DOCUMENT RESUME

ED 097 702 CS 201 625

AUTHOR Diederich, Paul B.
TITLE Measuring Growth in English.
INSTITUTION National Council of Teachers of English, Urbana, Ill.
PUB DATE 74
NOTE 107p.
AVAILABLE FROM National Council of Teachers of English, 1111 Kenyon Road, Urbana, Illinois 61801 (Stock No. 03460, $2.50 nonmember, $2.25 member)

EDRS PRICE MF-$0.75 HC-$5.40 PLUS POSTAGE
DESCRIPTORS *Composition (Literary); *English Instruction; Evaluation Methods; *Grading; Higher Education; Language Arts; *Measurement Techniques; Secondary Education; *Test Reliability; Writing Skills

ABSTRACT

The monograph is a complete outline for a program designed to help English departments institute logical and fair procedures for grading student essays. The contents in this monograph include "Factors in Judgments of Writing Ability," "The Effect of Bias," "Measuring Improvement in Writing," "Personal vs Staff Grading," "Standard Scores for Test Essays," "Computing the Reliability of Essay Grades," "Computing the Reliability of Objective Tests," "Design for an Examination in English Language Arts," and "Imitating Staff Grading of Test Essays." The appendixes, which comprise the second half of this monograph, include "Descriptions of Papers Rated High, Middle, and Low on Eight Qualities," "Topics for Essays," "Objective Items Based on a Central Theme," "Discrete Types of Objective Items," and "Learning to Write." (RB)

Measuring Growth in English

Paul B.
Diederich
Senior Research Associate
Educational Testing Service

National Council of Teachers of English

Library of Congress Catalog Card Number 74-84480
NCTE Stock Number 03640

National Council of
Teachers of English

Contents

BOSWELL: Sir Alexander Dick tells me that he remembers having a thousand people in a year to dine at his house: that is, reckoning each person as one each time that he dined there.

JOHNSON: That, Sir, is about three a day.

BOSWELL: How your statement lessens the idea!

JOHNSON: That, Sir, is the good of counting. It brings everything to a certainty, which before floated in the mind indefinitely.

BOSWELL: But *Omne ignotum pro magnifico* [Everything unknown passes for marvelous]: one is sorry to have this diminished.

JOHNSON: Sir, you should not allow yourself to be delighted with error.

BOSWELL: Three a day seem but few.

Boswell's *Life of Samuel Johnson*, April 18, 1783

Foreword

Somehow the teaching of English has been wrenched out of the Age of Aquarius and thrust into the Age of Accountability. Many of us view educational accountants in much the same spirit as we view the agent of the Internal Revenue Service coming to audit our returns. Theoretically, it is possible the agent will turn out to be a pleasant person, gregarious and affable, who writes poetry in his free time and who will help us by showing how we failed to claim all our allowable deductions, so that the result of the audit is the discovery of a new friend and a substantial refund. But somehow we doubt that possibility.

For the specialist in measurement and testing we have our image, too. In his graduate work, one of the foreign languages he studied was statistics. And he passed it. The other one was that amazing and arcane language the testing specialists use when they talk to one another. He passed it, too, and is fluent in it. He doesn't think of children except as they distribute themselves across deciles. He attempts with his chi-squares to measure what we've done without ever understanding what we were trying to do. Not so with the author of this monograph.

Paul Diederich, an eminent specialist in testing and measurement, is as pleasant a surprise as the IRS agent described above. The surprise begins with his academic background: three degrees in Latin and Greek classics from Harvard and Columbia. It extends through his first teaching assignment: high school Latin. It continues to this day. He still publishes articles on classical subjects and may be the only testmaker who reads Latin and Greek for pleasure at the age of 68.

The question remains, "But does he know anything about teaching and testing in English?" Fortunately, yes. Just after he began teaching Latin, the Great Depression set in, and soon both students and their parents became far more interested in survival than in the classical tradition. Noting

the decline in his classes, he projected that by 1940 he would be down to zero students. So—"like a rat deserting a sinking ship" as he expresses it—he swam over to a language that appeared to have a future, namely English, and soon became an Associate Professor and Examiner in English at the University of Chicago. Meanwhile he had been a member of the Evaluation Staff of the Eight-Year Study and helped develop several tests, including a measure of interests in twelve subjects, now called AIM (Academic Interest Measures), the only instrument inherited from that study that is still published by Educational Testing Service. At Chicago, the Board of Examiners was called upon to develop a large number of tests for the United States Armed Forces Institute, and over two million servicemen received school or college credit in English through tests developed by Diederich and his associates.

In 1949, soon after Educational Testing Service was formed by a merger of three non-profit testing agencies, its president Henry Chauncey went to the Middle West looking for fresh blood for his Research Division. He came back with Diederich and then discovered that they had been classmates at Harvard. During the teacher shortage, Diederich had a hand in promoting the employment of college-educated housewives to help high school English teachers deal with their overload of student compositions. These were first called "lay readers" but soon became "English assistants" when it was found that they were equally effective in supervising independent reading rooms. The latter enabled English teachers to cut their large classes in half by teaching one section Tuesday and Wednesday, the other Thursday and Friday; the section that was not in class went to independent reading. On Monday there was a large-group presentation in the auditorium, and the teachers who were not involved had this day free for conferences with students.

The initial conception that led to this monograph was not Diederich's. It started at ETS with a colleague who is a friend of both Diederich and NCTE. The thought was to gather together into a single collection a variety of manuscripts and published articles by Diederich to make available to English teachers ideas and insights from his lifetime of experience and research in the teaching and measurement of English.

As has happened so many times before, Diederich gave more than was asked for. Having consented to the original plan, he worried with us about the inevitable occurrences of repetition of ideas among papers on related topics. What we thought was an editorial problem he took as a writing problem. His solution was to write an entirely fresh manuscript. It follows.

Robert F. Hogan
Executive Secretary, NCTE

1

Introduction

As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student's writing, especially if the writing is done under test conditions in which one can be sure that each sample is the student's own unaided work. People who uphold the view that essays are the only valid test of writing ability are fond of using the analogy that, whenever we want to find out whether young people can swim, we have them jump into a pool and swim. If they can swim the length of the pool and back, the evidence is undeniable that they can swim.

But suppose we already knew that all of these young people could swim somehow or other—some well, others badly—and the test was to find out how well each one could swim. Then we might use five judges, each of whom would independently write on his scorecard a number from 1 (poor) to 5 (excellent) indicating his opinion of each person's swimming. Then suppose that over a long period of time, at every level from elementary school through college, and in several countries, everyone who tried this procedure reported that about a fifth of the swimmers received every grade from 1 to 5 and only a handful received less than three different grades from the five judges. Wouldn't this cast some doubt on the reliability of this test of swimming?

This is the situation we usually face in grading essays as a test of writing ability. We already know that practically everyone who is admitted to the test will write *something*. Our task is to determine how well each one writes. Then we must use judges, and their judgments are likely to scatter even more widely than judgments of performance in sports, since there are well-defined standards for most sports but standards for writing are neither well defined nor widely accepted. The principal task of this booklet will be to suggest ways of improving the reliability of grades on essays. We shall find that it is very hard to reach a desirable standard of reliability

through essays alone, and so we shall also consider the inclusion of a few sections of objective items on related parts of proficiency in English. Since objective items yield far higher reliabilities than essays per unit of time, they will usually increase the reliability of the total score on the examination to a level that is fair to students.

But why measure or grade at all? I hesitate to answer this question because, to anyone who buys or borrows a booklet with this title, the question is silly, the answer is obvious, and it is tedious to repeat the old twaddle about the need for accurate information on which to base educational decisions, and the like. But just now there is a vocal minority among English teachers who oppose any use of grades or measures that enter the permanent records of students—especially those that indicate weaknesses, and they are likely to introduce a resolution at the next NCTE meeting condemning the procedures recommended in this booklet unless something is said in defense of these procedures.

First let me surprise these critics by saying that I agree with practically everything they say. This is not a rhetorical trick, I really mean it. During my twenty-five years at the Educational Testing Service, one of my principal duties has been consulting with secondary schools on problems of measurement, grading, record-keeping, and reporting. I have had to visit more classes than I care to remember, and my predominant impression has been that these classes are fantastically over-evaluated. Students are graded on practically everything they do every time they turn around. Grades generate anxiety and hard feelings between students, between students and teachers, between students and their parents, and between parents and teachers. Common sense suggests that they ought to be reduced to the smallest possible number necessary to find out how students are getting along toward the four or five main objectives of the program, but teachers keep piling them up like squirrels gathering nuts. They appear to have no idea that there is any way to find out how much measurement of any objective is enough.

Of course there is, and they should have learned it in some course or unit on tests and measurements. If they have not, they will certainly know it by the time they finish reading this booklet. The answer is *reliability*. The concept will be fully understood only after studying and trying out the procedures I recommend, but the general idea is that there are quick and easy ways to find the amount of random variation in all measurement operations, and from that amount one can tell how much more evidence of the same kind is needed to reach a stable figure that will not change very much, or in very many cases, no matter how much more evidence is gathered.

Over the years I have come to accept a reliability of .80 in the measure (or series of measures) of an important objective as adequate for practical decisions in the ordinary course of schoolwork. In this booklet I suggest an

examination week at the end of each quarter or semester in which one day is reserved for English language and literature, another for foreign languages and literature, a third for history and social science, and so on (pages 41-42). The essay and objective parts of the examination on English language arts are virtually guaranteed to yield the desired reliability in just one day of testing. During the following week, most students are on vacation, but make-up examinations are scheduled in the same order for students who were absent or who wish to improve their grade. When students repeat an examination for this purpose, whichever grade is higher stands in the record. The recommended scoring procedure yields convincing evidence of the average amount of improvement in writing from one grade to the next within each curriculum, and it shows students how much their writing improves on successive examinations.

At this point, before we explain why it is necessary, some readers will be shocked to learn that such an examination requires two essays. Those written in the morning are graded independently by two teachers, and those written in the afternoon are graded independently by two different teachers. Whenever the two grades differ by more than a certain amount, the paper is referred to a small committee of the most experienced teachers, who substitute their own grade for whichever of the original grades was farther from their own. Before the examination, the teachers indicate how many students in each of their classes they expect to make each grade—not which students, but how many. These estimates are added and converted to percents as guidelines to the number of papers the teachers should expect to find at each level of merit. Their pooled judgments need not look anything like the normal curve. If they have reason to believe that the group is superior, they may aim at a distribution in which no one fails, only 10 percent get D's, 50 percent C's, 25 percent B's, and 15 percent A's. Ways of combining the four essay grades and four objective scores are suggested that will make the distribution of final grades conform to these expectations.

These procedures, we hope, will seem more and more reasonable and feasible as we proceed. Right now, however, many readers are probably thinking, "How unrealistic! We are already overworked, and it is hard enough to get our examination grades turned in on time when there is only one essay that we grade ourselves. Two independent ratings of two essays by each student plus a review of discrepant grades are out of the question."

But how much time does this procedure actually take? I recently introduced this type of examination in several junior high schools in which we secured accurate records of the time spent in grading, since most of the papers were graded in Saturday workshops. We encouraged the teachers to work rapidly and to trust their first impressions, since we found that this increased the reliability of grading. Besides, they could count on the

fact that any serious error in judgment would probably be caught by the second reader and the review of discrepant grades, not because the other readers are wiser but because all three are unlikely to err in the same direction. The essays were short, and there were no corrections or comments to write. In fact, they were forbidden to write anything at all on the papers lest it bias the judgment of later readers. Grades were recorded on separate work sheets.

The average grading time per essay proved to be two minutes. Two essays per student each graded twice came to eight minutes per student. Only 10% of the grades were far enough apart to require review, and since each review also took two minutes, the average grading time per student came to just under nine minutes. We had previously made a careful study of the time required to grade, correct, and comment on homework papers. It averaged eight minutes per student, and this result was confirmed by a similar study under different auspices in California. Since there are no classes during examination week, the teachers did not find this chore unduly burdensome. The objective exercises were scored by clerks and aides.

What did the teachers get in return? First of all, they had reliable scores on writing ability and other language arts, and they could prove it to the satisfaction of the Board, their principal, and their director of testing. They also had convincing evidence of the average amount of improvement in writing per year in each curriculum, and they could show students and their parents how much improvement in writing was revealed in successive examinations. They got such figures at the end of each quarter or semester—often enough to keep in touch with the progress of each student. If a student received a lower grade than his pride would accept, he could take the make-up examination, and whichever grade was higher would stand in the record.

Remember now that once your measures reach a satisfactory level of reliability, adding more measures of the same abilities will not change the position of many students, and none very far. Most of these teachers had required a paper every two weeks from their students, and if they were conscientious about it, it took about forty hours a week to grade, correct, and comment on them. Now that they had a reliable measure of writing ability, the grades were superfluous, and a careful study convinced us that the corrections were more damaging than helpful. Hence they refused to grade the homework papers; they cut out most of the corrections; and they concentrated on brief marginal comments, emphasizing what the student had done well. At the end, however, they might add one suggestion for the improvement of the next paper, but rarely more than one.

Here the Defense rests. In the rest of this booklet I outline a system for the evaluation of language arts that cuts out more than 90 percent of the grading that goes on day after day in almost every classroom. Fewer and better measures at longer intervals of time are enough to show students,

their parents, and their teachers how they are doing. At other times teachers should be free to devote their whole minds to teaching and students to learning. I firmly believe that measurement should be reduced to a properly subordinate role in education, but if you think you can get away with less than the minimum I have recommended in this booklet, the experience of a lifetime in the field warns me that you are unlikely to succeed.

2

Factors in Judgments of Writing Ability

Teachers who have never graded a set of papers that have previously been graded by another teacher seldom realize how commonly and seriously teachers disagree in their judgments of writing ability. The most impressive evidence I can offer on this point came out of a factor analysis of judgments of writing ability that John French, Sydell Carlton, and I performed at ETS in 1961. We secured 300 papers written by students in their first month at three different colleges and had them all graded by sixty distinguished readers in six occupational fields. As our academic judges we had ten college English teachers, ten social science teachers, and ten natural science teachers. As our non-academic judges we had ten writers and editors, ten lawyers, and ten business executives. For various reasons seven of these sixty judges were unable to complete their assignments, but all six fields were adequately represented by the fifty-three judges who remained. These were all outstanding people who were deeply concerned about the way students write.

In an actual examination, we bring all the judges together and spend a day or two discussing grading standards and rating sample papers until we reach an acceptable degree of consensus. But in this 1961 study we wanted to find out what qualities in student writing intelligent, educated people notice and emphasize when they are free to grade as they like. Hence we never brought these sixty judges together; they graded all the papers at home. Our only directions were to sort the papers into nine piles in order of general merit, using their own idea of what constituted general merit. The only rules were that all nine piles must be used, with not less than twelve papers in any pile. Then, on as many papers as possible, we asked them to write brief comments on anything they like or disliked.

Hence, the reliability of grading that was shown in this study should not be taken to represent the reliability usually attained in grading essays for the College Board, when we adopt strict rules and enforce them by close supervision. But it is probably typical of the amount of disagreement one would find in any large group of readers without such training and discipline that, out of the 300 essays graded, 101 received every grade from 1 to 9; 94 percent received either seven, eight, or nine different grades; and no essay received less than five different grades from these fifty-three readers. As the first step in our factor analysis, we had to compute the correlation—the amount of agreement—between the grades of each reader and the grades of each other reader. The median correlation in this large (53 x 53) table of correlations was .31.

This table was subjected to a complex mathematical procedure called "factor analysis," which has the effect of picking out clusters of readers from all over the table who agree within their cluster and disagree with every other cluster to a greater degree than could be attributed to chance. In effect, it determines how many different schools of thought exist among the readers as to what constitutes excellence in student writing. In this study we found five different schools of thought—five clusters of readers who were evidently judging the papers on somewhat different bases, since within each cluster there was a moderate amount of agreement on grades but a substantial amount of disagreement with every other cluster.

We have not yet taught the computer how to tell us what these clusters were agreeing on, so we resorted to a classification of the comments they had written on most of the papers. In a trial run, when we used a random sample of readers, the first result of this classification was utter chaos, for every cluster appeared to be commenting on everything. The picture only became clear and convincing when we restricted the classification to the three readers who stood highest on each factor—that is, who came closest to the central tendency represented by each factor—and to just those papers that these readers had graded either high (7-8-9) or low (1-2-3). Even with this restriction, we finally tabulated 11,018 comments on 3,557 papers under 55 headings, and we reduced the numbers of comments tabulated under each heading to percentages of the comments written by each of these selected readers, so that those who wrote the most comments would not unduly influence the interpretation.

Interpretation of the Five Factors

Then it became quite clear that the largest cluster (16 readers, drawn from all six occupational fields) was most influenced by the *ideas* expressed: their richness, soundness, clarity, development, and relevance to the topic and the writer's purpose. Notice how even this first finding bears

on a point that is often debated by English teachers. Some give little or no weight to the ideas expressed in student papers for two reasons. First, they hold that ideas are the product of God-given intelligence which teaching cannot alter; teaching can only help students express whatever ideas they may have more correctly and effectively. Second, they believe that students have an inalienable right to express any ideas or opinions they have, and any indication by the teacher that some are better than others, and hence deserve higher grades, borders on censorship. Other teachers reply that one *can* do something about the quality, development, and support of ideas in student papers by paying attention to them, raising questions about them, challenging them, and focusing attention on them in class discussion of selected papers. They add that students like it better when teachers take their ideas seriously and react to them than when they confine their attention to errors in expression. Such reactions are seldom intended or viewed as censorship. It is simply a fact that some papers are better thought out than others, and comments to that effect are intended only to encourage students to think carefully about what they write.

However that may be, it is an empirical fact that our largest cluster of sixteen readers from all six occupational fields had by far the highest percentage of comments on the ideas expressed, and lower percentages of comments on the qualities emphasized by the other four clusters. Hence we must accept it as a fact that a high proportion of intelligent, educated adults do pay attention to the quality, development, support, and relevance of the ideas expressed in student compositions and weight them heavily in their judgment of the general merit of these papers. This is certainly one basis on which the writing of our students will be judged, and English teachers will be well advised to give it considerable attention in their instruction and in their comments and conferences on papers.

The next largest cluster (13 readers) had by far the highest percentage of comments on errors in *usage*, *sentence structure*, *punctuation*, and *spelling*. It was no surprise that seven of the ten college English teachers stood high on this factor. This may be a good time to explain why I can cite a number like seven when we tabulated the comments of only the three readers who stood highest on each factor. That tabulation showed us what the factor *meant*—that is, the distinctive emphasis expressed in the comments of the three readers who best represented that factor. Then we could look at the occupational fields of the thirteen readers who belonged to this cluster—whose grades came closer to those given by the three highest readers than to those given by members of any other cluster—and seven proved to be college English teachers. Of the three who stood highest on this factor, however, just one was a college English teacher, another was a science teacher, and the third was a business executive.

The third cluster (9 readers) showed the highest interest of any group in

organization and *analysis*, which appear to be closely related. Four of the seven business executives who completed their assignments stood high on this factor. They were "organization men" in more senses than one.

The fourth cluster (also of 9 readers, but with no occupational bias) stood highest in comments on *wording* and *phrasing*—the choice and arrangement of words, including the deletion of unnecessary words. I suspect that this was at least in part a *vocabulary* factor—that these readers were more impressed than other groups by a large, mature vocabulary, but there was no way to prove it from their comments.

Finally, the fifth and smallest cluster (7 readers, four of whom were either writers or editors) emphasized style, individuality, originality, interest, and sincerity—the *personal qualities* revealed by the writing, which we decided to call "flavor," although they themselves called it "style." We avoided the latter as a label for this factor, since the people who emphasized wording and phrasing were also interested in "style," but in such a different sense that they came out on a different factor. They were interested in style in the use of language, but the fifth cluster was interested in style as the revelation of a personality in writing, as shown by such comments as "forceful," "vigorous," "outspoken," "sincere," or "inflated," "pretentious," "dogmatic," or "sentimental." In any large group of readers, these seven would probably be recognized as the devotees of creative writing, and the fact that four of the seven were professional writers or editors confirmed this impression. You know that the writing of Mark Twain and Edgar Allan Poe is so different in its general character that you could hardly mistake one for the other. It is this sort of difference in the personality revealed by writing that we decided to call "flavor."

If you are interested in numbers, you may have noticed that these five clusters of readers (15 + 13 + 9 + 9 + 7) add up to fifty-four readers, but we had only fifty-three who completed their assignments. This is not a mistake. Although this procedure minimizes overlap among the readers, it was inevitable that some stood almost equally high on two different factors, while a few did not belong to any cluster—they disagreed with everybody. Although it is conceivable that the latter were better judges than anyone else, the probability is higher that there was too much random variation in their grades to associate them with any distinct school of thought.

You may wonder why we did not classify the comments to begin with and call the largest group of comments Factor 1, the next largest Factor 2, and so on. The answer is clear and compelling. If you know only the percentage of comments that can be classified under a given heading, there is no way to tell how much influence this heading had on the way these readers graded the papers. You cannot simply ask them because few if any readers are conscious of what they are actually responding to in student

writing that makes them grade one paper higher than another. Some of the most common types of comments did not come out on any factor since they were made by every type of reader.

Hence you have to find clusters of readers who are judging the papers on somewhat different bases, since there are significant differences between the grades assigned by each cluster, yet a fairly high amount of agreement within each cluster. Then you know that whatever these clusters of readers are looking at has a demonstrable effect on their grades, since their grades do in fact differ. You find out what they are looking at by classifying the comments of the readers who best represent each cluster, and you find that one cluster has the highest percentage of comments on the ideas expressed, another the highest percentage of comments on mechanical errors, and so on. Then you know that these distinctive emphases actually influenced their judgments.

It was interesting and illuminating that we found five and only five distinct schools of thought among these fifty-three distinguished readers, emphasizing ideas, mechanics, organization, wording, and flavor respectively. There is some room for argument as to the exact interpretation of these five factors, but there is no reasonable doubt that our study revealed just five different bases for the judgment of our sample of 300 papers, or that the distinctive emphases of these five ways of looking at student writing could be described fairly accurately by the labels we chose. Another study using a different writing task, different students, and possibly a different age level might yield somewhat different conclusions, but the five factors we found in this particular study are a matter of knowledge, not opinion. We *know* that these five qualities in student writing influenced the judgments of this particular set of readers, and I use the word *know* deliberately. These results are far more convincing than any theoretical, armchair analysis of how students ought to write. We hope, however, that something like this study will be replicated by several different investigators as time goes on, since truth finally emerges only after several independent investigations reach essentially the same conclusions.

There was one other study of this sort, almost concurrent with our own, but we heard about it only after our study was completed. It was done by the Italian psychologist Remondino, using papers written in Italian by eleven-year-olds. Although his method differed slightly from ours, the factors he found could readily be translated into the labels we chose except that he found an additional factor that he called "graphics" and we called "handwriting, neatness." This addition was explained by the fact that he used the original handwritten papers, while we had to use typed copies. Later, when we were having teachers rate handwritten papers, we added Remondino's factor to our list.

You may think, "The reason for the unreliability of essay grades is now

clear: some readers are influenced mainly by the ideas expressed, others by the number of errors they notice, others by organization and analysis, and so on. They are looking at different things in the papers, or they are weighting them differently."

That is true, but it is not the whole story. The extent to which our fifty-three readers were influenced by these five factors is indicated by the sum of their "loadings" on these factors. On the average, the sum of these "loadings" explained 43 percent of the variance in grades; the remaining 57 percent was unexplained. Some of the remainder may ultimately be explained by factors which have not yet come to light or by more reliable measures of the factors we discovered. But most of it is probably due to two causes that are not amenable to factor analysis: unique ideas about grading that are not shared by any other reader, and random variations in judgment, which may be regarded as errors in judgment. The extent of the latter might be revealed by having the same judges grade the same papers six months later, after they had forgotten the grades they originally assigned. The correlation between their earlier and later grades might average no higher than .50, which would indicate a large amount of chance variation in grading. But this would be so expensive, and the readers would be so reluctant to tackle the same papers again that we did not dare to suggest it.

A more detailed explanation of the meaning of our five factors is given in Appendix A. A few research-minded readers of this report may want to examine the full, original report of this factor analysis. It was published (multilithed) by Educational Testing Service in August 1961 as Research Bulletin 61-15, but it has long been out of print. The only way to get a copy now is to ask ETS to make a Xerox copy of its file copy, but that would be very costly, and we advise against it. The full report is ninety-two pages long, extremely technical, and crammed with figures that are no longer relevant. The only use a researcher could make of it would be to study the mathematical procedures used in the factor analysis, but advances in computer technology since that time have made these procedures obsolete; there are now simpler, quicker, and less expensive procedures. One may take it on faith that the procedures we used were sound, and their results valid, because they were designed and supervised by Ledyard Tucker, whose authority in the field of factor analysis is unquestioned. All of the findings relevant to the grading of essays have been reported and explained in this summary.

3

The Effect of Bias

Another danger in grading essays that we must try to avoid is *bias* on the part of the readers—either for or against particular students, the views expressed (such as liberal or conservative), the way of writing (ornamented or plain, lengthy or succinct, etc.). There are even particular types of errors to which some teachers react so strongly that they are likely to fail any paper in which they appear, no matter how good it is in other respects. Bias appears most obviously when a teacher is grading the papers of his own students, knowing who wrote them. If a teacher reads the paper of a boy known to be dull, lazy, careless, and impertinent, it would take a remarkable paper to overcome the prejudice that the teacher has formed against him. On the other hand, if the paper was written by a model student, or by one with whom the teacher sympathizes because he has recently had serious trouble at home, the grade is likely to be higher than a dispassionate analysis of the writing would warrant.

Even when the paper of a given student surprises or disappoints us, we are likely to change too little. When I get a poor paper from a good student who generally writes well, I tend to think, "Too bad; he had an off day. I'm afraid that I'll have to lower his grade to a B." But if that same paper had been written by a poor student, it could easily get a D or an E.

The effect of this sort of bias was prettily illustrated by an experiment conducted in twelve school districts in the state of New York by another man at ETS, Dr. Benjamin Rosner. Since we were comparing four methods of improving writing, we wanted the grades on writing to be highly reliable so that we could detect significant differences, even if they were small. Hence we asked for one test paper per month on a topic selected by the central staff, written on the kind of paper that yields three sharp, clean copies. We kept one of these for our files, removed all identification except a code number from the other two, and sent them back to two different schools for grading.

The teachers who graded these papers knew nothing whatever about the writers—not even which school they attended. Soon they complained that they ought to have at least a little information, such as whether the paper came from grade 9 or 10 (the only two grades in our study), or from a regular or "honors" class, because the latter should be judged by higher standards.

Dr. Rosner said that this was a reasonable request, and it afforded an opportunity for a sub-experiment on the kinds and amounts of information about students that led to the most reliable grading. He promised that all papers would henceforth be stamped with one bit of information each month, such as whether it came from a boy or a girl, grade 9 or grade 10, a regular or "honors" class, and so on.

What the teachers did not realize until Dr. Rosner told them at the end of the year was that half of this information was true and the other half was false. Remember now that two copies of each paper were sent to different schools for grading. One month Dr. Rosner would stamp one copy of each paper "boy" and the other copy "girl." The next month he would stamp one copy "grade 9" and the other copy "grade 10." The next time he would stamp one copy "regular" and the other copy "honors," and so on with different bits of information each month.

The only bit of information that made any difference at all in average grades was whether the papers were stamped "regular" or "honors," and that difference was exactly opposite from what the teachers expected. They had argued that honors classes should be judged by higher standards, but the papers that were stamped "honors" averaged almost one grade-point *higher* than the other copies of the very same papers that were stamped "regular."

The explanation seems to be that grading is such a suggestible process 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. 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 this much effect on grades, think how much effect the full personality of the student must have when we grade papers knowing who wrote them, with all their past behavior and circumstances in mind. Some teachers argue that our knowledge of each student ought to have this effect—that a poor writer who has done his best ought to receive a higher grade, while a brilliant writer who has not come up to his usual standard ought to receive a lower grade than the actual merits of the paper would justify. I can see some justification for this treatment of the twelve to twenty papers per year that are written for practice, but not in the two to four test papers per year that are graded to determine how well each student actually writes. Then we are judging

writing, not students. Praise or blame enters at a later point. The poor writer who finally earns a passing grade of D may be congratulated; the brilliant writer who disgraces himself by getting a B (when he should have made an A) may be taken sternly to task, or comforted, or urged to repeat the examination.

4

Measuring Improvement in Writing

Bias in grading test papers is easily avoided by a procedure for measuring the amount of improvement that comes about in each year of a writing program. I have recommended it in articles in *English Journal* (Diederich, Paul B. "How to Measure Growth in Writing Ability," 55 [April 1966]: 435-49), and it has been adopted by many junior and senior high schools. For this purpose we ask all students in a span of three or four grades (such as grades 7-8-9, or grades 10-11-12, or even grades 9-10-11-12) to write a paper on the same topic on the same day, but not necessarily in the same hour. Each student numbers his paper with any number of six digits (like 928,401 or 003,256) that pops into his head and writes no other identification on his paper. He copies this number on a separate slip and adds his name, grade, class, teacher, and any other information that may be required. These name-slips are arranged in the numerical order of these self-chosen numbers and are locked up until the grading is finished.

Having the students choose their own numbers not only saves the trouble and expense of stamping code numbers on the papers and keeping a record of which student received each number, it also gives students greater confidence that their papers will be judged without knowledge of the identity of the writers. Duplicate numbers are no problem. When the name-slips are arranged in numerical order, the duplicate numbers come together. Then we match the handwriting on the name-slips with the handwriting on the papers bearing these numbers and change the number of the student who comes first in alphabetical order—usually by adding 1 to the last digit. If that results in another duplication, we add 2 or any other number that will distinguish papers bearing the same number.

The papers are also arranged in the order of these self-chosen numbers,

which puts them into an obviously random order—with all three or four grades scrambled together—and are divided into as many piles as there are teachers to grade them. Each teacher records his grades and comments on a separate work sheet and is forbidden to write anything at all on the papers, lest it bias the judgment of the second reader. He turns in both his work sheets and the papers he has graded to the person in charge of the examination, who locks up the work sheets. Then the papers are turned over to another teacher for a second, independent rating—with no knowledge of the grades given by the first reader. Again, the second reader records his grades and comments on a separate work sheet and writes nothing on the papers themselves. Both readers should rearrange the papers in their original numerical order before turning them in.

After all readers have turned in their second batch of papers and work sheets, the person in charge compares the two grades and pulls out all papers on which they differ by more than one full grade-point. That is, if one grade is B and the other C, they will simply be combined to get the final grade; but if one grade is B and the other C-, that is just over the one grade-point limit, and these papers should be reviewed by procedures that will be discussed later. If the "standard scores" for test essays that will be explained later are used, the two scores must be more than ten points apart to qualify for a review. In our experience, after a high school staff has had some practice in grading essays in this manner, only one paper in ten or twelve needs to be reviewed in order to iron out serious discrepancies in grades.

The main point I want to make now is that staff grading of papers written by all students in a given school—on the same topic and the same day—and identified only by numbers chosen at random by each student will completely eliminate bias either for or against particular students. The readers have no idea who wrote any paper—not even the grade in which it was written, nor whether it came from academic or vocational, regular or honors classes—since the papers are all mixed together in a random order. Incidentally, this mixing makes the task of grading the test papers easier, since the stack of papers given to each reader will probably include papers all the way from the top class in the highest grade to the bottom class in the lowest grade of his school. Hence differences in the quality of writing are far more gross and obvious than in the papers one gets from any one class.

Moreover, since each student's writing will be judged by at least four different readers in the course of a year, any bias toward liberal or conservative views, plain or fancy writing, and the like will almost certainly be cancelled out. Four readers are not necessarily wiser than one, but it is unlikely that all four will err in the same direction.

Results in One Senior High School

The grading of even one test essay in this fashion can provide powerful ammunition against our critics, who often charge that students learn nothing about writing in high school. Here are the results of rating 1,065 papers written on the same day in grades 10, 11, and 12 of a senior high school that stood almost exactly at the national average in general verbal ability.

	NONACADEMICS			ACADEMICS		
	Grade 10	Grade 11	Grade 12	Grade 10	Grade 11	Grade 12
HIGH	5%	8%	9%	22%	41%	53%
MIDDLE	34%	53%	63%	65%	52%	42%
LOW	61%	39%	28%	13%	7%	5%
AVERAGE	326	397	455	475	606	650

These papers were rated by eight English teachers on a "stanine" scale of 9 points, which we shall not explain because an easier scale will be explained later. Here it is sufficient to understand that, for clarity in presentation, we called the three top stanines (24%) a *high* rating, the middle three (52%) a *middle* rating, and the bottom three (24%) a *low* rating. The percents show the percentage of students in each grade of the nonacademic and academic curricula who received high, middle, and low ratings.

Since the papers written by nonacademics were mixed with those written by academics, the former could get very few high ratings in any grade; the competition was too formidable for nonverbal students. Their improvement is revealed more clearly by the percentage who received middle ratings: from 34 percent in grade 10 to 63 percent in grade 12. Best of all is what happened to the percentage who received low ratings, which declined from 61 percent in grade 10 to 28 percent in grade 12. Hence, although these nonverbal students could not hope to become really good writers, fewer than half as many in grade 12 wrote a paper that would really disgrace the school as in grade 10.

The improvement of the academics is best shown by the percentage who received high ratings: from 22 percent in grade 10 to 53 percent in grade 12. Since so many moved into higher brackets, their percentage of middle grades had to decline: from 65 percent in grade 10 to 42 percent in grade 12. This does not mean that the middle group of academics declined in writing ability. The three percents in each column have to add up to 100 percent, so if more than half finally achieve high ratings, less than half can remain in the middle group. Their percentage of low ratings declined from 13 percent to 5 percent for the same reason.

How about dropout of the less able writers as an explanation of the improvement shown in these percentages? The dropout rate in this school was negligible. There were only 5 percent fewer students in grade 12 than in grade 10—far too small a difference to account for the massive shifts in percentages across this table.

Could grade 12 have simply been brighter than grade 10? This is a question that the routine collection of standardized test scores year after year is well equipped to answer. The answer was a decisive "No!" There had been no significant difference in verbal ability in these two grade levels when they entered this school. There was, of course, a substantial difference in verbal ability between academics and nonacademics but not between one grade and the next within each curriculum.

The bottom line of the table, labeled "Average," refers to the average stanine scores of all students in each grade of the academic and nonacademic curricula—with decimal points omitted. This omission is a bit of strategy that at first seems dishonest but actually gives school board members and the public a truer picture of the amount of improvement from one grade to the next. Since stanine scores run only from 1 to 9, the "actual" averages in this bottom line would run from 3.26 to 6.50—not from 326 to 650 as we have written them. We first reported the "actual" averages, and the reaction of school board members and even teachers—who ought to know better—was shock and dismay. A typical comment was, "Look at the difference between the averages of 11th and 12th grade academics: 6.06 to 6.50, a difference of only .44 of a point, which is less than the difference between B- and B. Is it worth all the time and effort we put into teaching composition in grade 12 to produce an average difference of less than half a grade-point?"

What such critics do not realize is how sluggish the averages of large groups of students must necessarily be on a scale that has only 9 points. Given the wide range in ability within each grade, the uncertainty of the grading, and the tendency of students to write some papers better than others, would you expect the average of any of these six large groups to be less than 3? or more than 7? If not, the maximum attainable difference in such averages for large groups would run from 3 to 7, and this school came pretty close to it: from 3.26 to 6.50. The smallest difference (.44) between 11th and 12th grade academics is natural and inevitable. As one approaches the top of any scale, differences are bound to get smaller. Already in grade 11 the academics had received almost as many 8's and 9's as English teachers are willing to award. Hence grade 12 could not show much improvement because there was too little room left to detect further growth.

It then occurred to us that we need not call the lowest stanine 1 and the highest stanine 9. These are not entities like inches or pounds; they are dividing lines in distributions of scores; and we may call these dividing lines anything we like, provided they are successive numbers with equal inter-

vals between them. Many test publishers call their dividing lines 30, 40, 50, 60, and 70; the College Board calls them 300, 400, 500, 600, and 700. In this case, we decided to call the lowest stanine 100 and the highest 900. Then we could omit the decimal points with a clear conscience and save much fruitless, uninformed argument.

These corrected averages reveal two points of interest. First, note that the nonacademics finally reach an average of 455 in grade 12 while the academics start with an average of 475 in grade 10. In spite of this large difference in writing ability, note the relative amount of growth in these two groups: 129 points for the nonacademics, 175 for the academics. Before this little study, I asked the English teachers to guess how the improvement of the nonacademics would compare with that of the academics. Most of them guessed that the nonacademics would show no improvement at all, and the most optimistic estimate was that they might possibly show half as much improvement. That was far off the mark: they gained 5/7 as much as the academics. No one thereafter regarded the teaching of writing to these groups as a hopeless task.

The effect of even this first attempt at staff grading of unidentified papers on the morale of these English teachers was remarkable. They had been so beaten down by the complaints of colleagues and parents that they were almost ready to believe that no one learned anything about writing in high school. But after these figures were published on the education page of the local newspaper (surely an unusual outcome of any examination!) they went about with their chests out and chins up, saying, "How long will it be before the science or social studies teachers can show evidence of such substantial growth toward any objective of comparable importance? We didn't know whose paper we were grading, and there was no way to fake the percentages. So if anyone still thinks that students learn nothing about writing in high school, will he kindly explain how these shifts in percentages could occur?"

I should add just this caution in regard to such tables of percentages. I once conducted such studies on the same day in several junior high schools of one better-than-average school district, and one school showed far higher gains from grade to grade than any other. Since I had visited classes in these schools repeatedly and could not recall any difference in teaching methods or skill that could account for this finding, I had to look into their methods of rating the papers. The school with the highest gains had entrusted the task of rating all the papers to its two oldest teachers who had served for many years as College Board readers. The other schools had involved all their English teachers, even though they had done nothing to establish standards, and so their ratings were much less reliable. One can see why this would cut down the apparent gains from grade to grade if one imagines the extreme case in which all ratings were assigned by throwing dice. Then there would be no difference at all between the averages of

grades 7, 8, and 9. Thus any element of chance that enters into the ratings will reduce the apparent gain from one year to the next. This is another reason for trying to increase the reliability of essay grades and for learning how to compute their reliability before comparing gains per year in different schools.

Reporting Results to Students

Although you may agree that the procedure just outlined is a feasible, convincing way of measuring average improvement in writing from grade to grade, you may wonder how it can give a true picture of the status and progress of individual students, once it becomes a standard examination procedure. It seems unfair to younger, vocational, and remedial students, since the mixing of papers together without identification throws them into competition with all other students in the same span of grades. So it does, and for this reason we report two and occasionally three scores after this sort of examination.

First, we report a standard score (of a sort to be explained later) that shows each student where he stands as a writer in the total population of the school. This is a very important figure because it is the one that moves. Since there is a great deal of natural and induced growth in writing ability at this stage of development, an average writer should expect to stand in the lowest third of his school during his first year, in the middle third during his second, and in the highest third during his third. In the traditional grading system, he would get a C in all three grades, and no one on earth could tell him how much, if any, improvement that represented. Instead, we use rather large numbers to show him where he stands in the total school population on each examination, and how he works his way up through this population as he advances from grade to grade.

Second, we report another score showing each student where he stands in the group with which he may most reasonably be compared, such as tenth grade remedial vocational students or twelfth grade academic honors students. Hence, even if a student stands low in the total population of the school, his standing within his own group may be quite respectable. This is the score that corresponds most closely to the kinds of grades that are usually given.

Third, we may or may not report a growth score showing where a student stands in comparison with other students who started at or near the same point. This is not done routinely, however, because growth scores, while highly regarded, are the least reliable of all educational measures, and there are wide differences of opinion among testmakers about how to compute them for individuals. On the whole, I prefer to forget about comparative growth scores and content myself with showing students how far

they have advanced in the school population on each successive examination. How to do this will be explained in a later section. I should mention that if we ever figure out how to get "criterion-referenced" scores on writing ability, the comparison of one student's growth with that of another should present few difficulties. At present I see no way to do this, but so many bright people are working on the problem that there may be a breakthrough at any moment.

5

Personal vs Staff Grading

I have always taught writing (among other things) and have always believed that improvement in writing takes a great deal of practice and guidance. Hence I have nearly always required a paper a week from my students, and in high school I always graded these papers myself. The grading was the most difficult, time-consuming, and agonizing part of the whole teaching process. I did not mind writing brief comments on the good and bad parts of each paper, but deciding the grade was hard. Then, since I always kept office hours after school, the rest of the week would be filled by arguments with students who thought their grade was too low. Some argued, some blustered, some begged, and some broke down and cried. Some even brought in their parents, who were usually convinced that I was prejudiced against their child for some irrelevant reason. If it had not been for those grades, I would have found teaching a pleasant occupation.

Then by a lucky chance, I began teaching at the University of Chicago, which had an examining system somewhat like the one described above. There the opinion of the teacher had no effect whatever on the grades of his students. Grades depended entirely on six-hour comprehensive examinations that were given at the end of every quarter. Students could register for these examinations whenever they felt ready to take them, and if the grade first attained was below what their pride would accept, they could repeat them at the end of each quarter until they reached a grade that they regarded as satisfactory. To encourage such efforts to improve, we made it a rule that when a student repeated an examination, whichever grade was higher would stand in the record.

Although the students taking the writing examination were allowed three hours in both the morning and afternoon sessions of the same day, we tried to set topics that most students could complete to their satisfaction in about two hours. We encouraged them to spend about half an hour planning their paper, an hour writing it, and half an hour revising it. Of course, some students would write a complete paper during the first hour, tear it up, and then write another complete paper in the second hour. The third hour was allowed mainly to keep anyone from feeling hurried and to provide plenty of time for correction and revision.

These papers were identified only by code numbers and were handed out in a random order to all members of the composition staff for grading. The morning papers were graded independently by two teachers and the afternoon papers by two different teachers. Thus two samples of each student's writing were judged independently by four different teachers, selected at random. Papers on which the two grades differed by more than one full grade-point were referred to a small committee of the most experienced and trusted readers, who did not know what grades these papers had received; they knew only that the grades differed. One member of this committee would give each paper a third independent reading, and a clerk would substitute this grade for whichever of the two previous grades was farther from it. If they were equally distant, he discarded the grade nearest the mean, since combining or averaging grades pushes everybody toward the middle, and we want to keep them spread out as far as possible. But if the first two grades were B and D and the third was C, he discarded the lowest grade to give the student the benefit of the doubt.

What was the effect on teaching? After all the years I had spent arguing over grades with students, it was like coming out of a noisy tunnel into clear sunlight. I still required a paper a week, but I refused to grade them. What would be the point? The students knew as well as I that grades on these practice papers would have no effect on the official grade, which depended entirely on the examination. Hence what they valued more highly than grades were tips on what they were doing well or badly. I did not mind writing these bits of advice or talking them over with students in conferences on their writing. In thus dealing with about 24 practice papers written as homework, I could do nearly everything that elementary teachers try to do with personal grades. I could encourage the faint-hearted, challenge the over-confident, and praise everything a student had done that was even a little above his usual standard. I believe very strongly that noticing and praising whatever a student does well improves writing more than any kind or amount of correction of what he does badly, and that it is especially important for the less able writers who need all the encouragement they can get. After noting four or five things in their papers that I found interesting and making only one modest suggestion for improvement, I thanked my lucky stars that I did not have to put down a grade

that would make a liar out of me. Just try writing several favorable comments on a paper and then giving it a grade of D. Which will the student believe? And how much faith will he have in your comments thereafter? An elementary teacher might give the student an A or a B for trying hard, but a college teacher can't do it if the writing is below the minimum that other college teachers will accept. Hence, if we want to use "positive reinforcement" with the students who need it most, we had better rely on comments and conferences and forget about grades on the homework papers. If they finally pass the examination even with a grade of D- or its numerical equivalent, we can congratulate them warmly. "You passed! How perfectly splendid! Keep on writing as well as you can, but now you can give more attention to subjects in which you excel."

I honestly believe that those who defend the practice of personal grading as a holy cause are mistaken about its usual effects. To hear them talk, the teacher is a nearly perfect being who knows all, understands all, forgives almost everything, and encourages everybody. But when I examine whole files of papers that have been marked and commented on by teachers, many of them look as though they have been trampled on by cleated boots, and they must have a shattering effect on a sensitive student. I once wrote a whole paragraph on the sins against decency and tact that I had found in such comments, and the result was that most of my English-teaching friends would not speak to me. What I find it hardest to forgive is misinterpreting what the student wrote and then blaming him for something that he plainly did not say.

Effects of Excessive Correction

I can judge one of the main effects of personal grading by the attitudes of students who land in my remedial course in college. They hate and fear writing more than anything else they have had to do in school. If they see a blank sheet of paper on which they are expected to write something, they look as though they want to scream. Apparently they have never written anything that anyone thought was good. At least, no one ever *told* them that anything in their writing was good. All their teachers looked for were mistakes, and there are so many kinds of mistakes in writing that their students despair of ever learning to avoid them.

The attitude toward writing that these students have developed is well illustrated by a story told by the Russian writer Chekhov about a kitten that was given to his uncle. The uncle wanted to make the kitten a champion killer of mice, so while it was still very young, he showed it a live mouse in a cage. Since the kitten's hunting instinct had not yet developed, it examined the mouse curiously but without any hostility. The uncle wanted to teach it that such fraternizing with the enemy was wrong, so he

slapped the kitten, scolded it, and sent it away in disgrace. The next day the same mouse was shown to the kitten again. This time the kitten regarded it rather fearfully but without any aggressive intent. Again the uncle slapped it, scolded it, and sent it away. This treatment went on day after day. After some time, as soon as the kitten saw or smelled that mouse, it screamed and tried to climb up the walls. At that point the uncle lost patience and gave the kitten away, saying that it was stupid and would never learn. Of course the kitten had learned perfectly, and had learned exactly what it had been taught, but unfortunately not what the uncle intended to teach. "I can sympathize with that kitten," says Chekhov, "because that same uncle tried to teach me Latin."

If everything written by our less gifted writers gets slapped down for its mistakes, and if this treatment continues year after year, can we expect that their attitude toward writing will differ from the attitude of the kitten toward that mouse? I saw the result year after year in my remedial classes. If I asked them to write anything, they reacted as though I had asked them to walk a tightrope sixty feet above the ground with no net to catch them if they fell. It took some time to build up their confidence, to convince them that writing is as simple and natural as talking, and that no reader would mind a few mistakes if he got interested in what was being written about. For some time I never commented adversely on anything they wrote but expressed appreciation of anything I found interesting, no matter how badly it was expressed. After students gained confidence I continued to express appreciation but offered one suggestion for improvement at the end of each paper. If poor writers learn one thing about writing per paper, that is far above the average.

Allow me to insert two bits of advice about revision. Like most English teachers, I believe that rewriting an unsatisfactory paper teaches one as much about writing as writing a new paper, but most students hate it. They ought to get some sort of reward. The most effective reward I have found is to distribute a list of topics that I expect to assign during a quarter or semester, with certain topics starred. Then I tell my classes, "I shall expect all of you to write papers on the starred topics because they take up different types of writing, different rhetorical principles, and the like. But on all the other topics you have a choice. You may write either a new paper on that topic or rewrite a paper on an earlier topic if you were not satisfied with it and have since thought of a better way to treat that topic. If you choose to rewrite, I shall want to see both the original and the rewritten versions."

My second bit of advice is to duplicate copies of one paper on each assignment with wide spaces between lines and ample margins. Students study these papers as homework, grade them, and insert corrections and comments, including laudatory comments on anything they think was well done. In class the next day, we go through the paper paragraph by para-

graph, commenting on everything that was good or bad about it, and suggesting improvements. I have found this practice far superior to "buddy editing" in which pairs of students exchange papers and try to improve them. When only one other student sees a paper, he can usually find only three or four things wrong with it; but when the whole class gets copies of the same paper and has time to mark it up with corrections and suggestions before it is discussed, one student or another will notice everything that the teacher notices. This is the only situation in which I allow a paper to be ripped to shreds. I either ask the writer's permission to exhibit his writing (without identification) or, preferably, use a paper from a previous class. When students do the ripping, they enjoy it and probably learn more about revision than from rewriting one of their own papers, since their author's pride is not involved.

Whenever I suggest this practice, some teachers say, "I do the same thing, only I use a projector." I'm sorry, but that is *not* the same thing. Students cannot take the projection home to edit before the discussion; one cannot project the whole paper at once; and half the time the projection is unreadable. For this task, duplicated copies of the whole paper are indispensable.

In this section I have talked about the effects of bias in grading test essays and how to eliminate it. In so doing I have had to counter the arguments of teachers who believe that it is almost immoral to grade any paper without full knowledge of the student—his ability, background, and circumstances—so that one can adjust the grade to reasonable expectations. Such teachers think of grades as tokens of praise or blame, and that view may be all right up to the end of grade 6; at any rate, it is almost universally held by everyone connected with elementary schools. Above that point, however, both students and teachers come to look upon the results of important examinations as *information*—information that is valuable only to the extent that it is true. I have argued that praise and blame enter later—the poor writer who passes, the average writer who gets a C, and the brilliant writer who gets an A are equally entitled to congratulations. I have also argued that impersonal grading of unidentified papers by all members of a composition staff brings about better relations between students and teachers than the personal grading of elementary schools. In the staff grading system, the teacher is the student's friend and guide, never his taskmaster and judge. He would be delighted if every one of his students made A's, but he can't just give them A's; they have to earn their marks by the impression their writing makes on all other members of the department. At the lower end of the scale, if one of his students fails, or makes a lower mark than his pride will accept, the teacher feels it just as keenly as the student and does everything he can to help the student earn a satisfactory grade when he repeats the examination.

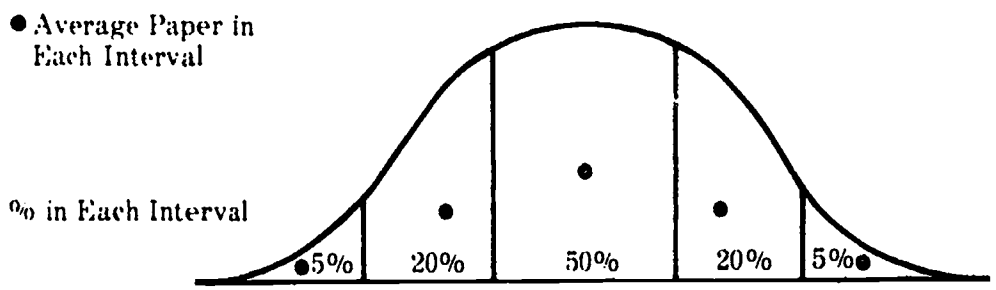
That possibility of repeating the examination if the grade first attained

is unsatisfactory, with the understanding that whichever grade is higher will stand in the record, does more than anything else to take the curse off the system. In secondary schools we have to offer a make-up examination in any case for students who were absent. If it is scheduled a week or so after the regular examination, and if students who were disappointed in their grades are allowed to take it, it will have the effect of reducing fear of the examination and offering a second chance to students who had an "off day."

6

Standard Scores for Test Essays

In staff grading of test essays, each reader gets a large random sample of papers from a large, heterogeneous student population in which it is reasonable to assume that writing ability is normally distributed. This means that each reader should expect to find small numbers of very good and very poor papers, larger numbers of good and poor, and a still larger number of average papers. As the number of papers graded by perfect judges approaches infinity, the distribution of their grades will come closer and closer to the "normal curve" that is crudely represented in the following diagram.



Standard deviation	−2	−1	Mean	+1	+2
Percentile	2	16	50	84	98
Standard score	10	20	30	40	50
Range of scores	1-14	15-24	25-34	35-44	45-59
Letter grade	E	D	C	B	A

In this diagram the distance from left to right represents the quality of the papers—from very poor to excellent—and the height of the curve above the base line represents the proportion of papers that we should expect to find at any given point on the scale of quality.

Of course, in testing any limited number of students, such as a thousand, there will be departures from this curve for two reasons: this sample may happen to include more students than usual at some points on this scale, or our imperfect measures may yield more scores at some points than would be found by perfect measures of the same characteristics. Since judging essays is a chancy business at best, the latter cause is more likely to affect the distribution of grades than the former. Hence, if we observe the proportions predicted by the normal curve in grading large numbers of test essays, we are likely to come closer to the truth than if we rely entirely on intuitive judgments.

The diagram of the normal curve has been divided into five intervals corresponding to letter grades of E, D, C, B, and A. The proportions of test essays in these intervals are 5, 20, 50, 20, and 5 percent. These differ from the proportions traditionally expected but seldom achieved in the United States—10, 20, 40, 20, and 10 percent—but they are common in New Zealand. I have come to accept the smaller proportions of top and bottom grades and larger proportion of middle grades for three reasons.

First, in staff grading of test essays, I have found teachers extremely reluctant to give as many as 10 percent of the papers either top or bottom grades, but they willingly settle for 5 percent. In spite of directions to the contrary, their middle grades always come closer to 50 percent than to 40 percent.

Second, differences in the quality of papers near the middle of the distribution are hardly perceptible. The closer to the mean one sets boundaries for the grade of C, the more differences one finds between the grades of pairs of readers. Teachers grade more confidently, cheerfully, and reliably if one sets these boundaries around the middle half of the papers. Then they want to indicate differences between papers that they regard as high C or low C, commonly expressed as C+ and C−. It is advantageous to have such distinctions in this large middle group, but they can be indicated more precisely by the numerical scores below the diagram, which will presently be explained.

Third, the proportions for the five grades that I now favor have a unique advantage: average papers in each of these intervals lie almost exactly one "standard deviation" apart. More precisely, the middle paper in the B and D intervals lies 1.07 standard deviations from the mean; the middle paper in the A and E intervals lies 2.06 standard deviations from the mean. The middle C, of course, stands exactly at the mean. These very slight departures from exact standard deviations could never be detected by even a

skilled reader, and they would never make a difference in our judgment of a student or in his placement and prospects in school. It is impossible to come closer to exact standard deviations than this without resorting to proportions that teachers would be unable to remember or compute. But nothing could be simpler than first sorting the papers into three piles—top quarter, middle half, and bottom quarter; then, on a second reading, picking out a fifth of the top papers for a grade of A, and a fifth of the bottom papers for a grade of E, or the numerical equivalents of these grades.

Teachers who know some statistics often tell me that I should set the boundaries of the C interval half a standard deviation above and below the mean along the base line, and those beyond the B and D intervals 1.5 standard deviations from the mean. Knowing how teachers grade papers, I am more anxious to have the middle paper in each interval anchored to the standard deviation than to have the boundaries set at mid-points between standard deviations. In the divisions I have chosen, the average distance from the mean of all papers in the B and D intervals is one standard deviation; of all papers in the A and E intervals, two standard deviations.

The Standard Deviation

Now it is time to explain what the standard deviation is and why it is useful. It is an average of the distances (deviations) of all scores or ratings from the mean, but a special kind of average. In the usual kind of average, you would add all the distances from the mean, disregarding whether they were plus or minus, and divide by the number of distances to get the average distance. But in this special kind of average, you first square each distance from the mean, add all these squares, divide by the number of squares to get the average *squared* distance from the mean, and then take the square root. At first you may think that this gets you right back to the average distance from the mean, but it does not. It gives greater weight to scores or ratings that are farther from the mean. Hence it yields a number that is larger than the average distance from the mean, and this number is called the standard deviation.

This computation takes a lot of time, and for most purposes it is unnecessary. A very close approximation of the standard deviation of scores on objective tests (assuming that all are positive numbers) is given by the formula:

$$\text{Standard deviation} = \frac{1.8\,(\text{sum of high fifth of scores minus sum of low fifth})}{\text{Number of students}}$$

This computation is easier than finding the average score—the mean—because you do not even need to add all the scores; only the top and bottom

fifth (rounded to the nearest whole number). You subtract the low fifth from the high fifth, multiply by 1.8, and divide by the number of students to get the standard deviation. In a comparison of several short-cut formulas for the standard deviation (*Journal of Educational Measurement*, Winter 1971), this one proved most accurate.

You have already seen another way to approximate the standard deviation in the case of grading essays. If large numbers of test essays are sorted into five piles in order of merit in the proportions of 5, 20, 50, 20, and 5 percent, the average distance from the mean of the B and D piles will be one standard deviation; of the A and E piles, two standard deviations. Hence you may say that the middle papers in these piles lie one standard deviation apart.

Translating These Letter Grades into Numbers

Since the grades on test essays will have to be added, averaged, combined with objective test scores, and subjected to other computations in what follows, it is necessary at some point to translate them into numbers. When schools and colleges compute grade-point averages, they most often use numbers from 0 (E) to 4 (A), with tenths representing positions between and beyond these whole numbers. I used to prefer numbers from 1 (E) to 5 (A), also with tenths, for two reasons. First, it is unnecessarily insulting to award a student a grade of 0. Second, when large numbers of students are tested, a few of their scores will extend as far as three standard deviations above and below the mean, but only three students in a thousand will score above or below three standard deviations if the distribution is normal. It is possible to indicate these extremes by .1 for the lowest score and 5.9 for the highest if you use numbers from 1 to 5, but there is no way to indicate positions lower than two standard deviations below the mean if you use 0 to 4. Incidentally, 0 is a handy symbol for "no data": the student was absent, was too ill to do himself justice, misunderstood the question so badly that his paper could not fairly be compared with the others, or was suspected of cheating. Such zeros should not be averaged with other grades; they should be omitted until the student takes the make-up examination, which will supply the missing grade.

After using the scale of 1 to 5 (with tenths) for several years, I found that many teachers were having trouble with decimal points in complex computations and regarded them as a nuisance. Students and their parents also regarded tenths as trifling amounts and complained bitterly if they missed a higher grade by what they called "one lousy tenth of a point." It did not alter the true situation in any way, but it made computations easier and everyone happier to call the midpoints of the five intervals 10, 20, 30,

40, and 50 from low to high. We are free to call them whatever we like; many publishers call them 30, 40, 50, 60, and 70. The second digit is understood to refer to tenths of the standard deviation. The range of scores equivalent to each letter grade is then 1-14 for E, 15-24 for D, 25-34 for C, 35-44 for B, and 45-59 for A, as shown below the diagram. Ranges for A and E are slightly extended to get out to three standard deviations above and below the mean, but very few students will ever be found at these extremes.

Since there are now ten points between the midpoints of grade intervals, teachers soon begin using these points to indicate their judgments of test essays more precisely. For example, if a paper is just a shade above a straight C, they may give it a 32; if it is almost on the borderline between C and B, they may give it a 34. I have not found it necessary to set quotas for the number of papers that may be placed at these intermediate points, since repeated combinations with grades of other readers, grades on other test essays, and scores on objective tests bring the final distribution of standard scores close enough to the normal curve for practical purposes. In any case, the second digit does not mean very much, since the "standard error" of such ratings (with the reliabilities usually attained) is roughly 5 points on this scale. This means that if the same essays were graded repeatedly in exactly the same way, and we kept averaging the ratings until we were sure what the true rating was, about two-thirds of these ratings would lie within 5 points of the true rating, but 5 percent of them would be more than 10 points off. Hence all that the second digit can tell us—after all the combining that an examination permits—is whether the final score is closer to B than C, closer to C than B, and so on for the other intervals.

Some teachers speak with scorn of "grading on a curve," but they are thinking of single classes of twenty to thirty students, graded by their own teachers. Everyone knows that some classes of this sort are brighter, better prepared, and more highly motivated than other classes. Perhaps 50 percent of such students ought to get A's, 40 percent B's, and 10 percent C's. In the staff grading situation, in which we are typically dealing with something like 1,000 students, graded by eight different teachers, those are probably the grades that the best classes will get, since their papers will be compared with those from much less gifted and industrious classes. With a number as large as 1,000—usually the total population of a school, or of three grades—it is reasonable to assume a normal distribution of writing ability, and grades may be distributed in accordance with that assumption. But if all the best writers have been placed in one class of thirty students, and their papers are mixed in with the other 970 and graded without identification, nearly all of them should get either A's or B's—barring errors of judgment—and most of these should be caught by the machinery of double grading and review of discrepant grades, as previously explained.

In a larger perspective, sophisticated use of the normal curve is the best guide I know to the proportions of the various grades that different classes should be expected to achieve. Although there are complications that are too technical to explain, and professional judgment may modify the result, the general idea may be conveyed by the following example. It is well known to testmakers that the best predictor of general verbal ability is usually a standardized test of reading comprehension plus vocabulary, taken routinely by all students in most schools. Suppose the distribution of scores on this test in your school looks like this:

Reading + Vocabulary Scores of All Students in This Grade

Lowest 5% (E)	Next 20% (D)	Middle 50% (C)	Next 20% (B)	Highest 5% (A)
0-13	14-24	25-36	37-47	48-60

What percent of your students stood within these ranges of scores?

0%	10%	50%	25%	15%

If your students are indeed as superior to the general run of students in their grade as these reading and vocabulary scores indicate, and if they work up to their ability, then—on tests that are closely related to verbal ability—no one should be expected to fail, only 10% should be expected to get D's, 50% C's, 25% B's, and 15% A's. Such figures, of course, should be taken as only a rough guide to what you should expect, since no short standardized test for a small number of students is a good enough predictor to trust very far. Still, if you gave 25 percent of these students failing grades, your principal would be justified in raising questions.

Professor Edward Gordon of Yale tells about an examination he once conducted for the College Board. He explained and illustrated the scale of five points that was to be used and had the readers practice using it by grading copies of a set of sample papers.

When the actual grading began, he noticed that one military-looking gentleman—an instructor from West Point—was obviously not using the scale. His grades were all two-digit numbers: 53, 71, 83, and so on.

"How do you get these numbers?" asked Dr. Gordon.

"Well, Dr. Gordon," replied the military gentleman, "I'm too old a dog to learn new tricks like that new-fangled scale you wanted us to use. So I just went back to my usual way of grading papers, knowing that you're smart enough to translate my grades into any scale you please. I just count the number of mistakes and subtract that number from 100 percent."

"But what do you call a mistake?" asked Dr. Gordon.

The man's astonishment was obvious. "Why surely, Dr. Gordon, you know what a *mistake* is!"

Setting Grade-Lines in Accordance with Teachers' Predictions

Although standard scores for test essays are nothing more than a translation of letter grades into numerical equivalents, there may be no immediate prospect of getting your school or department to adopt them. Let us see, then, how to get nearly the same results with letter grades, using as predictors the pooled judgment of several teachers as to the number of students in each of their classes who are likely to make each grade on the examination. Suppose their predictions turn out as follows:

Class	E	D	C	B	A	Total
1	0	0	8	10	7	25
2	0	1	7	9	8	25
3	0	1	10	9	5	25
4	1	2	10	7	5	25
5	2	4	11	6	2	25
6	2	5	10	5	3	25
7	3	9	12	1	0	25
8	2	8	12	3	0	25
Totals	10	30	80	50	30	200
Percent	5	15	40	25	15	100

These totals and percents are neater than one would find in actual predictions. They are intended only to illustrate the point that there is nothing wrong about asking teachers to aim at a distribution of grades in which there are far more A's and B's than D's and E's if, in their judgment, the students taking this examination are brighter and better prepared than the general run of students in their school. Such deviations from the normal curve are often recommended by directors of testing.

As these teachers grade the test essays, they should expect to place *about* 15 percent of the papers they receive in their A pile, *about* 25 percent in their B pile, and so on. If they deviate from these predictions by more than 5 percent, they should expect some heated arguments from their colleagues before the grades are turned in. For example, if one teacher fails 15 percent of the papers he grades, he should be prepared to explain why, because the others think that not more than 5 percent of this group should fail. These predictions are based on a great deal of prior experience with these students and should not be disregarded. On the other hand, predictions should not be followed slavishly because, in a group as small as 200, most of the papers that deserve failing grades might fall into the hands of one reader. If the second reader of these papers agreed, then readers of the other papers should find less than the predicted 5 percent of failures.

Without such guidelines, there is no way to tell whether the grades

turned in by the readers are in line with reasonable expectations. With them, each reader will know when he is straying very far from the standards and expectations of his colleagues, and this may cause him to reconsider some of the grades he has assigned. If he still thinks they are correct, he will probably formulate his reasons carefully, because he knows that they will be challenged. When teachers formulate reasons carefully for the sake of (a) explaining grades that are out of line and (b) combatting arguments with colleagues, such teachers are also gradually bringing about closer agreement on grading standards. Such agreement reduces the unfairness to students that often results from insufficient thought and care in grading. Over a period of time, it also makes all members of a department more vividly aware of what they are trying to teach.

Remember that the predictions indicated only how many students were likely to make each grade, not which students. Consequently, even if the final distribution comes out exactly as predicted, there will be many surprises when the teachers find out which students received these grades. Some that they thought were sure to get A's will get B's, and some that they thought would fail will pass.

Although these surprises cause some dismay and argument before the grades are recorded, it is unfair to change the grades of particular students—once their identities are known—simply because their teacher thinks they deserve a higher or lower grade. Such changes would reinstate all the forces of bias, prejudice, favoritism, and idiosyncratic judgments that the staff grading procedure was designed to avoid. They add an unknown allowance for hard work, compliance with requirements, attention in class, sympathy with the student's misfortunes, etc. to the original meaning: a simple measure of competence in English. Such changes also cause the staff an endless amount of trouble. It is almost impossible to keep it a secret that some grades were changed at the insistence of a teacher. Then, as soon as Carlos learns that Emile's grade has been changed, he comes to his teacher with tears in his eyes and begs him to insist that his paper be reconsidered also. Soon almost all students except those who were agreeably surprised by their grades will besiege their teachers with requests to have their papers reviewed. Unless all the teachers hold the position that the only way to change a grade is to take the make-up examination, they will probably revert to personal grading the following year.

Even though they do hold the line, it often happens that teachers feel some injustice was done to their students. They may win support for weighting teacher evaluations in the final grade decision. The weight most often adopted is half for the course grade, determined by each teacher, and half for the examination grade, determined by the staff grading procedure plus scores on the objective sections. Incidentally, if the staff wishes

to have the course grade count as much as the examination, it is wise to insist that course grades be turned in before examination grades are reported. Many teachers are so uncertain about their judgments that if they think a student should get a B but the examination says C, they change their minds and put down a C. I recall two colleges, one of which secured course grades in freshman composition before the final examination, the other after the examination grades were reported to teachers. In the former, the correlation between the two grades was usually about .60; in the latter, about .80.

7

Computing the Reliability of Essay Grades

To find out whether the reliability of grades on test essays in your department stands in need of improvement, and whether the procedures I have recommended or any other procedures bring about improvement, you need a way of computing reliability that is easy to understand and takes very little time. English teachers are often allergic to computation and have neither time nor interest enough to learn the complex method of computing reliability that is explained in books of statistics.

Fortunately there is a quick and easy way to do it that I call "top-quarter tetrachories." It applies to any set of papers that has been graded independently by two readers. For this purpose both must indicate which papers they would place in the top quarter in general merit. This must be precisely the top quarter, rounded to the nearest whole number. For example, if there are 215 papers, both readers must indicate which 54 papers they regard as the best. This causes no extra trouble because, in the grading procedure I have recommended, one starts by sorting the papers into three piles: top quarter, middle half, and bottom quarter. Since the papers are usually identified only by code numbers, the first reader arranges the top quarter in numerical order and sends a list of their numbers to the person who will compute the tetrachorie. When the second reader gets this batch of papers (rearranged in their original random order), he or she does the same. The person in charge has a list of all 215 numbers in numerical order, and puts a check after each number that the first reader put into the top quarter, then a check after each number that the second reader

put into the top quarter. The person in charge counts how many numbers have two checks: that is, how many papers were placed in the top quarter by both readers.

Let us suppose that twenty-six numbers appeared in both of their lists. To change the number to the percent of this group that both readers placed in their top quarter, divide 26 by 215, the total number of students. Here it is 12 percent, corresponding (in the following table) to a tetrachoric correlation of .50. This is an estimate of the amount of agreement between the two readers, which is regarded as the reliability of *one rating*. Since we intend to use the sum or average of both ratings as the grade on each paper, that reliability must be "stepped up" by the Spearman-Brown Prophecy Formula: twice the correlation (between the two ratings) divided by one plus that correlation. That brings the reliability of this set of essay grades up to .67, as shown below 12% in the third line of this table:

Percent in top quarter of both														
06%	07%	08%	09%	10%	11%	12%	13%	14%	15%	16%	17%	18%	19%	20%
Tetrachoric correlation														
.00	.07	.17	.26	.34	.42	.50	.57	.64	.70	.75	.80	.85	.89	.92
Reliability of sum or average														
.00	.13	.29	.41	.51	.59	.67	.73	.78	.82	.86	.89	.92	.94	.96

The standard but more difficult way of computing correlations between two sets of essay grades or other measures is called "product-moment" correlation. Roughly speaking, tetrachoric correlations mean the same thing as product-moment correlations, but they are less precise and more subject to chance variation. In technical terms, the standard error of a tetrachoric is approximately twice as large as that of a product-moment correlation for groups of the same size.

Still, tetrachorics are better than nothing, and if they are computed routinely in all essay testing operations—between pairs of readers, between morning and afternoon essays, and the like—they will tell you whether the reliability of essay grades in your department is improving, and whether it has reached a level that is adequate for practical decisions in the ordinary course of school work.

Over the years I have come to accept a reliability of .80 as adequate for that purpose, especially when the examination includes objective sections that yield far higher reliabilities than essays per unit of testing time. But the reliability of the essay grades we computed as an example on page 33 was only .67. That is far from satisfactory, but it is typical. Even after working with an English staff for some time, I have rarely been able to boost the average correlation between pairs of readers above .50, and other examiners tell me that this is about what they get.

Fortunately there is another form of the Spearman-Brown Prophecy Formula that tells how many times to increase the length of a test—a number usually represented by *k*—to attain any desired reliability.

$$k = \frac{\text{(the reliability you want) times (1 - the reliability you got)}}{\text{(the reliability you got) times (1 - the reliability you want)}}$$

Since we want .80 and got .67, this becomes:

$$k = \frac{.80 \times (1 - .67)}{.67 \times (1 - .80)} = \frac{.80 \times .33}{.67 \times .20} = \frac{.2640}{.1340} = 2 \text{ (times longer)}$$

True, the fraction does not exactly equal 2, but that is due to "rounding error." The .67 and .33 obviously represent two-thirds and one-third. If we substituted fractions for decimals, the numerator would be 4/5 x 1/3 = 4/15. The denominator would be 2/3 x 1/5 = 2/15. Since 4/15 is exactly twice as large as 2/15, our conclusion is sustained.

So we have to double the length of our test in order to attain a reliability of .80. What does this mean? In objective tests, exactly what it says: you make up twice as many items of the same kind. But in the special case of essay tests, there are three possible interpretations, one of which is wrong, another correct but not feasible, and a third that is both feasible and more informative. If we simply doubled the time allowed for the essay but still got only one grade on it from each reader, it would have little, if any, effect on reliability. If we had each essay read by four instead of two readers, it would indeed increase the reliability of grades on this particular essay to .80, but it is hard enough to get two independent ratings of each essay, and few schools could afford the time or expense of giving each essay four independent ratings. Besides, the result would not indicate how consistent the students are in the quality of their writing from one topic to another, or from one time to another. The most fruitful interpretation, therefore, is "Do the same thing over again." Have the students write a second essay on a different topic, but one that requires the same mode of writing and is equally familiar to all students, and have a different pair of readers rate this essay independently. In my experience, having students write two short essays in the same session of an examination does not constitute two genuinely independent samples of their writing. They rarely differ more than the first and second pages of the same essay. There must be some separation in time as well as in topic before one can judge the average quality of a student's writing on different occasions. It has been the experience of many examiners in different colleges that the shortest possible separation in time for this purpose is to have one essay written in the morning and the other in the afternoon of the same day, and the examination schedule of most colleges does not permit any longer separation in time than this. That is why so many examining boards have adopted this policy

if they intend to attach any real weight to the essay grades. Of course, you will find examinations that require just one short essay, but in such cases the examiners rely on the objective sections to carry practically the whole burden of reliability.

If you have a director of testing, one procedure that I have recommended may worry him when the time comes to compute tetrachoric correlations. I said that whenever two grades on a test essay differ by more than one full grade-point (or more than 10 points if you use standard scores), refer the paper to the most experienced reader who has not already graded it for a third independent rating. A clerk will substitute this grade for the previous grade farthest from it (see page 20). Such revisions of discrepant grades necessarily increase correlations above the level that your director of testing expects when he correlates uncorrected grades, and he may cry "Foul!" Remember that correlations tell you how closely two sets of measures agree, and if you take all pairs of grades that disagree sharply and substitute a third grade that is closer to one or another, you automatically increase the correlation. But what else can you do? It would be stupid to correlate just the original grades, because the grades you discard have no effect on students' grades. What you want to compute is the reliability of students' grades, and for that purpose you have to correlate the two ratings that actually determine the grade. In any case, your director of testing or statistical consultant has little cause for complaint. He is used to getting correlations of .30 to .40 between sets of uncorrected grades, and they make the reliability so low that the essay grades are practically meaningless. If you discard about 10 percent of extremely aberrant grades and substitute genuinely independent grades of a more experienced reader, you will probably get tetrachorics in the neighborhood of .50, and they bring the reliability of students' grades on one essay up to .67. Then, as we have seen, all you have to do is to secure a second essay, graded in the same way, to attain a reliability of .80.

The reliability of *grading* is one thing, however; it shows how closely four readers agree in judging the merits of two essays. What it leaves out is the reliability of the *students*. To what extent do they tend to write as well on one topic as on another, and on different occasions? The only way to find out is to correlate the sum or average of their grades on the first essay with the sum or average of their grades on the second. This is computed in the same way as the reliability of the grading (as explained on page 33): find the percent who stood in the top quarter of final grades on both essays, look down to the corresponding tetrachoric, and below that to the reliability. In my experience, if the average reliability of the grading is .80, the stepped-up correlation between final grades on the two essays is likely to be about .70. That is the over-all reliability of the essay part of the examination, including both the variation in readers' judgments and the variation in quality of writing from one topic to another.

Although that final reliability of .70 is lower than I like, I do not know any examiner who consistently does better than this in any sort of essay examination that is administratively feasible—unless he adopts rules that artificially constrain the grading. Of course, in essay tests designed to measure information and understanding, as in history, one can do better, but not much better in tests designed to measure writing ability. If you need a reliability of .90 or better to determine the outcome of a controlled experiment, you will have to get eight or more test essays. Otherwise, that final reliability of .70 on the essay part of the examination can be offset by the higher reliability of the objective sections, as I shall now explain.

8

Computing the Reliability of Objective Tests

The last section should lodge forever in your memory the basic meaning of test reliability: the amount of agreement between two sets of independent measures of the same characteristic, taken at about the same time. You have more confidence in a test if you measure the same thing twice and get approximately the same result both times.

It was easy to see how to do this in the case of essays: correlate two sets of independent ratings of the same essays, or correlate grades on one essay with grades on another essay written by the same students.

But how do you do it in the case of objective tests: for example, a vocabulary test of sixty items? There you have only one measure—a single score for each student. How do you know how close you would come to getting the same scores for these students if you gave them another test of the same kind? There is not enough time in ordinary school testing to administer two comparable forms of every test.

Let me explain how professional testmakers do it in constructing such a vocabulary test, not because you want to learn how to construct vocabulary tests but because the laborious procedures they employ are the basis for the quick and easy formula for objective test reliability that I shall presently explain, and they will help you understand what it means.

The vocabulary testmaker usually wants to produce two comparable forms so that you can use one before instruction and one after, or one in

the regular examination and one in the make-up examination. Since he knows that many of the items he writes will be discarded after tryout because they are too hard, too easy, have either two right answers or no right answer, or have some other defect, he writes perhaps 200 items like the following:

exploit: A. go off with a loud noise
B. make use of for one's own benefit
C. run away and hide
D. throw away

Each tryout form has 100 items of this sort, which nearly all students can finish in 35 minutes or less. A good trick to remember in trying out a new test is to arrange the forms in each package in what testmakers call a "spiral" order so that the first student in each tryout class will get Form A, the next Form B, and so on. Thus both forms are administered simultaneously in each tryout class, but each student takes only one form. If there are as many as eight tryout classes (and there are usually more than this), one can be pretty sure that the average ability of students taking Form A is equal to the average ability of students taking Form B, since a random half of the students in each class took each form. That would not be the case if four classes took Form A and another four Form B.

From the results of the tryout, the testmaker discards items that are too hard, too easy, or defective and arranges the rest in order of difficulty. From this arrangement he selects items 1, 3, 5, 7, 9, etc. for Final Form A; items 2, 4, 6, 8, 10, etc. for Final Form B. They will probably not be arranged in order of difficulty in the published forms, because then students tend to give up as soon as the items get hard, but if they keep finding easy items interspersed with harder ones, they are more likely to finish the test. Hence the selected items are often rearranged in the alphabetical order of the words to be defined. Let us suppose that there are sixty items in each Final Form, and one can be reasonably confident that they are equal in difficulty. The testmaker also tries to make the two forms equal in discriminating power by using a figure called "biserial r" that is routinely computed for each item. It would take too long for present purposes to explain precisely what this means, but in general it answers the question: to what extent did high-scoring students on the total test do better on this particular item than low-scoring students?

The final step is to get as many teachers as possible in different schools to give both Final Forms to the same classes on successive days. Then the testmaker can compute the correlation between scores on the two Final Forms, since the same students took both. He does not "step up" this correlation by the Spearman-Brown formula because he does not expect any teacher thereafter to give both forms to the same students in one examination. The correlation between scores on Forms A and B is itself the reliability of either form. This is called "parallel form reliability," and it is the

most highly esteemed, especially if the testmaker reports a range of reliabilities for groups of class size. It clearly conforms to the definition of test reliability: the amount of agreement between two independent measures of the same characteristic, taken at about the same time.

Since teachers do not have time to apply this procedure to their own tests, but still ought to have some easier way to compute their reliability, it first occurred to someone that, if you have only one form, you can break it up into something like parallel forms by getting one score on odd-numbered items and another score on even-numbered items. The correlation between scores on these random halves is the reliability of the half-test and has to be "stepped up" by the Spearman-Brown formula to get the reliability of the whole test. This is called "split half" or "odd-even" reliability, and it is still widely used. It should not be used with speeded tests because students get the same score—0—on all items that they do not reach, and this spuriously increases the correlation between odd-even halves.

Next, Kuder and Richardson devised a long series of formulas that yielded almost the same results as the split-half method. Their Formula 20 is most often used today by large testing organizations like ETS to determine the reliability of their objective tests. The only trouble with it is that you have to know how many students answered each item correctly, and unless you have data-processing equipment, that takes more time than teachers can afford.

RELIABILITY

In his *Biographia Literaria* (Everyman Edition, p. 36), Coleridge pays this tribute to his friend, the poet and essayist Robert Southey:

"No less punctual in trifles, than steadfast in the performance of highest duties, he inflicts none of those small pains and discomforts which irregular men scatter about them, and which in the aggregate so often become formidable obstacles both to happiness and utility; while on the contrary he bestows all the pleasures and inspires all that ease of mind on those around him or connected with him, which perfect consistency; and (if such a word might be framed) absolute *reliability*, equally in small as in great concerns, cannot but inspire and bestow; when this too is softened without being weakened by kindness and gentleness."

According to the *Oxford English Dictionary*, this is the first recorded use of the term *reliability* (1816), even though it is regularly formed from *reliable* which goes much farther back (1569). The sense in which it is used by Coleridge, where it stands for consistency and stability, is not too far removed from the sense in which it is applied to test scores.

The mind-boggling sentence in which it appears is typical of Coleridge. What he means is, "You can always count on Southey. He's *reliable*."

Their Formula 21, however, is made to order for teachers. It takes only a few minutes to compute after you know the mean and standard deviation, which you ought to compute anyway for the purposes discussed earlier. If you have forgotten the short-cut formula for the standard deviation, it is given on page 26.

Here is a slightly simplified version of the Kuder-Richardson Formula 21, which yields a close approximation of the reliability of objective tests in which all items have equal weight: that is, each correct answer gets one point and each incorrect or omitted item gets 0. It must be applied only to raw scores on such tests, not to standard scores, percentiles, or numbers corresponding to letter grades.

$$\text{Reliability} = \text{ONE minus } \frac{\text{MEAN times (number of items minus the MEAN)}}{\text{Number of items times standard deviation squared}}$$

If you prefer symbols to formulas written out in words, it is:

$$r_{xx} = 1 - \frac{M(n-M)}{ns^2}, \text{ in which}$$

r_{xx} = reliability

M = MEAN

n = number of items (NOT number of students)

s^2 = standard deviation squared

Suppose that, on the vocabulary test of sixty items, the MEAN is 40 and the standard deviation 10. This becomes:

$$r_{xx} = 1 - \frac{40\,(60-40)}{60 \times 10^2}$$

$$= 1 - \frac{40 \times 20}{60 \times 100}$$

$$= 1 - \frac{800}{6{,}000}$$

$$= 1 - .133$$

$$= .867 \text{ (rounded to .87)}$$

The most common mistake in applying this formula to your own tests is to get so involved in manipulating the rather large numbers in the fraction that you forget to subtract the resulting decimal from ONE. What should alert you to the mistake is that the fraction usually turns out to be a relatively small number, like the .133 above. If that were the reliability, it would be terrible, but it is not: it is the error, the random variation, the UNreliability. The reliability is ONE minus this decimal, which is .87.

Although this reliability is quite high for an objective test that most students will finish in 20 minutes or less, one must not expect other objective

sections that are often included in English language arts examinations—reading comprehension, listening comprehension, and ability to detect errors in sentences—to do as well. Vocabulary is nearly always the most reliable objective section of any verbal test for two reasons: the items go so fast that one can get in a large number in minimal time, and they yield a large standard deviation, since people vary a great deal in the range and precision of their knowledge of words. One has to allow about one minute per item for reading and listening comprehension, and since there is rarely more than 30 minutes available for these tests, their reliability is likely to be in the sixties. Usage items (ability to detect errors in sentences) take about half a minute apiece; hence you can include forty in a 20-minute test, and its reliability is likely to be about .70.

The average reliability of these four objective tests—reading and listening comprehension, vocabulary and usage—may well be no higher than .70. Is that the reliability of the total objective part of the examination? By no means. The reliability of the total has to answer the question: if you gave comparable forms of these four tests to the same students tomorrow, how close would their *total scores* come to the *total scores* they got today? Therefore you must add together their raw scores on these four tests, find the mean and standard deviation of these total scores, and then apply the Kuder-Richardson Formula 21. Do not try to give extra weight to items that take longer and seem more important, or you can't use Formula 21. Anyway, the total number of correct answers in all four tests is an adequate basis for computing the reliability of the objective part of the examination. In the time usually available for the objective sections one can get in at least 160 items, and it is virtually impossible to attain a reliability lower than .90 for total scores on 160 objective items that are as highly correlated as these are likely to be.

In the last section (page 35) we concluded regretfully that the overall reliability of the essay part of the examination was unlikely to exceed .70, but it would be offset by the higher reliability of the objective part. Let us assume now that the reliability of these two parts turned out to be .70 and .90 respectively. How do we combine these to get the reliability of final grades on the examination as a whole, assuming that the essay and objective sections are to have equal weight? The chief statistician at ETS devised a formula for it, but it turned out that a simple arithmetical average of the two reliabilities "stepped up" by the Spearman-Brown formula gave the same result. The average of .70 and .90 is .80. Twice this correlation divided by one plus this correlation is 1.60 over 1.80, or 16/18, or 8/9, or .89. This is a conservative estimate of the reliability of final grades on the examination as a whole, and it is eminently satisfactory.

It is almost a pity that teachers usually insist on adding a course grade: their personal estimate of the amount and quality of work done during the course. There is no way known to mathematics to estimate the reliability of

that. Still, in the imprecise field of education a mathematically rigorous system of measurement may need a bit of looseness somewhere to make it comfortable to live with, and the course grade determined by each teacher may do just that.

9

Design for an Examination in English Language Arts

In the section on the reliability of essay grades, especially on pages 34-35, it was shown that, if you want to give real weight to the essays, you must secure two test essays from each student with some separation in time as well as in topic, and the least possible separation in time is to have one test essay written in the morning and the other in the afternoon of the same day. Many colleges have found that the only feasible way to get two essays per day in fields that use essay examinations is to schedule an examination period of one week at the end of each quarter or semester. In this week, one day is assigned to each field in which the examination requires a good deal of writing and half a day to fields in which the examination consists of objective or short-answer questions, like science and mathematics. Monday may be reserved for all courses in English language and literature. Tuesday for all courses in foreign languages and literature. Wednesday for history and social science. Thursday for mathematics and natural sciences. Friday for the fine and practical arts, and Saturday for vocational courses. There is usually a week following the examination period in which most students are on vacation, but make-up examinations are scheduled in the same order for the few who were absent or who want to improve their grade. Only by special permission are students allowed to take two courses in fields tested on the same day or half-day. They must take the make-up examination in one of these fields with the understanding that they waive the privilege of repeating that examination until the next time it is offered—at the end of the next quarter or semester.

A comprehensive examination in English language arts might be arranged as follows:

A. First objective section: maximum time, one hour
   Reading comprehension, 30 items, 30 minutes
   Vocabulary, 60 items, 20 minutes

B. First essay: maximum time, two hours (but most students finish and leave the examination room in 90 minutes or less)

LUNCH

C. Second objective section: maximum time, one hour
   Listening comprehension, 30 items, 30 minutes
   English usage, 40 items, 20 minutes

D. Second essay: maximum time, two hours (but again most students finish and leave in 90 minutes or less)

These time estimates are based on the time students usually take to finish such items or tasks when there is no pressure of time. The essays are scheduled at the end of the morning and afternoon sessions so that students may leave as soon as they have finished. They vary far more in the time they are able or willing to spend on their papers than in the time they take in answering objective items. Able and conscientious students usually spend more time than the average, especially in planning and revision, and we do not want to cut them off before they have completed the task to their satisfaction. There are always a few compulsive students, however, who keep hacking away at their papers long after everyone else has left the room. No matter how much time one allows, they always want more. At the end of the scheduled time, one has to take their papers away as gently as possible and shoo them out.

Combining Scores on Comprehensive Examinations

If the foregoing outline of a comprehensive examination is accepted as a workable model, we next face the problem of combining four grades on the two test essays and four numerical scores on the objective sections. Of course there will be variations in this outline to suit different courses of study, but the problem of combining grades and scores will remain.

As a first approach, let us assume that the composition staff has been using letter grades with plus and minus signs, and that they have agreed to aim at the distribution of grades predicted on page 30: 5% E, 15% D, 40% C, 25% B, and 15% A. To combine the four essay grades for each student, we have to translate the letter grades into numbers as follows:

E−	E	E+	D−	D	D+	C−	C	C+	B−	B	B+	A−	A	A+
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Adding four of these numbers for each student to get his total score on the essay part of the examination presents no problems, but then we have to combine these totals with much larger numbers representing scores on the four objective tests, and we want to give the essay and objective sections equal weight.

This problem is usually ignored in elementary textbooks on measurement, so there is no standard procedure, but the simplest and most satisfactory method I have found is to turn the raw scores on each objective test into letter grades in accordance with the predictions already applied to the essays. Thus, the top 15 percent of scores on each test get A's; the next 25 percent get B's; the next 40 percent get C's; the next 15 percent get D's; and the lowest 5 percent get E's. We may have to vary these proportions a bit when, for example, ten students get the same score at the lower boundary for B, but the predicted 25 percent will take in just three more students at this point. We can't give three of them, chosen at random, a B and the other seven a C if they all made the same score. In such cases, I go for whichever grade makes the smaller difference in the prediction. Since only three of the ten came within the B range, I would give all ten some variety of C (probably C+). If seven of the ten had come within the B range, I would have given all ten some variety of B (probably B−). After assigning letter grades to all four objective tests in this fashion, the staff translates them into the numbers corresonding to each grade in the table above that was used for the essay grades. Incidentally, this gives each objective test equal weight, even though the vocabulary test has twice as many items as the reading comprehension and listening comprehension tests.

In carrying out this procedure, someone is sure to object, "Those were the proportions for the various grades that we predicted for the essays. How can we apply them to the objective scores as well?"

This is natural, since at that point we were trying to set guidelines for the readers that would make the distribution of essay grades conform to reasonable expectations. But if you look carefully at what I said on page 30, you will see that I asked each teacher to predict how many students in each of his or her classes would make each grade *on the examination*. I said so repeatedly. Since they knew that the examination would include objective tests, it is reasonable to apply the same predictions to scores on these tests.

Now we have eight numbers for each student corresponding to letter grades on the essays and objective tests. We know that none of these components is highly reliable (except the vocabulary score) and some are not very highly correlated with others. The effect of averaging eight numbers of this sort is to shove everybody closer to the mean than we intended. If we just add the eight numbers and then divide by 8 to get each student's final (average) grade on the examination, it is virtually impossible for anyone to get a final average higher than 11, which means B, or lower than 5, which means D. We predicted that 15 percent would make A's, 25 percent B's, and so on, but even though we made the grades on each component come out that way for the group as a whole, the numbers each student gets depend so largely on chance that, if we take straight averages, no one will get an A;

about 15 percent will get B's; about 75 percent C's; about 10 percent D's; and no one will fail.

Any mathematician would have foreseen this result, but English teachers are not mathematicians, and their first reaction is always shock, incredulity, and dismay. Someone must have made a mistake in adding or averaging! No; the figures have all been checked, and they are accurate. Then some say that since these numbers represent our own judgments, we are morally obliged to abide by them. Others say no; the averages make no sense; and we must re-examine the papers and raise or lower enough grades to make final grades come out in the intended proportions.

Neither faction is right, and the solution is simpler than either one imagined. *Do not average those eight numbers.* Simply add them and make a distribution of total scores. Draw a line under the top 15 percent of these totals. Any student above that line gets an A; the next 25 percent get B's; the next 40 percent get C's; and so on. This is the first point at which those predictions make any real difference, and here above all we should abide by them.

The essay grades might just as well have been standard scores like those discussed earlier, based on the normal curve, with a mean of 30 and a standard deviation of 10, and with no attention at all to the presumed superiority of this group. Each reader could then divide the papers he received into top quarter, middle half, and bottom quarter; then pick out a fifth of the high papers as the very high, and a fifth of the low papers as the very low. The objective scores could be translated into standard scores in the same manner, or by actual calculation of standard scores. Once teachers get used to it, this is far easier than observing the predicted proportions for the various grades at every point. The result would be that each student would have eight numbers after his name representing standard scores, different from and larger than the numbers representing letter grades. But if one simply added the eight standard scores, made a distribution of the totals, and drew a line under the top 15 percent for a final grade of A, the chances are slight that any student who received an A from the first set of numbers would not also receive an A from the second. All we need to be sure about is that all eight numbers are on a common scale: either the standard score scale or the letter grade scale. It is only when we get the totals on either scale and make a distribution of these totals that we really need to think about our predictions—but then we should stick to them like glue.

Suppose the staff insists on giving some of the eight scores more weight than others. Suppose they decided that reading comprehension was the most important of the objective scores and should have a weight of 1.5; and that the score on error-detection was least important and should have a weight of .8. Very well; a clerk simply multiplies the numbers representing either standard scores or letter grades on reading comprehension by

1.5; then the numbers representing error-detection by .8. He adds the eight scores, some weighted in this fashion, for each student, draws a line under the top 15 percent for a grade of A, under the next 25 percent for a grade of B, and so on. At the end, it would be advisable to correlate the weighted with the unweighted totals. Over the years it has been found that weighting rarely makes any serious difference; students come out in nearly the same rank order regardless of weighting. Hence my advice would be to give all parts of the examination equal weight unless, for pedagogical reasons, you want to emphasize the importance of some part of the course by saying that it will get extra weight in the examination. It will probably make little if any difference in students' grades, but it may get them to work harder at something that they might otherwise neglect.

After the examination grade has been determined in the manner just explained, there is still the problem of combining the examination grade with the course grade, determined by each teacher. Here again simple averaging will push everybody closer to the mean than the staff intended, and again the remedy is the same: add the two numbers for each student corresponding to his examination grade and course grade. Make a distribution of these totals and award final grades of A to the top 15 percent, B to the next 25 percent, C to the next 40 percent, and so on. It is desirable to report all three: examination grade, course grade, and final grade. If any student or parent objects that the final grade is not precisely the average of the other two, explain that these are "adjusted averages."

A Note on the Significance of Differences

Since so few English teachers conduct controlled experiments, and those who do have statistical help, I shall not devote much attention to the significance of differences between the averages of groups taught with different materials or methods. But since books and articles on the teaching of English often state that the difference between the results achieved by Method A and Method B was not significant, or was significant at the .05, .01, or .001 level, I want you to have some notion of what it means.

The basic idea is that there is a good deal of chance (random) variation in all educational measures, and the amount of variation you would find in two out of three repetitions of the same measurement operation is called the "standard error" of that measure. This has nothing to do with mistakes, with bias, or with external conditions (such as an infernally hot day); it most commonly refers to chance variations from one sample of tasks or performance to another. For example, I said on page 28 that the standard error of essay grades was about 5 points on the standard score scale that I proposed (with a mean of 30 and a standard deviation of 10).

That is, if you had the same essay graded repeatedly by different competent readers and kept averaging the grades until you were certain what the true grade was, you would find that about two-thirds of the grades leading to this final average lay within one standard error (5 points) of the true grade, and 95 percent of them lay within two standard errors (10 points).

I shall say no more about the standard error of individual scores because they are so large that I find it the best policy to disregard them. But the standard error of the average of large groups—more than 100 students—generally used in educational experiments is much smaller: it is the standard deviation divided by the square root of the number of students. I mentioned that our hypothetical vocabulary test of sixty items (page 39) might have a standard deviation as large as 10. If it were given to 100 students, you would divide the standard deviation (10) by the square root of 100 (10), and so the standard error of the average of this group would be just 1 raw-score point.

I also said that the average score (mean) of my illustrative group was 40. Suppose that another group, treated in a different way, made an average score of 45 on this same test and also had a standard deviation of 10; hence a standard error of 1 point. Is that difference of 5 points between the two averages a true (significant) difference, or is it within the range of chance variation that one should expect in two administrations of the same test?

To find out, you have to compute the standard error of the *difference*. You square the standard error of the first average (1 x 1 = 1), square the standard error of the second average (1 x 1 = 1), add the two squares (1 + 1 = 2), and take the square root of the sum (2), which is 1.41, as you can find in any table of squares and square roots. Then the significance (reality) of the difference is judged against four standards:

1. If the difference (5 points) is less than twice as large as the standard error of that difference (1.41 x 2 = 2.82), it is not significant. This does not assert that it is, but that it could be, a chance variation. But since 5 is much larger than 2.82, it passes this first test.
2. If the difference is between 2 and 2.6 times as large as its standard error, it is significant at the .05 level, meaning that there are less than 5 chances in a hundred that a difference this large would be found if there were no true difference. But 2.6 times 1.41 is 3.67, and the difference of 5 is larger than this, so we can go on to the next level of significance.
3. If the difference is between 2.6 and 3 times as large as its standard error, it is significant at the .01 level, meaning that there is less than one chance in a hundred that it was a fluke. But 3 x 1.41 is 4.23, and 5 is larger than this, so we go on to the next level.

4. If the difference is more than 3 times as large as its standard error, it is significant at the .001 level, meaning that there is less than one chance in a thousand that it was a fluke. As we have just seen 3 x 1.41 = 4.23, and 5 is larger than this, so it is significant at the .001 level.

There are many different types of standard errors: of correlations, proportions, regressions, etc., each computed in a different way and yielding results of different orders of magnitude. There are also many different ways of computing the significance of differences between experimental groups: chi-square, analysis of variance and covariance, regression analysis, etc. Once you get into this statistical maze, you will never get out without help. But for most of the articles you will read that refer to the significance of a difference between two groups, the basic idea of "significance" is conveyed by the classical procedure that I have just explained: if the difference between the two averages is 2, 2.6, or 3 times as large as its own standard error, it is significant at the .05, .01, or .001 level respectively, referring to the chances in a hundred or a thousand that a difference this large would be found if there were no true difference.

One final point: "significant" does not necessarily mean "important"; it means only "non-chance." In statewide testing programs in which several hundred thousand students are involved, one group of 10,000 might be compared with another group of 10,000. To get the standard error of each average, you would have to divide the standard deviation by the square root of this number, which is 100. That would make the standard error so small that a difference of a tenth of a point might be significant, in the sense that it could not be attributed to chance, but it would have no educational or practical importance. Perhaps I should add that the .001 level does not mean that the difference was 10 times as large as at the .01 level; it only means that you are ten times as sure that there was *some* difference.

10

Initiating Staff Grading of Test Essays

It is no easy matter to introduce staff grading of unidentified test essays on the same topic in a staff that has four or more teachers of English. Teachers may be so sensitive to possible criticism of their results that they will not let anyone else even see the essays written by their students, let alone grade them. You may reassure them that no one will know which papers were written by their students because they will be identified only by numbers chosen at random by each student. Then they will want to know how anyone can possibly grade a paper fairly, not knowing the student. You may reply that in such examinations we are grading the writing, not the student, and that a final grade of D for one student may represent a triumph, while a final grade of B for another may represent a shattering disappointment. If we profess to be teaching composition, we ought to be able to tell which papers are better than others, regardless of who wrote them. Still, the argument goes on.

I see little hope of winning over such people by argument or persuasion. One has to introduce a series of experiences that will open their eyes to the extent of disagreement in the staff on the worth of selected papers that they all grade independently. I used to do this by getting one paper per month, each time from a different teacher, making typed copies with all identification, comments, corrections, and grades removed, and having each teacher grade it, comment on it, and return it to me at least one day before our next staff meeting. In that meeting I would write on the blackboard what grades the paper had received, and at the start I was pretty sure to get four or five different grades. The teachers were dismayed, but I tried not to be. I explained that such differences in grading standards always came to light whenever a staff began to study the reliability of its essay grades, and the only way to improve was to discuss our differences, examine the reasons behind them, and gradually develop standards that

would bring our grades closer together. I said it would be foolish to expect anything like perfect agreement in judgments of writing ability; all we could hope for would be the amount of agreement represented by a correlation of about .50 between grades assigned independently to each set of test essays by pairs of readers. Since that is the usual correlation between height and weight among adults of the same sex, it would still leave plenty of room for legitimate differences of opinion. But we were starting with a correlation of about .30 in our grades on this paper, and that was altogether too low to be fair to students.

I would then call upon some respected staff member to explain why he gave this paper an A. Next I would ask a friend of his to explain why he gave it a D or an E. Other teachers would express agreement or disagreement with these explanations and tell why they gave the paper some other grade. Thus we would move toward an elucidation of the grading problem presented by this paper and what policy we should adopt if we found such a paper in an examination. These discussions, which were amicable but spirited and often witty, proved to be more interesting than what we had previously done in staff meetings, and they gradually moved the staff toward acceptance of the idea that maybe more than one point of view should be represented in grading such important essays as those written in examinations.

We next tried out this idea in the least threatening case: each teacher chose one other teacher with whom he was willing to exchange papers on a topic that both had assigned to at least one class at the same level. Each graded the papers of both classes independently, and without writing anything on the papers. Then they compared their grades and resolved differences of more than one full grade by discussion. We learned the easy way to compute the correlation between the two sets of grades (before resolution of

John Stalnaker, long president of the Merit Scholarship Foundation, recalls this incident from his early days as Examiner in English at the University of Chicago.

In one of his experiments he had a few hundred papers to grade. He called in four of his most experienced readers and told them, "I want you to grade these papers but not on your regular scale of A to F. I know that you all have different ideas about what those letters mean. Just sort these papers into five piles in order of merit. Then mark the highest pile 4, the next pile 3, and so on down to 0."

They agreed to do so, but about a week later they came to his office and said, "We're sorry, John, but we could not do what you wanted. It turned out that there weren't any "4" papers. But we did the best we could. We sorted them into five piles, but we had to mark them 3, 2, 1, 0, and 00."

differences) that was explained on page 33. These figures gradually convinced us that a single essay, graded independently by two readers, was not enough to yield the reliability that we wanted in our examinations, so we gradually developed the type of examination outlined on pages 41-42, in which morning essays were graded by one pair of readers and afternoon essays by another pair. Later, as the staff gained experience with this method of grading, they decided that it would be a good idea to expose themselves to a wider range of viewpoints than that of their best friend in the department, so they let the department head assign sets of papers to pairs of readers that were either chosen at random or systematically rotated.

This is a shortened and simplified account of the development of the staff grading procedures I have recommended—with all the mistakes, setbacks, and wasted motion left out. Some of these procedures represent changes from those I suggested in earlier publications; more recent studies have changed my mind. Those that you adopt must be suited to your course of study, your student population, and the convictions and preferences of your staff. But one requirement is almost universal. At some point someone with authority—usually the principal or dean—must tell the staff to stop arguing and try something—no matter what. Without that push, nothing will happen.

Appendices

A Descriptions of Papers Rated High, Middle, and Low on Eight Qualities

Some readers may be disappointed that the procedures recommended for ascertaining and improving the reliability of essay grades all involved the cooperation of at least two teachers. What they probably hoped to learn was some way of rating papers that would improve the reliability of their own grades so that they could have greater confidence in their fairness and accuracy and could explain to students exactly why their grade was high or low. In other words, what they wanted was a list of things to look for in student compositions and how many points to give for this or take off for that.

A collection of readings offering suggestions of this sort was published by the National Council of Teachers of English, 1111 Kenyon Road, Urbana, Illinois 61801, in 1965: *A Guide for Evaluating Student Composition*, edited by Sister M. Judine, IHM. It is a paperbound volume of 162 pages, and sells for $2.75.

Although these papers contain much practical wisdom, I have never had much confidence in any scheme for rating papers that does not involve comparison with independent ratings of another person and discussion of papers on which there is a substantial difference of opinion. I have never seen any solid evidence in print that any of these schemes improves reliability.

If you want to use some sort of checklist to improve the consistency of your ratings, the only help I can offer is an example of the way in which guidelines for rating papers might be developed. After our factor analysis of judgments of writing ability, described on pages 5-10 of this booklet, we proceeded to a study of writing improvement in twelve school districts in the state of New York. All students in grades 9 and 10 who were involved in this study wrote one test paper per month on a topic set by us—the same topic for both grades. As indicated on page 11, these test essays were

written on paper that yielded three sharp, clean copies, two of which were sent back to different schools for rating on the following type of rating slip.

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	_____

Teachers encircled one number after the name of each quality to indicate their rating of the paper on that quality. At first the numbers all ran from 1 to 5, but since their courses concentrated on ideas and organization, they persuaded us to give double weight to those ratings by doubling the numbers representing each scale position. This weighting had no basis in research, but it seemed reasonable to give extra credit for the qualities these teachers wished to emphasize.

These eight qualities are short forms of the names of the five factors in judgments of writing ability revealed by our factor analysis, except that the mechanics factor is broken up into its logically distinguishable components—usage, punctuation, and spelling—and we added Remondino's factor (see page 9), here called "handwriting." At the right are spaces for subtotals of ratings on the first four factors, which we called "general merit," and on the last four, which we called "mechanics," and then a space for the sum of these two, the total rating. Note that, if a student gets the lowest possible rating on everything, his total will be 10; if all his ratings are in column 2, his total will be 20; and similar totals for the other three columns are 30, 40, and 50. These coincide with the standard scores of 10, 20, 30, 40, and 50 corresponding to letter grades of E, D, C, B, and A as explained on pages 27-28. Thus they were compatible with and led into the later use of standard scores; meanwhile they developed a clear idea of what the standard scores meant in terms of factors that make a difference in the grades of skilled readers.

Although these factors are represented on the rating slip only by short forms of their names, we developed an initial understanding of what they meant in all-day Saturday workshops that these teachers were paid to attend. We also gave them practice in rating sample sets of papers that had previously been rated on these eight qualities by expert readers. We kept

rating sets of these papers and discussing differences of opinion until a reasonable consensus was reached.

After the test papers had been rated in this fashion for one school year, heads of these departments met in a week-long workshop during the summer. Each brought a small sample of test papers on each topic that had been rated high (top quarter), middle (middle half), or low (bottom quarter), and that he or she regarded as typical papers at these levels of merit. We made photocopies of these papers and studied them together until we were able to agree upon brief descriptions of their salient characteristics. These descriptions were used throughout the following year as a guide in rating the monthly test papers, and particularly in training new teachers to rate papers on these qualities. At the end, the department heads met again and revised the descriptions, chiefly by cutting out parts that had been more confusing than helpful. The revised descriptions are reproduced in the following pages.

By the end of this study, we had come to look upon these guidelines as a training device that teachers may well use for a year or two to develop a common set of standards and a systematic way of thinking about the qualities that should enter into their judgment of a paper. After two years (at most) they move easily and naturally into the use of standard scores as a quicker and easier way to indicate their judgment of the general merit of a paper. We call this "rating on general impression," but it is no longer a blur: it is a quick summing up of characteristics that determine whether a paper is high, middle, or low in general merit. The teachers also have a common vocabulary for discussing the merits and defects of papers on which their grades disagree. They quickly recognize their agreement on perhaps six or seven of these eight qualities and "zero in" on the one or two that accounted for the discrepancy in their grades.

I. GENERAL MERIT

1. Ideas

High. The student has given some thought to the topic and writes 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 convey. 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 understand 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, if he had only stopped to think, he would have realized that they made no sense. He is only trying to get something down on paper. He does not explain his points; he only asserts 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, has a sense of movement, gets somewhere, and then stops. The paper has an underlying 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 standard and conventional. There is usually a one-paragraph introduction, three main points each treated in one paragraph, and a conclusion that often seems tacked on or forced. Some trivial points are 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 started to write. The paper seems to start in one direction, then another, then another, until the reader is lost.

3. 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 and 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 in a childish vocabulary may also get a low rating on this quality, even if no word is clearly wrong.

4. 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 not mistake this writing for the writing of anyone else. Although the writer may assume 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. Sometimes the unconscious revelation is so touching that we are tempted to rate it high on flavor, but it deserves a high rating only if the effect is intended.

II. MECHANICS

5. Usage, Sentence Structure

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

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

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

6. Punctuation, Capitals, Abbreviations, Numbers

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 clauses, around nonrestrictive clauses, or

between short coordinate clauses unless their omission leads to ambiguity or makes the sentence hard to read. Contractions are acceptable—often desirable.

Middle. There are several violations of rules that have been taught—as many as usually occur in the average paper. Counts of such errors in high, middle, and low papers at various ages and socioeconomic levels would be desirable in order to establish standards.

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

7. Spelling

High. Descriptions of spelling levels are most often used in grading test papers written in class. Since there is insufficient time to make full use of the dictionary, spelling standards should be more lenient than for papers written at home. The high paper (at ages 14-16) 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 high rating on spelling, even if no difficult words are used.

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. Spelling standards differ so sharply from grade to grade and from one socioeconomic level to another that each school would do well to make a distribution of spelling errors per hundred words (at least for test papers written in class) and relate its ratings to this distribution.

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. It may be excellent in other respects and still get a low rating on this quality.

B

Topics for Test Essays

If you have to set topics for test essays that will be written by the students of several teachers, you should have a way of securing ratings by these teachers of quite a long list of topics—preferably those that you or they have used and found appropriate for short, impromptu papers that can be planned, written, and revised in the time available and under the pressure of an examination. Unless these topics are selected from a list that teachers have approved, they almost always complain that the students would have written much better had it not been for the awful topic you gave them. Either it was too difficult and beyond their experience or it was so dull and hackneyed that no one could get interested in it.

The following topics are typical of those suggested by teachers for test essays. Most of them can be handled successfully by students in secondary schools (ages 12-17), but those near the end of the list seem more suitable for college students. Although I have no objection to your using any of these that seem interesting, I hold no brief for this particular list. I assume that you will compile a similar list of topics that you and the other teachers have found that your students can handle. Often the topics are suggested by papers that students have written on topics of their own choice. I make copies of such lists and hand them out to teachers at the first staff meeting of the year. I ask them to put a 2 before the topics they like best, a 1 before those that they accept, a 0 before those that they reject, and no mark before those about which they have no opinion. At the next meeting I hand out a shorter list of acceptable topics that received the highest ratings. It is understood that topics for all examinations concerned with writing ability will be taken from this list, but I try to keep the topic for any given examination a secret until the day of the test. Otherwise it sometimes happens that the less secure teachers give their students such broad hints about the nature of the topic that some write the essay beforehand, or get a friend to

write it, and commit it to memory. Other teachers may assign a topic that is almost like the one to be used in the test and then give detailed instructions on how to write such a paper. In one examination we found thirty-five papers that all started with the same topic sentence. If it is even suspected that some teachers are giving their classes more direct preparation for the examination than others, students will lose confidence in the fairness of the grades. Hence the only safe policy is secrecy. If the teachers keep their lists of approved topics, it is easy to pass the word just before the morning essay, "Topic 8." Then, if there is to be an afternoon essay, you wait until after lunch to announce "Topic 12." Since these topics are usually short, each teacher writes the selected topic on his blackboard. But if the topic is lengthy, and there is "stimulus material" on which students are to comment, the examination papers must be duplicated and handed out in sealed envelopes on the day of the test. Then it is understood that the seal may be broken only in the presence of the students who are ready to take the examination.

Here is the illustrative list of topics suggested by teachers:

1. I saw it happen
2. What I learned from experience
3. What I'll be doing ten years from now
4. If I could do it over
5. On being alone
6. My day in the palace
7. Flight to Planet X
8. Robbie the Robot
9. If an ancient Greek came to town
10. What happened when some machine went berserk
11. My idea of happiness
12. What scares me
13. My own standard of living
14. Were people happier in days gone by?
15. Some things do not change
16. The trouble with families
17. Mistakes parents make with children
18. Why teenagers rebel
19. Are teenagers conservative?
20. When should teenagers be treated as adults?
21. There's nobody like _____
22. Who should go to college?
23. My idea of an educated person
24. What I like about life in my country
25. What I dislike about life in my country

26. My country's contributions to mankind
27. In what ways are all men equal?
28. Is peaceful coexistence possible?
29. Can a world government prevent war?
30. What is the spirit of our time?

CHOOSING A SUBJECT

I was privileged to attend the last regular lecture at Harvard of the great teacher of the Bible, Kirsopp Lake. It was the day before the final examination, and I think he tried to ease the tension by telling this story.

"Gentlemen, I had a wonderful dream last night. I dreamed that I was sitting on a cloud at Judgment Day, watching all the tribes of earth assemble. They all came together in a great plain and sat down.

"Then, out of the circumambient mist, a great hand arose and began writing on a celestial blackboard in letters that all the world could read.

"It wrote out the Ten Commandments, and then—in typical examination fashion—it added: STUDENTS CHOOSE SIX."

C

Objective Items Based on a Central Theme

If this short course on grading essays written in examinations is widely used, other short courses will be written that will deal with the preparation, review, tryout, selection, scoring, and analysis of objective items far more extensively than we can do here. It seemed wise, however, to include a brief appendix on types of objective items that teachers of English will accept, since so many of them have a deep-seated prejudice against any use of objective tests. The previous discussion may have convinced you that short sections of objective items ought to be included in any final examination on English language arts for at least two reasons. First, the course is bound to include reading and listening comprehension, vocabulary, and grammar or usage, all of which can be tested more quickly, easily, and reliably by objective items than by written answers. Second, we have seen that the highest over-all reliability that American examiners can consistently attain in grades on essays written in final examinations is about .70, and this is too low to be entirely fair to students or to detect improvements in the course. It is most commonly raised to acceptable levels by scores on the objective sections, which yield much higher reliabilities per unit of testing time.

Still, teachers of English tend to regard these objective sections as, at best, a disagreeable necessity which can test only the most superficial aspects of proficiency in English. To help you convince your colleagues that objective tests need not be stupid, I should like to show you a test that I wrote some years ago and used in one of my examinations at the University of Chicago. Its distinctive characteristic is its unity. In almost all objective tests, no item has any connection with any other item, but here the whole test deals with a single problem of universal concern. The problem is discussed in three short passages that present contrasting points of view, and students must answer twenty items that test not only comprehension of

each passage but also an understanding of relationships between these passages. Next, there is a short but complete paper written by a student who was asked to compare the views expressed in these three passages and then state his own position on this issue. Note that the twenty items following this paper will deal with larger aspects of writing than mechanical errors. (A few examples of discrete items on ability to detect errors in sentences will be given later.) Finally, there is a writing assignment dealing with one important issue that is a part of the general problem discussed in the three passages.

The test as it stands is probably too hard for high school students. In fact, it was a bit too hard even for my college students. I chose a hard test as an illustration so that intelligent and well-prepared teachers of English would themselves get interested in it and find it hard to answer some of the questions. I think they will agree that, whatever else it may be, it is *not* superficial. It is intended only as an illustration of a possible format for objective tests of reading and writing that you and your colleagues may want to prepare for your own examinations, using easier material and simpler types of objective items. I have found it effective as what might be regarded as propaganda for some objective sections in tests of English language arts that rely chiefly on essays. Many fine teachers of English have said to me, "I have never had any use for objective tests, but I can't despise this one."

The Reading Test

Directions. Read all three passages before answering the questions that follow.

Passage I

The nation, with all its so-called internal improvements, which are all external and superficial, is just an unwieldy and overgrown establishment, cluttered with furniture and tripped up by its own traps, ruined by luxury and heedless expense, by want of calculation and a worthy aim; and the only cure for it is in a rigid economy, a stern and more than Spartan simplicity of life and elevation of purpose. It lives too fast. Men think it essential that the *Nation* have commerce, and talk through a telegraph, and ride thirty miles an hour, whether *they* do or not; but whether we should live like baboons or like men is a little uncertain. If we do not get out sleepers [large pieces of wood to which railroad tracks are nailed], and forge rails, and devote days and nights to the work, but go to tinkering

upon our lives to improve *them*, who will build railroads? And if railroads are not built, how shall we get to heaven in season? But if we stay at home and mind our business, who will want railroads? We do not ride on the railroad, it rides on us. Did you ever think what those sleepers are that underlie the railroad? Each one is a man, an Irishman or a Yankee man. The rails are laid on them, and they are covered with sand, and the cars run smoothly over them. They are sound sleepers, I assure you. And every few years a new lot is laid down and run over; so that, if some have the pleasure of riding on a rail, others have the misfortune to be ridden upon. And when they run over a man who is walking in his sleep and wake him up, they suddenly stop the cars and make a hue and cry about it, as if this were an exception. I am glad to know that it takes a gang of men for every five miles to keep the sleepers down and level in their beds, for this is a sign that they may sometime get up again.

Passage II

Myself when young did eagerly frequent
Doctor and Saint, and heard great argument
  About it and about: but evermore
Came out by the same door where in I went.

With them the seed of wisdom did I sow,
And with mine own hand wrought to make it grow;
  And this was all the harvest that I reaped—
"I came like water, and like wind I go."

Into this universe, the *why* not knowing
Nor *whence*, like water willy-nilly flowing:
  And out of it, as wind along the waste,
I know not *whither*, willy-nilly blowing.

Waste not your hour, nor in the vain pursuit
Of This and That endeavor and dispute;
  Better be jocund with the fruitful grape
Than sadden after none, or bitter, Fruit.

The moving finger writes; and, having writ,
Moves on: nor all your piety nor wit
  Shall lure it back to cancel half a line,
Nor all your tears wash out a word of it.

Passage III

No man can serve two masters: for either he will hate the one and love the other; or else he will hold to the one and despise the other. Ye cannot serve God and mammon.

Therefore I say unto you, Take no thought for your life, what ye shall eat, or what ye shall drink; nor yet for your body, what ye shall put on. Is not the life more than meat, and the body than raiment? Behold the fowls of the air: for they sow not, neither do they reap, nor gather into barns; yet your heavenly Father feedeth them. Are ye not much better than they?

Which of you by taking thought can add one cubit unto his stature?

And why take ye thought for raiment? Consider the lilies of the field, how they grow: they toil not, neither do they spin: and yet I say unto you that even Solomon in all his glory was not arrayed like one of these.

Wherefore, if God so clothe the grass of the field, which today is, and tomorrow is cast into the oven, shall he not much more clothe you, O ye of little faith? Therefore take no thought, saying, What shall we eat? or, What shall we drink? or, Wherewithal shall we be clothed? For after all these things do the Gentiles seek; for your heavenly Father knoweth that ye have need of these things. But seek ye first the kingdom of God and his righteousness; and all these things shall be added unto you.

Take therefore no thought for the morrow, for the morrow shall take thought for the things of itself. Sufficient unto the day is the evil thereof.

Directions continued. Mark the best answer to each question. Remember that no short answer to a question about a literary work can be completely correct. The best answers to the following questions need be only a little better than the other answers.

1. Which of the following questions is the central concern of all three passages?
   1. Is the pursuit of pleasure a desirable goal in life?
   2. Is hard work necessary for success in life?
   3. What should be our chief purpose in life?
   4. Is the pursuit of material values contrary to religion?

2. Which of the following best represents the goal stated in Passage I?
   1. The development of the Nation
   2. Simplicity and elevation of purpose
   3. To ride upon the railroad rather than to be ridden upon
   4. To keep the sleepers down and level in their beds

3. Which of the following stands for the opposite of the goal of Passage I?
   1. The Nation
   2. Spartan simplicity
   3. The sleepers
   4. Building railroads

4. Which of the following best represents the goal stated in Passage II?
   1. To sow the seeds of wisdom
   2. To come like water and to go like wind
   3. To be jocund with the fruitful grape
   4. To do whatever the moving finger writes

5. Which of the following stands for the opposite of the goal of Passage II?
   1. Doctor and Saint
   2. Sowing the seeds of wisdom
   3. Whatever the moving finger writes
   4. Endeavor and dispute over This and That

6. Which of the following best represents the goal stated in Passage III?
   1. The kingdom of God and his righteousness
   2. Sufficient unto the day is the evil thereof
   3. Take no thought for your life
   4. Refrain from any sort of labor

7. Which of the following stands for the opposite of the goal of Passage III?
   1. Mammon
   2. The morrow
   3. Food and clothing
   4. Hard work of any kind

8. Which of the following descriptions of man's role in life as conceived in these passages is LEAST accurate?
   1. Passage I: Man is a tool-using animal.
   2. Passage II: Man is a puppet of fate.
   3. Passage III: Man is a child of God.

   Remember: Which interpretation of each passage is LEAST accurate?

9. Which passage expresses concern over the exploitation of workmen in the pursuit of material values?
   1) Passage I 2) Passage II 3) Passage III 4) None of them

10. Which passage places its chief emphasis on *service to others*?
    1) Passage I 2) Passage II 3) Passage III 4) None of them

11. Which passage or passages regard *simplicity* as essential to a good life?
    1. All, about equally
    2. None of them
    3. Passages I and III
    4. Passage II

12. Which of these views is based on a conviction that there are no ar ers, that effort is futile?
    1) Passage I 2) Passage II 3) Passage III 4) None of them

13. Passages II and III both deny the value of "taking thought." How do they differ?
    1. II regards thought as unrewarding; III as a necessary evil.
    2. II refers to thought about philosophic issues; III to thought about making a living.
    3. II prefers action to thought; III prefers faith.
    4. II refers to thought about fate; III to thought about God.

14. All three passages seem to regard material possessions as unimportant. Which statement of their reasons for thinking so is LEAST accurate?
    1. Passage I: We should reduce our wants rather than increasing our means of satisfying them.
    2. Passage II: It is pleasanter to drink wine.
    3. Passage III: Striving for worldly goods interferes with the service of God.

    Remember: Which interpretation of each passage is LEAST accurate?

15. In which ways are the "sleepers" of Passage I like the "lilies" of Passage III?
    1. Both are subjects of parables.
    2. Both illustrate how men should act.
    3. Both illustrate what happens to people who concentrate on material things.
    4. Both illustrate the advantages of simplicity.

16. Which of the following pairs of passages are closest together in point of view?

    1) I and II 2) I and III 3) II and III

17. Which passage or passages emphasize the thought of the following quotation:

> The world is too much with us; late and soon,
> Getting and spending, we lay waste our powers.

    1) All of them 2) None of them 3) I and III 4) II

18. Which passage agrees with the thought of the following quotation:

> In the fell clutch of circumstance
> I have not winced nor cried aloud
> Under the bludgeonings of chance
> My head is bloody, but unbowed.

    1) Passage I 2) Passage II 3) Passage III 4) None of them

19. Which passage agrees with the thought of the following quotation:

Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think: every effort we can make to throw off our subjection will serve but to demonstrate and confirm it.

1) Passage I 2) Passage II 3) Passage III 4) None of them

20. Which passage agrees with the thought of the following quotation:

The great cry that arises from our manufacturing cities, louder than their furnace blast, is all in very deed for this, —that we manufacture everything there except men; we blanch cotton, and strengthen steel, and refine sugar, and shape pottery; but to brighten, to strengthen, to refine, or to form a single living spirit never enters into our estimate of advantages.

1) Passage I 2) Passage II 3) Passage III 4) None of them

The Writing Test

Directions. This student was asked to summarize and compare the views expressed in these three passages; then to state and defend his own position on this issue. His paper is reproduced here exactly as he wrote it except that each sentence is numbered. The questions that follow this paper deal with larger aspects of writing than correctness of expression; they call for the judgment of a critic rather than the skill of a proofreader. It would be wise to read the paper as a whole before starting to answer the questions, but you need not watch for errors in usage, punctuation, or spelling, since ability to detect such errors is not tested in this part of the examination.

(1) The three authors regard success in a job as unimportant because many in obtaining success use others as stepping stones. (2) Success is seeing the good in others and living a good life.

(3) Passage I considers any improvement in mechanical things as unnecessary and unsuccessful because thousands of people are often hurt in making the improvement. (4) Passage II says learning is important; it also says that if you're going to do anything, don't do something you'll regret, for what's done can't be undone. (5) Passage III stresses the point that you shouldn't struggle for material things; food and clothing are nothing compared to everlasting life. (6) All the authors agree that in success there is happiness, and there is no happiness in gains made crookedly.

(7) I believe success in work can't be the most important element in life but is very important. (8) Being successful in business doesn't necessarily mean that you're leading a good life. (9) Many successful people have reached their goal by robbing and cheating others. (10) Success in business often leads to conceit, and many successful people can't see the beauty in life for thinking only of themselves.

(11) Success in business is important in that it proves you can accomplish something. (12) It is a good thing if you reach your goal honestly and get happiness out of your success. (13) Many successful people aren't happy. (14) The real success in life is happiness and making others happy. (15) Many people are so busy rushing toward their goal that they haven't time to be happy. (16) I believe success in business is important if you don't let it obstruct your vision so that you can't see good in people, and it takes up all your time.

Questions on This Paper

1. In items 1-9 assume that the student's purpose is to show that *success in work is important*, provided that—and he mentions all of the following but one. Which one does he leave out?
   1. Provided that it is honestly attained
   2. Provided that it brings happiness and leaves time for other forms of happiness
   3. Provided that it makes a constructive contribution to the common welfare
   4. Provided that it does not inflate the ego and prevent seeing good in others

2. In the light of this purpose, his review of the passages is
   1. adequate, for he answers their objections to regarding success in work as important.
   2. adequate, for he points out that their only fundamental objection is to *dishonest* success in work.
   3. inadequate, for he includes only what is relevant to his purpose and leaves out many other points that could be made.
   4. inadequate, for he neither recognizes nor refutes important objections to his position that can be found in these passages.

3. In the light of this purpose, the opening sentence
   1. starts at a good point in reviewing the passages by showing their only serious objection to his own position.
   2. starts at a good point but immediately falls into a misinterpretation.

3. starts at a bad point; he should first point out what these passages say in favor of his position.
4. starts at a bad point; he should first tell what each passage said before pointing out any conclusion that they hold in common.

4. In the light of this purpose, sentence 14 is
1. the logical conclusion toward which his whole argument is directed.
2. one of the major reasons on which his conclusion is based.
3. only a restatement of his conclusion in slightly different terms.
4. irrelevant to and inconsistent with his conclusion.

5. The student tries to show that "success in work is important" by
1. first refuting the objections of the three passages and then building up his own case.
2. misrepresenting the arguments of the passages and then refuting them.
3. overlooking or misstating objections and then asserting and qualifying his view.
4. the propaganda devices of name-calling, begging the question, exaggeration, and reiteration without proof.

6. The student misinterprets at least one point in his summary of each passage, but everything he says about one passage is a misinterpretation. Which passage is that?
1) Passage I 2) Passage II 3) Passage III

7. At what point in the paper does the student's development of his own position begin?
1) Sentence 6 2) Sentence 7 3) Sentence 11 4) Sentence 14

8. There is one logical argument in support of the student's conclusion. In which of the following sentences is it stated?
1) Sentence 6 2) Sentence 7 3) Sentence 11 4) Sentence 14

9. Which of the following is the best comment on the student's arguments in support of his conclusion?
1. They are true as far as they go, but the argument is incomplete.
2. They are repetitions of his conclusion in different terms, not arguments to support it.
3. They sound plausible but commit many logical fallacies.
4. There are about twice as many statements opposed to his conclusion as there are in favor of it.

10. In sentence 1, "use others as stepping stones" was probably suggested by
    1. the remarks about the "sleepers" in Passage I.
    2. a misinterpretation of what Passage II means by "the moving finger."
    3. the reference to Solomon in Passage III.
    4. nothing that is stated or implied in any of the passages.

11. Sentence 2 is
    1. intended to summarize the positions of the three passages.
    2. intended to state the student's own position.
    3. intended to state a point on which the passages and the student agree.
    4. not clear as to which position is intended.

12. Sentences 7-10. This paragraph
    1. is a fair statement of the main point at issue.
    2. misses the point, which is whether even honest success in work is an essential element of a good life.
    3. misses the point, which is whether individual success makes for social progress.
    4. misses the point, because none of the passages mentions "conceit."

13. Sentence 11 is
    1. good, because it gives a reason for regarding success in work as important.
    2. good, because it answers the objections raised by Passage I.
    3. poor, because the last word, *something*, is vague.
    4. poor, because no one needs to be told why success in work is important.

14. Compare sentence 2 with sentence 11.
    1. The student is inconsistent in these sentences.
    2. The student is consistent because these sentences mean the same thing.
    3. The student is consistent if sentence 2 refers to views stated or implied in the passages while sentence 11 refers to the student's own position.
    4. Even so, the student is inconsistent because success is not the same thing as living a good life.

Directions continued. Items 15 to 20 are concerned with precision and accuracy of expression. Since we have already read and can refer to the passages that the student is trying to summarize, we can judge which answer to each of these items gives the most accurate interpretation.

15. Sentence 3: because thousands of
    1. people are often hurt
    2. workmen are injured
    3. investors are defrauded
    4. lives are used up

16. Sentence 4: Passage II says learning is
    1) important 2) vital 3) insufficient 4) useless

17. Sentence 4 (after the semicolon): it also says that
    1. if you're going to do anything, don't do something you'll regret, for what's done can't be undone.
    2. if you have to decide on a course of action, be very careful, because one mistake can ruin you.
    3. striving to accomplish anything is futile, because everything that happens is determined by fate.
    4. life should be devoted to pleasure, because it will end soon enough anyway.

18. Sentence 5: food and clothing are nothing compared to
    1. everlasting life.
    2. health and success in life.
    3. the birds and the lilies.
    4. the service of God.

19. Sentence 6: there is no happiness in
    1. gains made crookedly.
    2. ill-gotten gains.
    3. material wealth.
    4. the fruitful grape.

20. Sentence 16. Which of the following endings of this sentence comes closest to what the student probably meant?
    1. and it takes up all your time.
    2. and it does not take up all your time.
    3. and if you let it take up all your time.
    4. and if you don't let it take up all your time.

Answers

The three passages: 1-3 2-2 3-4 4-3 5-4 6-1 7-1 8-1 9-1 10-4 11-3 12-2 13-2 14-2 15-1 16-2 17-3 18-4 19-4 20-1

The student paper: 1-3 2-4 3-2 4-4 5-3 6-2 7-2 8-3 9-4 10-1 11-4 12-2 13-1 14-3 15-4 16-4 17-3 18-4 19-3 20-4

A Related Writing Assignment

Passage III is from the New Testament in the King James translation of the Bible, and it has always made thrifty Christians uncomfortable. The injunction that is hardest to take literally is "Take therefore no thought for the morrow." How can we reconcile this advice with the following passage from the Old Testament in the same translation of the Bible?

> Go to the ant, thou sluggard;
> Consider her ways, and be wise:
> Which, having no guide,
> Overseer, or ruler
> Provideth her meat in the summer
> And gathereth her food in the harvest.
>
> How long wilt thou sleep, O sluggard?
> When wilt thou arise out of thy sleep?
> Yet a little sleep, a little slumber,
> A little folding of the hands to sleep:
> So shall thy poverty come as a robber,
> And thy want as an armed man.

Write a paper in which you explain and, if possible, resolve the seeming contradiction between these two passages. You may approach this task in any way you like, but it may help you to get started if you consider the following suggestions. First, you might explain what the apparent contradiction is, and show the dilemma in which a devout believer is placed. Then you might write a careful explanation of what you think these passages mean, supporting your interpretation with relevant quotations. You might examine the case for the "ant," then the case for the "lilies," giving reasons for acting in accordance with each position, and then showing what difficulties an extreme adherence to either position would entail. Finally, you might try to work out a resolution of the conflict: either a way of reconciling the two positions or some middle ground between them that you would regard as a tenable position. Remember that both passages are *translations*, first published in 1611. The words are not those of the original writers; some expressions may have changed their meaning or connotations in the centuries that have gone by since this translation was made; and even as they stand, these passages may be interpreted in different ways. To show you how widely scholars differ in their interpretations of these texts, here is a recent, authoritative translation of the last paragraph in Passage III: "So do not worry about tomorrow: tomorrow will take care of itself. Each day has enough trouble of its own."

We hope this assignment will not offend either devout believers in the Bible or followers of the other great religions of the world. It is not our

purpose to show that the Bible offers contradictory advice. On the contrary, we believe that a careful interpretation of these passages will reveal no contradiction but only a difference in emphasis: a difference that exists among the followers of all religions.

You need not worry that a recent decision of the Supreme Court of the United States forbade compulsory reading of the Bible as a devotional exercise in public schools. The same decision explicitly permitted and even encouraged voluntary study of the Bible as literature, philosophy, or history. Here the purpose is literary: the interpretation and comparison of two passages of singular beauty.

A Comment on This Assignment

It is not necessary for the essay topic to be as closely related as this to the objective exercise, nor is this a common practice in college examinations. Indeed, if the objective sections consist of discrete items, unrelated to any central theme, as is usually the case, no such connection in thought is possible. But if you and your colleagues go to the trouble of preparing objective exercises on interpretation and criticism that are unified around a single topic or problem (along the lines of those you have just seen), you will naturally want the essay written in this session of the examination to deal with some aspect of the same theme. The students will be "warmed up" by answering questions on one or more passages dealing with this theme and on a student paper based on the passages. By that time they will have given a good deal of thought to the topic and will probably have generated some ideas of their own that they would like to express.

After all, it is somewhat unnatural and artificial to assemble a group of students on a given day and ask them all to write a paper about some unexpected topic that they may never have thought about before. We have to do it because, if we announce the topic several days in advance in order to give them time to study it and think about it, they may get varying amounts of help from their family or their friends. Keeping the topic a secret until the examination begins is the only way to make sure that each paper is the student's own unaided work. It is not wholly unreasonable, because this is not a test of creative writing; it is a test of ability to write something coherent and sensible on demand, as those of us who work in offices have to do every day. Still, it takes students some time to generate ideas about an unexpected topic; to discard those that, after consideration, seem irrelevant, inconsistent, or indefensible; and then to arrange the rest in a logical and effective order. It is no wonder that most of them do not write as well in this situation as they do on papers written at home, to which they have devoted a good deal of time and thought. Only after the

examination do many of them think of all the good things they ought to have written.

Whether or not this "warming up" makes enough difference to justify the time and work involved in preparing such unified examinations, the foregoing assignment illustrates the sort of extended assignment with a good deal of "stimulus material" that is often used in college examinations. You can see how much more food for thought it provides than the brief topics listed on pages 60-61.

D Discrete Types of Objective Items

Vocabulary. The most common type of vocabulary item was illustrated at the top of page 37: the word to be defined is underlined and is followed by a choice of three or four defining words and phrases. Every word in these definitions should be more familiar than the word to be defined. Both Edgar Dale and I, who have made extensive studies of the familiarity of English words to American students, have found that three-choice vocabulary items work as well as four-choice. The greater element of chance in the three-choice item is offset by the larger number of responses one can get per unit of time. Some teachers have the idea that all the choices must be single words. Such a restriction is pointless; I prefer several words as in the definition of *exploit* on page 37: "make use of for one's own benefit." Here are some other common types of objective vocabulary items:

Completions. Which pair of words best fits the meaning of this sentence?
From the start, the islanders, despite an outward _____ _____, did what they could to _____ the ruthless occupying power.
1. harmony, assist 2. enmity, embarrass 3. resistance, destroy 4. acquiescence, thwart

Opposites. Which of these is the opposite of the italicized word?
chronic: 1. slight 2. temporary 3. wholesome 4. patient

Analogies. Which pair of words is related in the same way as *trigger: bullet*?
1. handle: drawer 2. holster: gun 3. bulb: light 4. switch: current

Right-wrong sentences. Mark each sentence R (right) if the italicized word is used correctly; W (wrong) if it is used incorrectly.

I *adjure* you to treat the matter confidentially. (R)

A barely *culpable* heartbeat showed that the victim was still alive. (W)

Listening comprehension. Listening comprehension passages and items are prepared in the same way as reading comprehension passages and items, and we have seen plenty of examples of the latter on pages 63-68. The main differences are that the passages (which are read aloud by the teacher) should be material of a sort that is normally listened to rather than read: stories, conversations, lectures, directions, short and relatively simple poems, etc. The test booklets that students mark have only the four answers to each question, but not the questions themselves, which are read aloud by the teacher. For example, the first story in a test of this sort is about an eagle and a fox. The first item in the test booklet has only this: 1) Looking for food. 2) Sleeping on a rock. 3) Trying to hide from the eagle. 4) Drinking from the stream. These make no sense until the teacher finishes the story and reads the first question: What was the fox cub doing when the eagle saw it? Then it is clear that the correct answer is 2) Sleeping on a rock. This device keeps the students from marking their answers during the reading of the passages.

English usage, sentence structure, and punctuation. There are innumerable ways of testing students' knowledge of the rules and conventions of a language and no clear-cut superiority of one way over another in terms of correlations with carefully determined grades on samples of the students' own writing. I used to use student papers with a large number of errors, including some that I inserted myself. These were printed in the left-hand column of a divided page with certain portions underlined or enclosed in brackets. Opposite each marked portion were from two to four ways of writing, arranging, or punctuating it, always starting with the one that appeared in the left-hand column. To keep students from assuming that this first choice was always wrong, I would sometimes put the best choice on the left side and transfer what the student had written to the right-hand column as one of the choices. Sometimes the intended answer was to transfer that part of the sentence to some other place; sometimes it was to omit that part entirely. Although this was a realistic way of testing correctness of expression, since it virtually duplicated the act of proofreading, I was never able to prove that it yielded results that were superior to those of other item-types that were easier to prepare and assemble. By using actual student writing, I was stuck with whatever errors a particular student happened to make, plus a few that I inserted, and these might or might not reflect the weaknesses of the class or the rules we had been studying.

I therefore abandoned this effort at realism in testing and substituted discrete items in which no sentence had any connection in thought with any other sentence. I had a long list of the most common errors in the writing of American students that persist through the freshman year in college (age 18). I embodied each error in a sentence and broke up the sentence into three lines of about equal length, making sure that the whole error lay within one of the three lines. The directions were simply to mark the line that contained an error or 0 if there was no error. One can test the ability to detect almost any type of error in usage, word choices, sentence structure, and punctuation in this format. At first I included spelling errors, but even good students and teachers tended to overlook them in this type of test; they were looking for bigger game. Hence I cut out the spelling errors and made separate spelling tests of 100 words each, about half spelled correctly and the rest incorrectly, to be marked R (right) or W (wrong).

Here are just a few examples of the three-line sentence item-type:

1. She asked whether	1. His last address
2. we would be ready	2. was seventy-four
3. to leave by noon?	3. Poe Lane, Albany.
1. Last Saturday Chester and	1. "Please don't do
2. Bud went fishing and	2. that", said Mary
3. brought back ten of them.	3. to her sister.
1. She is one of those rare	1. If I had known that the
2. women who never cares about	2. assignment was important,
3. wearing stylish clothes.	3. I would of done it quickly.

It is obvious that such items are easy to write, assemble, reproduce, and score. They approximate the act of proofreading one's own work, since there are no marked portions drawing attention to possible errors, and one is not told what kinds of errors to look for; one has to be ready for anything. Such items do not test the ability to correct such errors or avoid them in one's own writing, but students who are good at detecting them tend also to be good at correcting and avoiding them. If my memory is correct, this item-type was first suggested by S. Donald Melville when he was the director of the Cooperative Test Division of ETS. It makes the work of preparing objective tests of English usage a great deal easier than any other item-type I have used for this purpose, and it works as well as any other.

Common Errors in Usage and Sentence Structure

In a tryout of 580 items of the three-line sentence type in secondary schools, I found that the items most frequently missed (marked incorrectly) could be classified under the following 20 headings. When the name of the error is universally understood by teachers of English, I give only the name; otherwise I give a brief statement of the rule that was violated, sometimes with a warning that modern linguists and editors accept certain constructions that were formerly regarded as errors.

1. Sentence fragment, incomplete sentence (if unintentional)
2. Comma splice, fused sentence (main clauses joined only by a comma without a conjunction, or by nothing at all)
3. Run-on or strung-together sentences (more than two main clauses unless they are short, of the same pattern, or separated by semicolons)
4. Carelessly omitted words or parts of words, especially endings
5. Careless or needless repetition
6. Adjective for adverb and vice versa
7. Confusion of subject and object forms of six pronouns, *I, we, he, she, they, who*. Many linguists accept *who* as an object form, especially in questions, but *whom* is not accepted as a subject form.
8. *Shall-will, should-would*. The rules governing these word choices are so complex and so rarely mastered that some linguists advise using *will* and *would* regularly; *should* only in the sense of *ought to*. In current American speech, *will* occurs 217 times for every *shall*; *would* nine times for every *should*. British usage differs from American on this point and uses *shall* and *should* more frequently.
9. Subject-verb agreement, especially after *there* and after a compound subject joined by *and* or *or*. Speakers of some American dialects often omit final -s in writing because they neither hear it nor pronounce it.
10. Indefinites such as *anyone, anybody, someone, everybody, each, either, neither* and *none* take a singular verb and following pronoun if the meaning permits; but *none* and *neither* are often plural, and sometimes both singular and plural follow, as in "Everybody was there, but they have gone home," and "If anyone calls, tell them to call back."
11. Pronoun-antecedent agreement: two antecedents with *and* usually require the plural; with *or* the pronoun agrees with the nearer antecedent.
12. Pronoun reference: what a pronoun refers to should be clear from the sentence structure, meaning, or context; but *it, this, that,* and *which* may refer to the whole preceding clause if no ambiguity results.
13. Tense: wtong form, improper sequence, needless shift.

14. Parallel structure: sentence elements having the same function should, if possible, be parallel in form (e.g., not a clause, a gerund, an infinitive, and a noun as members of the same series).
15. Misplaced modifiers (especially dangling participles and *only*) should be counted as errors only if they appear to modify something they cannot logically modify, often with ludicrous effect.
16. Abbreviations: the safest rule is to avoid abbreviations in sentences except *Mr.*, *Mrs.*, *Ms.*, *Dr.*, *St.* (*Saint*), *a.m.*, and *p.m.*; *Hon.* and *Rev.* may be used only when the first name, initials, *Mr.* or *Dr.* precedes the surname.
17. Contractions (such as *don't*) are permissible in anything less formal than a dissertation, but some students have to be cautioned against excessive use of them.
18. Possessives: omitted or misplaced apostrophe: *her's*, *it's*, *your's*, *their's*, and *who's* are incorrect. There has been a long controversy over whether the possessive should be used before *-ing* forms, but our editors now tend to accept either "I'm surprised at his saying that" or "I'm surprised at him saying that."
19. Numbers: some publications now use figures for even small numbers like 2 or 3, but most prefer writing out numbers in sentences unless more than two words are required, unless several numbers occur in the same sentence, and unless they are pages or divisions of a book, street numbers, dates, and time of day if followed by a.m. or p.m. Numbers like $10 million are now common. A number beginning a sentence must be written out.
20. Capitals: although usage varies, we generally capitalize names of persons, places, languages, organizations, days, months, holidays; historical periods, events, or documents; titles before names; first word and all others except articles, prepositions, and conjunctions in titles of publications and papers written by students (but not always in bibliographic entries); first word in every line of poetry (or as printed); first word in every sentence including quotations and inserted statements.

The remaining types of three-line sentence items that gave American students the most trouble were connected with the use of the following punctuation marks: comma, dash, semicolon, colon, question mark, apostrophe, ellipses, quotation marks, and breaks within quotations. I omit all errors in word choices, since there are too many to classify.

My final word of advice on such tests is not to despise them. Objective test items can easily, quickly and reliably test a student's knowledge of the rules and conventions of English. In writing, if a student is not sure that he knows how to use a certain construction, he can change his sentence to

avoid using it, but in an objective test, you can give him a sentence with that construction in it, and he has to decide whether it is correct or incorrect. Arguments over whether answering objective items "is the same thing as" actual writing or speaking are futile. For that matter, writing one essay is not "the same thing as" writing another essay, even on the same day; we have seen that even the most carefully determined grades on such essays rarely correlate higher than .70. Then the really astonishing thing is that scores on a good objective test of English usage often correlate about .70 with averages of the two essay grades. It does not matter that they do not "really" measure "the same thing." If students who are good at one also tend to be good at the other, and vice versa, then it is a good indicator of proficiency in written English. Call it an editing test if you like, but I can promise you that students who do well on it also *tend to be* good writers.

A Short Test of Knowledge of Grammar

Although I have frequently inveighed against the teaching of English grammar, since most students refuse to learn it, and research in several countries over a long period of time has shown little, if any, connection between any type of grammar, traditional or modern, and improvement in writing, I have to admit that most teachers of composition devote an inordinate amount of time to it. I often suspect that they run away from the problem of teaching writing and teach grammar instead. Wondering how this time could be shortened, I wrote out the rules governing standard usage in the most common types of errors (described in the last section) and counted the number of technical grammatical terms that I had to use in stating them. I found that I could get by with forty, which I arranged in five groups as follows:

1. active, passive, linking; subject, verb, object, complement; helping verb
2. phrase, clause (independent, subordinate, coordinate); simple, compound, complex
3. noun, pronoun, adjective, adverb, preposition, conjunction, article, interjection
4. singular, plural, possessive; tense, perfect; modify, agree, apposition
5. number, case, person; infinitive, participle, gerund; conditional, parenthetical

Some linguists insist that there are only four parts of speech, but they treat pronouns as a subclass of nouns; they begin talking about prepositions when they get to phrases, and conjunctions when they get to clauses; and they call articles "determiners," but I can see no advantage over the familiar term. I included *interjection* only because I had to use it in the rule about setting it off with a comma or exclamation point.

Many linguists treat the passive as a transformation, but in my experience young students do not grasp it unless it is included in the list of basic sentence patterns. The term "transitive," however, seems to me to make more trouble than it is worth, and I doubt that young students need to distinguish direct and indirect objects. When a sentence contains both, I describe it as subject verb object object.

Some teachers may want to add a few terms to my list, but I doubt that anyone would really need more than fifty. The quickest way I know to find out whether students can use such terms in describing a sentence is illustrated by the following test.

The "Shadow" Test

Directions: Encircle the number of the best answer to each question.

The test is based on *one* sentence:

I have a little shadow that goes in and out with me and what can be the use of him is more than I can see.

1. This sentence may be hard to read because one comma has been left out. Where would you put a comma to break up the sentence into two main parts?

1. After *shadow*
2. After *me*
3. After *him*
4. After *more*

2. What kind of sentence is this?

1. Simple
2. Complex
3. Compound
4. Compound-complex

3. What is *I have a little shadow*?

1. The subject of the sentence
2. The first independent clause
3. The first subordinate clause
4. The subject of *him* (line 3)

4. What is *that goes in and out with me*?

1. The first independent clause
2. A subordinate clause, object of *have*
3. A subordinate clause modifying *shadow*
4. A subordinate clause modifying *goes*

5. What is *and*?

1. A coordinating conjunction
2. A subordinating conjunction
3. A relative pronoun
4. A preposition modifying *what*

6. What is *and what can be the use of him*?

1. The second independent clause
2. A subordinate clause modifying *shadow*
3. A subordinate clause, subject of *is*
4. A subordinate clause, subject of *see*

7. What is *than I can see*?

1. The second independent clause
2. A subordinate clause, object of *is*
3. A subordinate clause, object of *more*
4. A subordinate clause modifying *more*

8. What is *is*?

1. Verb of second independent clause
2. Verb of second subordinate clause
3. Verb modifying *more*
4. A verb that does not have a subject

9. What is *more*?

1. A coordinating conjunction
2. A subordinating conjunction
3. An adverb modifying *than I can see*
4. A linking-verb complement

10. What is the subject of the first independent clause?

1. *I*
2. *shadow*
3. *I have a little shadow*
4. *that goes in and out with me*

11. What is the subject of the second independent clause?

1. *shadow*
2. *that goes in and out with me*
3. *what can be the use of him*
4. *more than I can see*

12. How many subordinate clauses are there in this sentence?

1. One
2. Two
3. Three
4. Four

13. What is the subject of the first subordinate clause?

1. *shadow*
2. *that*
3. *what*
4. *more*

14. What is the subject of the second subordinate clause?

1. *what*
2. *use*
3. *him*
4. *more*

15. What is the subject of the third subordinate clause?

1. There is no third subordinate clause.
2. *what*
3. *use*
4. *I*

16. What is the verb of the first independent clause?

1. *have*
2. *goes*
3. *can be*
4. *can see*

17. What is the verb of the second independent clause?

1. *goes*
2. *can be*
3. *is*
4. *can see*

18. What is *shadow*?

1. Subject of the whole sentence
2. Object of *have*
3. A linking verb complement
4. Object of the preposition *little*

19. What are *in* and *out*?

1. Prepositions
2. Adverbs
3. Objects of *goes*
4. Adjectives modifying *with me*

20. What does *with me* modify?

1. *shadow*
2. *have*
3. *goes*
4. *in and out*

21. What is *what*?

1. A relative pronoun
2. An interrogative pronoun
3. An indefinite pronoun
4. A personal pronoun

22. What is *of him*?

1. Object of the verb *use*
2. Prepositional phrase modifying *use*
3. Prepositional phrase, subject of *is more*
4. Prepositional phrase modifying *can be*

23. What is *than*?

1. A coordinating conjunction
2. A subordinating conjunction
3. An adverb modifying *can see*
4. A relative pronoun, object of *can see*

24. *Can be* is a different form of the same verb as

1. *have*.
2. *goes*.
3. *is*.
4. *can see*.

25. What is *can* in *can be* and *can see*?

1. An adverb
2. An auxiliary
3. The subject
4. The object

26. The subordinate clauses in this sentence have *three* of the following functions. Which one do they *not* have?

1. Noun
2. Verb
3. Adjective
4. Adverb

Here is the sentence again: I have a little shadow that goes in and out with me and what can be the use of him is more than I can see.

Rewrite this sentence in as many of the following ways as you can. Use the same words that are in this sentence but change the form and order of these words as required. Try not to change or omit any of the ideas expressed by this sentence. Each rewritten version should be a single complete sentence.

27. Start with *I had a little shadow.*

28. Start with *I cannot see the use.*

29. Start with *The children had.*

30. Start with *Do you have.*

31. Start with *What can be the use.*

32. Start with *Going in and out with me.*

33. Start with *More than I can see.*

34. Start with *Go in and out.*

E

Learning to Write

I do not want to end this booklet with treatments of mechanical errors and grammatical terms, because teachers devote altogether too much time to them already. To give a broader view of what students need to learn about writing—at least by the end of the freshman year in college—I have decided to conclude with a list of ninety-six things that I have tried to teach in one way or another: by direct instruction, by comments on papers, and in conferences with students. They may be regarded as an extended list of objectives, but I wanted my students to read it so that there would be no mystery about what I intended to teach. Hence I could not use the maddening repetition of "Ability to . . . Ability to . . . Ability to . . .," nor the form of statement advocated by Magers and others: "Given a set of twenty sentences, students will indicate which ones contain colorful words or expressions with not more than four errors." Even teachers would refuse to read ninety-six statements of that sort. I therefore decided to state my goals in the form of advice to students on learning to write, with as much variety of statement as possible. I began with the following paragraph to show that I did not expect all students to follow all these injunctions all of the time:

"No general statement about writing, including this one, is 100 percent true. The following statements are probably true of 10 to 90 percent of good writing. They are no less useful because they are not universally true. What even 10 percent of good writers do most of the time, or what all good writers do even 10 percent of the time, is likely to be suggestive and helpful."

A. The Writer

1. Students should form a definite and serious intention of becoming good writers, fully realizing the difficulty, the feasibility, and the value of the enterprise. They should not take this intention for granted. They should consider the question seriously and at length, make up their minds deliberately, and mark their resolution by some outward act. It may be necessary to start from a conviction of sin: an awareness of the limitations of their present writing, and a deep concern about it. At the other end of the scale they should recognize excellent writing when they see it and wish to emulate it.

2. Students should feel a glow of exultation when they have written a good phrase, sentence, paragraph, or paper. They should care enough about the quality of their writing to spend the time necessary to do a good job. They should realize that practiced writers will gladly spend an hour or more over each page.

3. Writers must be willing to throw away hard-written paragraphs or pages, even though they are clever, once it becomes clear that they do not belong. They must cultivate the art of waste-basketry.

4. When students have to write something, they should set about it promptly, with confidence that they can do it well. They should not postpone the task indefinitely because they feel that they "can't write."

5. The first step in writing is to think about the problem or topic—not to begin writing anything that comes to mind, not to search through books for an idea, and not to run away from the problem and write about something else. Fifteen minutes of honest thinking about any problem will usually yield some idea about it that is worth writing down. The way to interest people is to have an idea.

6. The ideas about a problem or topic that occur to one in the process of thinking about it are the only things worth writing down—not what someone else has said about it, what people usually say about it, or what you think the teacher would like you to say about it. Information about a topic should never be used in place of an idea; it should be used only to support or illustrate an idea. Students should not be dismayed if the ideas that occur to them do not solve the whole problem, and they should not expect to present very many or very important ideas. One small idea per paper is above average.

7. Students should be cautious about transferring to a new problem the thinking they have done about a previous problem. It is well to see relationships, but not to save wear and tear on the brain tissue by using an old idea over again. In too many cases the old idea does not really fit the new problem.

8. Writing should give assurance that the writer is capable of looking a fact in the face, of taking a definite stand, of telling the truth rather than what he thinks people will like. It should not leave the impression that the writer wants above everything else to avoid trouble—even at the cost of saying nothing.

9. The most tiresome writing in the world is that which tries to protect itself from every possible attack by putting in every possible exception, qualification, and condition. It is like the aged spinster who still looks under the bed—but no man is sufficiently interested to hide there.

10. Writing should cut through the obvious, conventional, easy thing to say to the real issues underneath: to true feeling, fresh perception, independent thinking, on however humble a level. Pretentious writing is the most likely to miss this quality. Writing should mean something, not just mouth words.

11. Writers should be willing to reveal themselves, not as they would like to be, but as they are, confident that qualified readers will understand and be interested. The model to imitate is the honest candor of a conversation between friends.

B. The Whole Paper

12. A paper ought to have a plan that will be apparent to the discerning reader.

13. A paper ought to have one central purpose, point, or idea, which we shall refer to hereafter as the "theme." The student should consider very carefully what he wants to accomplish: what impression or conclusion he wishes to leave with the reader. In the beginning he should practice formulating his central point or purpose in a single sentence and writing it down.

14. The title should be related to the theme. It should delimit the field of the paper as sharply as possible without sacrificing other desiderata. It should be brief, and the words chosen should be in keeping with the tone of the paper. If the subject warrants it, the title may be arresting—but young writers strain too hard to make it arresting.

15. Apart from the introduction and conclusion, there should rarely be more than three or four main divisions in the short papers that students write. The student should list the points he wants to cover, eliminate those that are not essential to the theme, and group the rest under not more than three or four main headings. He should mark the points he wants to emphasize and consider what point will furnish the best entrance into his subject.

16. Each main point should be clearly related to the theme, and should reveal the way in which it is related: e.g., as illustration, proof, application, etc.

17. The points in a paper should be arranged in the order that fits best (a) the purpose in writing, (b) the logical requirements of the subject, and (c) the requirements of the audience—what they already know, what they will accept without question, what they will oppose, criticize, or misunderstand, and what will move them most powerfully.

18. There should be a clearly marked beginning and ending.

19. The beginning should (a) be clearly related to the theme, (b) catch the reader's interest, (c) show that the topic deserves consideration; that it is interesting, important, or timely, and (d) state or suggest the purpose, scope, and general method of organization.

20. The paper may begin with a direct reference to the title (never with "this" or "it" intended vaguely to refer to the title), with a statement or quotation bearing on the subject, with a pertinent narrative, with background information, with an explanation of the timeliness or importance of the topic, or in other ways too numerous to mention. One writer suggests: "A paper that begins on a moralizing tone will never come to anything."

21. The paper should stick to the scheme of organization stated or implied in the beginning, or to the underlying pattern of organization, even when it is not indicated in advance. A paper should not start out as one thing and then turn into something else—except for good reason, and with appropriate indications of the shift. A combination of two types of organization, however, is not necessarily inconsistent: cause and effect, for example, frequently requires a chronological organization as well.

22. Some of the common methods of organization are by time, space, cause and effect, familiar to unfamiliar, classification, division, definition, comparison and contrast, analogy, the order of impressions, the order of climax, etc. These are *not* the only possible types of organization. They rarely exist in a pure form; most actual schemes of organization could be described only in terms of two or more of these headings.

23. Students should be able to organize the same material in different ways to suit different purposes, occasions, or audiences.

24. Within a chronological organization it should be noted that usually events cannot be related in a strict time sequence without confusing two or more trains of events. One train should be followed to a convenient break in the narrative before starting another.

25. A story should be told from a consistent "point of view," using only events that could have been observed from that point of view. If other events are necessary to the story, the observer should have some plausible way of learning about them.

26. A long paper may need to be enlivened by changes of pace: e.g., by examining some parts slowly and analytically, then quickly sketching out several others that present no new problems, etc.

27. Students should be able to clarify and illuminate an abstract discussion by the use of analogy without relying upon it as proof.

28. Students should be able to write an accurate literal definition, without circularity, and to expand the meaning of a key term or concept by an extended definition, developed by classification, function, distinctions, historical causation, etc.

29. The most important parts of the paper should be treated at greatest length or with the greatest emphasis—by position, choice of words, or manner of statement. If necessary, one may say directly—in so many words—that a given part is important. The other parts should be treated in proportion to their importance, difficulty, or interest.

30. The ending should (a) if necessary, recall the chief points that have been made, (b) state or suggest the conclusion that has been reached, the resolution of the conflict or problem, (c) (possibly) show some application of this conclusion, suggest next steps, etc., (d) point up or heighten the emotional and imaginative significance of what has been said, (e) show what has been said as one thing, even though it has been presented in related pieces. Sentimental and moralistic endings should be avoided.

C. Paragraphs

31. Paragraphs should be distinct, each dealing with a clearly separable phase of the theme, and unified, with every sentence clearly related to the topic sentence or central idea.

32. Paragraphs should be joined by smooth transitions that indicate or reflect the relationship of the paragraphs to the central theme and to one another.

33. Transitions may be made by connectives, by direct statements of relationship, by repetition of key terms, and, above all, by a close connection in thought.

34. A paragraph should have a beginning and an end, and should move in an orderly fashion between the two. The sentences should go from one consideration to another like a train of thought. The discerning reader should be able to see the connection between each sentence and the one preceding. He should never feel that a sentence should have occupied some other position in the paragraph.

35. The topic sentence or central idea of a paragraph may be developed by definition or explanation of terms, by distinguishing it from some other idea with which it may be confused, by repetition with variation, by details, instances, examples, comparison, contrast, analogy, proof, cause, effect, chronological development, and by other means too numerous to mention. The student should collect as many and as forceful details as are necessary to explain or support the central idea, in proportion to its importance in the paper as a whole.

36. The development of an idea should include references to common and familiar things to make the thought clear and the emotion lively. The clumsy do it mechanically, first stating an idea abstractly, then giving an example. The adroit can develop the idea concretely from the beginning.

D. Sentences

37. It is frequently said that a sentence should usually put the idea that is to be emphasized in the main clause, subordinate ideas in dependent clauses and modifiers. This rule is highly questionable. Note that the very sentence that states the rule does not obey it, nor does this sentence or the next, and none of them would be greatly improved by following this principle. Perhaps a better rule is that the form of a crucial sentence should be so manipulated that the idea to be emphasized will come either first or last.

38. A sentence should fit smoothly into its context by its choice and arrangement of words. In a long sentence, the first part should grow out of the preceding sentence; the last part should lead into the following sentence.

39. The ways in which sentences are linked together, without overworking trite connectives like "however" and "therefore," is an important and time-consuming subject of study. The chief means is a close connection in thought, so that each sentence has some logical relationship to the surrounding sentences. No new terms or ideas that are likely to be strange to the reader should be introduced without preparation or explanation. A helpful device is the repetition of a key term, or synonym for it.

40. The modern sentence tends to be "loose" in construction. In general, students should not try to imitate the "balanced" or "periodic" style of the seventeenth century except in occasional sentences designed for special effects. On the other hand, they should be able to get necessary qualifications out of the way before making their main point.

41. The structure of a sentence should be simple and easy to follow. A large number of subordinate clauses may be used only when they all have the same pattern or function (e.g., "That man has had a liberal education who . . . , who . . . , who . . . , and who . . ."). Clauses subordinate to subordinate clauses should be used in moderation, and hardly ever a third order of subordination.

42. Subordinate clauses should be introduced by connectives that clearly and correctly indicate the relationship of the subordinate idea to the main idea.

43. One should be able to write sentences in many forms to fit the mood, to make the meaning clear, to flow into the surrounding sentences, or to make a point stand out. The length, order, and pattern of successive sentences should be varied except when repetition is desired for emphasis.

44. A sentence usually consists of a subject, verb, and (maybe) an object or complement. Each of these elements may be modified by words, phrases, or clauses. Then there *may* be a comma followed by "and," "or," or "but," a semicolon followed by a conjunction like "therefore," or a semicolon without any other connective. These may be followed by another subject, verb, and (maybe) an object or complement, and each of these elements may be modified by words, phrases, or clauses, as before. But then, except in most unusual circumstances, it is well to stop. There should hardly ever be three main clauses except when they are short and of the same pattern: "I came; I saw; I conquered," or "He came, and we told him, but he would not listen."

45. Another limitation on the length of a sentence is that it should contain only one idea. The idea may have several parts, but when it becomes two ideas, it requires a second sentence. In practice, of course, it is sometimes hard to tell where to draw the line, but criticism on this point will develop judgment.

46. On the other hand, a style composed almost exclusively of very short sentences sounds choppy and immature. Several adjacent sentences of this sort are usually related to one central idea; one tells the cause, another the time, a third the consequence, etc. With practice, one can turn most of these sentences into subordinate clauses or modifiers.

47. A sin that is almost unforgivable in college is the joining of two separate sentences by nothing but a comma, that betrays an abysmal lack of "sentence sense."

48. The flow of thought within a sentence, except in unusual circumstances (and for special effects, as in the works of Henry James), should not, as in this one, be interrupted (again and again!) by the insertion of too many, possibly unnecessary, parenthetical elements.

49. A sentence should come to the point with reasonable dispatch. The necessary qualifications may be subordinated, buried in the middle, or added later.

50. If a sentence lends itself to climactic order, the climax should not be spoiled by revealing the most powerful idea before the end, or by adding qualifying words and phrases after it.

51. A primary quality of good writing is energy—not to be confused with a facade of exclamation points, violent language, exaggeration, etc. Whether poised or exuberant, the sentences should have a go about them.

52. Constructions within a sentence should be consistent with one another. There should be no unnecessary shifts in subject, voice, tense, person, or number. Phrases and clauses having the same function should usually be parallel in form.

53. The reference of pronouns and of modifiers should be clear. When starting with a participle, it should not be left dangling, as in this sentence.

54. In general, related words should be placed near one another. A good trick to learn, for example, is that of placing an adverb directly before or after the verb it modifies, whenever its normal position toward the end of the clause makes trouble with the following clause.

55. A sentence should not contain any word that can be omitted without spoiling the intended effect. On the other hand, constructions must be complete; necessary words must not be omitted. "Of," "that," and the second member of a comparison are frequently omitted without justification.

56. A sentence should not be so ambiguous that a qualified and well-disposed reader will have any serious doubt as to what is meant. On the other hand, the attempt to remove every possible ambiguity results in a tiresome, legalistic style. Precision should be sought only where it is important, and to the degree necessary for the end in view. It is achieved even more by manipulation of the context than by choice of words.

57. One should learn to use controlled ambiguity (a) to avoid unnecessary argument, (b) to arouse emotion, and (c) to enrich meaning. Perhaps its most common use in daily life is the "white lie" and the "face-saving formula." At the other end of the scale, something like the "Four Freedoms" can command devotion where a bill of particulars would provoke dissension.

E. Words, Phrases, Figures of Speech

58. Words should be chosen with an eye to (a) *clarity*, aiming at the degree of precision appropriate to the context; (b) *appropriateness* to tone and purpose; (c) *effectiveness*, using specific, vivid, forceful, or unexpected words at points of emphasis; (d) *euphony*, avoiding words that are hard to pronounce together; unintended rhyme, alliteration, or assonance; and awkward, choppy rhythm.

59. One should learn to use a few words in unexpected senses and contexts that awaken a fresh perception of their meaning (e.g., a fine, *large* morning). A failure in this attempt is a malapropism, but the risk is worth taking.

60. In general, little words are better than big words, but sometimes a big word is indispensable.

61. A word should not be repeated within or near a sentence except for good reason, such as clarity, emphasis, or connection. This rule does not apply to articles, prepositions, conjunctions, or pronouns. On a larger scale, a sentence should not go over the same ground twice.

62. Adjectives and adverbs should be used in moderation.

63. In general, active verbs are better than passive verbs.

64. One should avoid jargon: words and phrases that mean nothing, unnecessary technical terms, and words too often profaned.

65. One should not mix levels of usage. If a paper is formal, it should not use colloquial or slang words or constructions. If it is informal, it should not include words, sentences, and constructions which, in that context, sound pompous and out of character.

66. A figure of speech should be capable of being reduced to a proportion that will reveal the intended relationship.

67. Successive figures of speech should be consistent with one another. A metaphor should not come in like a lion and then proceed to gild the lily.

68. One should realize that all language is metaphorical: that words could not cover the flux of experience without metaphorical extensions of their root senses. Figures of speech are not mere ornaments; they are economical ways of conveying meanings.

69. One should be able to distinguish the literal meaning of a metaphor from the intended meaning. Since Richards' terms, "vehicle" and "tenor," have not become current, the terms "literal meaning" and "figurative meaning" may help to make this distinction.

F. Semantic Considerations

70. Words should not be used as though they were identical with the things they represent. "This is X" should be understood as "For present purposes this may be classified under X because in certain respects, but not in all, it is like other things that we classify under X."

71. A word usually carries several different meanings. The context should indicate which of these meanings is intended and should warn a qualified reader against meanings that are not intended.

72. One should not impute a single, fixed meaning to a word and base a position upon it when other meanings may be intended or understood.

MEANING AND MEANINGS

Shortly after I. A. Richards became University Professor at Harvard, I had the privilege of serving for one year as one of his assistants.

He had many distinguished visitors, some of whom questioned his more paradoxical opinions. One of them said, "I can accept your general position that any English word can be given almost any meaning by its context, but surely there are limits. How, for example, could anyone make the word *house* mean *bread*?"

Without hesitation, Richards quoted a line from "The Bugler's First Communion" by Gerard Manley Hopkins, referring to the communion bread:

"Hiding in leaf-light house his too huge godhead."

Another visitor said, "I recognize that words have different meanings in different contexts. For example, in one context the word *rest* may mean *remainder*; in another context it may mean *repose*. But you seem to be saying that sometimes a word can carry two such meanings simultaneously. Apart from puns, which are trivial, how could such a word as *rest* in a given context mean both *remainder* and *repose*?"

Richards rolled his eyes heavenward for just a moment and then quoted the dying speech of Hamlet:

"The rest is silence."

73. One should not impute greater specificity of meaning to a word than is indicated by the context. When a word is used loosely, with several possible meanings in mind, one should not assume that it is intended to mean one quite definite thing.

74. One should recognize and allow for shifts in the meanings of words from one context to another. This is not only inevitable but highly desirable except within a single train of deductive reasoning.

75. One should be sensitive to the need for a clear definition or understanding (through context) of crucial terms in statements intended to be precise, or to lead to important decisions.

76. One should not use or be misled by the trick of securing assent to a proposition using a key word in one sense, and then extending this agreement to another proposition using the same word in a different sense.

77. One should not hope to carry meaning solely by a careful choice of terms. One should also indicate the sense in which one is using them by a context that makes them unambiguous.

78. One should not stretch the meaning of a term beyond the probable capacity of one's audience to grasp and retain. One should expect a common term used in a technical sense to revert many times in the course of a discussion to its common range of meanings.

79. In dealing with general statements or abstractions one should be able, if challenged, to point to concrete things or operations on which the abstractions are based.

80. All language is both "referential" and "emotive": it produces a response that is a blend of thought and feeling. Neither function is "higher" than the other; they are inseparable, and a defect in either will impair the other. Students should watch the emotional coloring of the words they use, making sure that it is in harmony with the thought, and on the highest level that the thought will sustain.

G. Argument and Rhetoric

81. Students should be able to classify arguments as inductive or deductive and recognize that both are usually involved in persuasive writing.

82. Students should be able to construct an inductive argument with careful regard to adequacy of sampling, statistical significance (when necessary), and limitation of the generality of the conclusion.

83. Students should be able to construct a deductive argument with careful regard to the validity of the premises, the consistency of terms and propositions, the avoidance of fallacies, and the soundness of each step in the reasoning.

84. Students should be able to recognize, refute, and avoid common fallacies.

85. Students should recognize the role of definitions and assumptions in argument and should be able to bring to light hidden assumptions by supplying missing premises.

86. Students should be able to adapt an argument to a given occasion and audience by such means as organization, establishing an appropriate character for the speaker, modifying the patterns of sentences, using appropriate words and figures of speech, etc.

H. Style

87. Students should realize that, in one important sense, style is not the natural and inevitable expression of a personality in writing but the gradual discovery and adoption of successful ways of achieving certain purposes in writing. It becomes habitual and recognizable only to the extent that the writer's purposes are fairly constant, and he keeps using and developing the same means of achieving them. This view of style is more fruitful than the personality theory because it dispels mystery and gives students something to do besides waiting for their personalities to achieve their predestined form. They should clarify their purposes in writing and set about discovering successful ways of achieving them.

88. Students should realize that the selection of details is an important element of style and is conditioned by the purpose in writing.

89. Students should recognize and be able to produce the effects achieved by selection of words: by various levels of usage, by concrete (image-bearing) vs. abstract words, by emotionally charged vs. neutral words, by the proportion of content to structure words, etc.

90. Students should be able to use appropriate figurative language to clarify an idea, to add interest, and to intensify emotion.

91. Students should recognize and be able to produce the effects achieved by various patterns of sentences: long or short, hard or easy to follow, normal, interrupted, or inverted patterns, few or many connectives of the various types, etc.

92. Students should be aware of the sound and rhythm of their sentences when read aloud. They should be able to write sentences in which similar metrical patterns are repeated, or suddenly changed, for emotional effects. They should equally avoid harshness and the musical effects of poetry that are inappropriate to prose. They should watch vowel and consonant sounds so that there is a pleasing variety without incongruity or awkwardness.

93. Students should be able to recognize and produce the effect of a change of pace in the movement of sentences, from quiet and deliberate to hurried, excited, or passionate.

94. Students should be able to adapt their style to various literary forms such as parable, fable, dialogue, familiar essay, criticism, scientific report, etc.

95. Students should be able to adapt their style to their attitude toward the subject (what Richards calls "tone"): admiration, irony, invective, objective appraisal, etc.

96. Students should develop a sustained interest in stylistic effects which they come upon in reading, and in discovering the means by which they were achieved, to the end that they may gradually achieve a readable prose style of their own. It is almost too much to expect that any students except born writers will achieve a mature prose style before graduation from college, but a foundation may be laid and habits may be built toward the establishment of a mature prose style by the age of thirty.

HUMOR

In a letter commenting on qualities in student writing that he almost never found in mediocre or bad papers, Professor Macklin Thomas, formerly Examiner in English at Chicago State College, concluded with the following point:

"HUMOR. Not clowning, of course, though a good writer must be allowed to snap at a good trifle; rather a knack for intellectual or dramatic (as opposed to merely verbal) irony or incongruity (as in *New Yorker* wordless cartoons). Standards here would naturally be hard to fix—one teacher's idea of what is funny often seems pawky to another. But adrift on the shoreless sobriety of student writing, we needn't drive a hard bargain. Any overt intention, however feebly executed, to indicate that the writer has some ground for thoughtful amusement should be credited as humor."

Glossary

bias is the influence on grades of irrelevant considerations such as liking or disliking the student, disagreement with his views, etc.

cluster as used in this booklet is a group of readers whose grades agree within their group and disagree with the grades of every other group to a greater extent than can be attributed to chance.

combining scores or grades suggests several procedures for putting essay grades and objective scores on a common score-scale and combining them in ways that yield total scores that conform to reasonable expectations.

correlation is a mathematical procedure that shows to what extent it is true that, the higher a student stands on one measure, the higher he stands on another. The measures need not be of the same characteristic nor on the same scale: one can correlate height in inches with weight in pounds. But one must correlate two sets of measures of the same students; there is no way to correlate two groups on the same measure.

The standard but difficult way of computing correlations is called "product-moment" correlation. Since correlating the grades assigned independently by different readers is the basic procedure in computing the reliability of essay grades, a quick and easy way that yields approximately the same results is called "top-quarter tetrachorics" and is explained for the first time in this booklet.

The correlation between two sets of essay grades for the same students is regarded as the reliability of *one rating*. Since one expects to use the sum or average of both grades as the final grade, this correlation must be "stepped up by the Spearman-Brown Prophecy Formula" to get the reliability of this set of final grades. But all the teacher has to do is to compute one percent; then he can look up the corresponding tetrachoric and the reliability in the table presented in the section on Computing the Reliability of Essay Grades.

distribution of scores, grades, etc. usually takes the form of a list of all possible scores or ratings from high to low with a tally after each score for each student who made it.

factor analysis as discussed in this booklet is a complex mathematical procedure that requires a large number of readers to grade copies of a large number of essays written by the same students. One then computes the correlation between the grades of each reader and those of every other reader. A computer can then pick out clusters of readers whose grades agree pretty well within their cluster but disagree with the grades of every other cluster to a greater extent than can be attributed to chance. A classification of comments written on these papers by the readers who best represent each cluster can then reveal the qualities in student writing that each cluster emphasized, such as ideas, organization, wording, and correctness of expression. Each of these distinctive emphases is called a *factor*.

grade-lines are usually lines drawn across a *distribution* (q.v.) of total scores on an examination to mark the dividing lines between the final letter grades or their numerical equivalents. The staff usually tries to make the percentages awarded the various grades conform to reasonable expectations.

holistic grading or scoring is a term not used in this booklet, but it refers to what is called "rating on general impression." It consists of giving a single grade or score to each essay rather than a number of ratings on various qualities. The latter is called "analytic grading."

independent grades or other measures most commonly refer to the practice of having each reader record his grades and comments on a separate work sheet and write nothing on the essays themselves. Thus no reader knows what grade any other reader has given a paper. The term "independent" was used in a different sense later, where it was argued that there must be some separation in time as well as in topic between the writing of two essays to make them genuinely independent samples of each student's writing. Two short essays written in the same session of an examination rarely differ more in quality than pages 1 and 2 of the same essay.

Kuder-Richardson Formula 21 is a quick and easy formula for computing the reliability of objective tests. All you have to know is the mean, the standard deviation, and the number of items.

loading as used in this discussion of factor analysis is basically the correlation of each reader's grades with the central tendency represented by each factor. The higher his loading on a given factor, the more he has been influenced by the distinctive emphasis of that cluster of readers.

make-up examination requires no definition because it is almost universally provided for students who were absent. In this section it is argued that students who were disappointed in their grade on the regular examination should be allowed to take the make-up, and whichever grade is higher should stand in the record.

mean as used in this booklet is the same as a mathematical average.

median is the middle score or other measure when they are ranked in order from high to low.

name-slip is the sheet on which each student identifies his paper only by any number of six figures chosen at random. He includes his name, grade, teacher, date, and any other information that may be required. These name-slips are locked up until all grades are turned in, so that no reader has any way of finding out which student wrote any paper.

normal curve is a curve representing the theoretical distribution of an infinite number of perfect measures of characteristics that are the product of more than four independent causes.

normally distributed refers to abilities or characteristics that may reasonably be expected to occur in the proportions predicted by the normal curve in large populations. Male and female are obviously not such characteristics, but complex abilities such as reading and writing are.

norms is a term not used in this booklet but is so common in discussions of objective tests that a definition may be useful. A published test is usually given to a large, representative sample of the kinds of students for whom the test is intended. The test manual contains tables showing the percent of students in each grade who fell below each score on the test.

percentile refers to the percent of students who fell below each score in the tables of norms described above.

positive reinforcement is a term popularized by the Harvard psychologist B. F. Skinner to refer to the principle that recognizing and rewarding whatever a student (or animal) does right usually has a stronger effect on learning than any kind or amount of punishment of what he does wrong.

random variation usually refers to differences in scores on measures designed to measure the same ability or characteristic, depending on the sample of tasks included in each measure. Students may just happen to be more familiar with or adept at one sample of such tasks than another.

raw scores are the number of items in objective tests that each student answered correctly.

reliability is the amount of agreement between two sets of independent measures of the same characteristic in the same students, taken at about the same time. In objective tests, it is usually an estimate of how close they would come to getting the same scores on a second test of the same kind. It differs from *correlation* (q.v.) in that the measures must be designed to measure the same characteristic, while *correlations* may be computed between quite different characteristics, such as height and weight.

Remondino's factor emphasized handwriting and neatness, which did not appear in our factor analysis because we had to use typed copies of the essays written by students. Later, when we asked teachers to rate handwritten essays, we added Remondino's factor to our list.

review of discrepant grades is a procedure in which essays that received far different grades from the original readers were referred to a small committee of the most experienced and trusted readers. They were not told what the original grades were; they knew only that the grades differed. One member of this committee gave each of these papers a third independent reading, and clerks substituted this grade for whichever of the original grades was farther from it. This procedure usually had the effect of increasing the reliability of the essay grades by at least 10 points.

significance of differences is the result of various mathematical procedures that determine the chances in a hundred or a thousand that differences between scores, averages, and other measures can reasonably be attributed to chance, given the *standard error* (q.v.) of these measures.

spiral order refers to arranging several different forms of a test in such an order that the first student in each class tested will get Form 1, the next Form 2, and so on.

staff grading is the grading of test essays by all members of the staff of that course (or that department), usually after some training in grading by common standards. The essays are identified only by code numbers so that no reader knows which student wrote any paper. Each essay is graded by two readers, who record their grades and comments on separate work sheets and write nothing on the essays themselves, so that no reader knows which grade any essay received from any other reader.

standard deviation is a special kind of average of the distances (deviations) from the mean of all scores on any measure. It shows how far the scores are spread out from the mean. If the scores are normally distributed, about two-thirds will lie within one standard deviation from the mean, and 95% within two standard deviations. This figure is more important than most teachers realize, since nearly all computations applied to test scores have the standard deviation in their formulas. It is also the basis for the standard scores used by many test publishers, in which scores lying one standard deviation apart may be designated by such numbers as 30, 40, 50, 60, and 70.

standard error may be thought of as the limits within which scores on any given measure may vary by chance. If any measurement operation were repeated a large number of times (without students' learning or forgetting anything), and we kept averaging the results until we were quite sure what the true measure was, we would find that about two-thirds of the scores leading up to this final average lay within one standard error of the true measure, and 95% within two standard errors. This computation is most important in determining whether differences between the averages of different groups are "real" or could be attributed to chance. The formula for the standard error of such averages is the standard deviation divided by the square root of the number of students.

standard scores are scores based on the standard deviation (q.v.). The standard scores for test essays recommended in this booklet are 10, 20, 30, 40, and 50, in which the mean is arbitrarily called 30 and the standard deviation 10. The second digit in such scores is understood to refer to tenths of the standard deviation.

stanine is a score scale of 9 points, based on the standard deviation, in which the mean is 5 and the standard deviation 2, so that each point in this scale covers half a standard deviation. This scale was widely used by our Armed Forces in World War II, and for some years we made strenuous efforts to get teachers to adopt it, but they were so used to thinking in terms of a scale of 5 points that they soon reverted to it. We then adapted such a scale to standard scores by the procedures discussed in the section on Standard Scores for Test Essays.

teachers' predictions may be defined as a procedure in which teachers predict how many students in each of their classes are likely to make each grade on an approaching examination. The average of these predictions indicates what percentage of students we should expect to make A's, B's, C's, etc. or the numerical equivalents of these grades. These percentages serve to keep readers from straying too far from the standards and expectations of their colleagues.

validity is mentioned only once and not discussed in this booklet since samples of student writing are direct measures of the ability we wish to measure and hence are valid by definition. If they were judged by some eccentric standard, such as their

conformity to Marxist doctrines, the grades would no longer be valid measures of writing ability, but such eccentricities are usually ironed out by the procedures of double grading and review of discrepant grades discussed in this booklet.

The validity of objective tests of reading comprehension, listening comprehension, and vocabulary is seldom questioned, and it has been sustained by numerous studies showing that they predict what they may reasonably be expected to predict. Only the sort of test of error-detection discussed in Appendix D has often been questioned as a valid test of writing ability. Here it is defended only as a quick and easy method of measuring familiarity with the rules and conventions of informal standard English. This is a valid objective in its own right, even if it had nothing to do with writing ability, but in fact such tests often attain high correlations with carefully determined grades on essays.

The writer would not defend the short test of knowledge of grammar, as exemplified by The "Shadow" Test (Appendix D) as a valid test of writing ability. Researchers in several countries over a long period of time have been unable to show that any kind of grammar, traditional or modern, has any consistent or substantial effect on writing ability in one's native language. This test measures only what it purports to measure: ability to describe a sentence in grammatical terms.

weighting is the process of giving some parts of a complex examination more credit than other parts in arriving at the total score. The weights are usually determined by discussion and are easily applied by multiplying scores on one part by something like 1.5, and scores on another part by .8. It was mentioned that researchers seldom find that weighting makes any serious difference in final grades. Students tend to come out in about the same rank order regardless of weighting.