someone pulls out the papers on which the two ratings differed by more than a certain amount that the department will learn by experience. Usually it is an amount that will cull out not more than ten percent of the papers for a third reading by a small committee of the most experienced and trusted readers. Some departments average all three ratings; others substitute the third rating for whichever of the two previous ratings is farther from the rating of the committee. If the ratings are recorded on separate small rating-slips, I myself like the practice of recording the committee rating in red on the rejected rating-slip and filing this slip under the name of that rater. This practice frightens teachers when they first hear about it, but they soon find that nothing disagreeable happens. Everyone must expect to have some of his ratings rejected, but usually just two or three of the newer members of the department accumulate a considerable number of these "rejects." Some time after the examination, one member of the review committee goes over these papers with the staff members whose ratings were rejected and explains why the committee thought that their rating was too low or too high. He listens politely to anything they may have to say in reply and agrees with their good points but tries to correct any misunderstanding that comes to light. There is no reason whatever to regard these private sessions as a reproach, and no one else need know about them. The newer staff members just have to learn the standards prevailing in the department, and this takes time and help. I can think of no more tactful or effective way of doing it. Usually these readers are brought within reasonable distance of departmental standards within a year, and only once in several years do we find a reader whose judgment is so erratic that he probably ought not to grade these essay questions in departmental examinations. Even that is no calamity, for there are plenty of other things for him to do. For example, he may be particularly good at devising objective questions, or he may be a superb director of plays. In these days of team teaching, we should not expect teachers to be equally good at everything.

By way of contrast, some administrators have a touching faith in what they call the "training" of readers by some consultant on measurement. They sometimes invite me to meet with their English department between 3:00 and 4:00 some afternoon, and in that time they expect me to show them how to grade papers in a way that will yield fabulous agreement. There is no magic secret that can be taught in one session. In that time I can only get them to worry about the problem, but they have to work out a solution for themselves, and it takes a long time. If I became a department head, I should expect it to take about three years before we could establish reasonably uniform standards in grading even those few examinations that we all worked on together. Even these standards

(Continued on page 590)

Cooperative Preparation and Rating of Essay Tests

(Continued from page 584)

can hardly be regarded as a straitjacket. Remember that in rating anything that is very complex, I expect only as much agreement between two independent ratings as we usually find between height and weight. If even that amount of agreement seems unduly restrictive to some of your proud, independent spirits, I wonder why we should pretend to be able to teach anything like good writing if no two of us can agree even this much on what it is.