ED 395 028 TM 025 046

AUTHOR Braun, Henry I.; Wainer, Howard
TITLE Making Essay Test Scores Fairer with Statistics. ETS Program Statistics Research Technical Report No. 89-90.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-43
PUB DATE Oct 89
NOTE 14p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essay Tests; Estimation (Mathematics); *Interrater Reliability; *Scoring; *Statistical Analysis; Trend Analysis; *True Scores
IDENTIFIERS *Advanced Placement Examinations (CEEB); *Calibration; Fairness; Variability

ABSTRACT

A desirable goal would be to develop a methodology for scoring essays so that the final grades are less affected by when or by whom each essay was read. It seems sensible to derive such grades by somehow adjusting the ratings originally given by each reader. This essay describes a solution that relies on statistical adjustment, using the context of the College Board's Advanced Placement program. Nonstatistical provisions, such as rater training, are in place to minimize the potential impact of rater differences on grades, but there is no simple way of getting a true score for an essay. The basic idea in using statistical thinking to help is to reduce the effect on scoring reliability of some of the sources of variability through calibrating readers and days on which essays are read. Estimating the relative stringency of raters and the scoring trends across time is made possible by the choice of experimental design developed by statisticians. An example illustrates the approach. Calibration experiments on five different Advanced Placement examinations showed that, in general, calibrated scores enhance reliability, but there are obstacles to overcome before the approach can be operationalized with actual essays. (Contains three tables and three references.) (SLD)

RR-89-43

Making Essay Test Scores Fairer With Statistics

Henry I. Braun
and
Howard Wainer

Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

TECHNICAL REPORT NO. 89-90

BEST COPY AVAILABLE

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants. Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

MAKING ESSAY TEST SCORES FAIRER WITH STATISTICS

Henry I. Braun
and
Howard Wainer

Educational Testing Service

Program Statistics Research
Technical Report No. 89-90

Research Report No. 89-43

Educational Testing Service
Princeton, New Jersey 08541-0001

October 1989

Making Essay Test Scores Fairer With Statistics

Henry I. Braun *Educational Testing Service*
Howard Wainer *Educational Testing Service*

INTRODUCTION

As the graded tests were handed back, a crescendo of groans echoed through the classroom. After the initial shock was registered, the long-suffering teacher smiled benignly and stated, "Your poor performance, relative to previous classes, indicates that this form of the test was more difficult than I had anticipated. I'll have to curve the scores." The students' relief was palpable.

This sort of scene is common. "Curving the scores" is the transformation of the usual rules of correspondence between percent correct and its associated letter grade. In classroom tests the effect of curving almost always allows a score to qualify for a higher grade than would ordinarily be expected. While almost everyone knows this, the question of why teachers grade on a curve is shrouded in mystery. The answer, in its simplest terms, is that we curve (adjust) test scores to allow fairer comparisons among individuals who take different forms of the test.

A similar problem of adjusting test scores for fairness occurs in the subjective scoring of essays. When a large collection of essays is to be graded, it is common to engage a number of individuals to carry out the scoring, with a different sample of essays assigned to each reader. The difficulty of an essay question involves both the inherent difficulty of the question (for example: "Describe your activities over the Christmas holidays" versus "Compare Kant's metaphysics with Aristotle's") and the strictness of the reader who scores it. We can control differences of the first kind by asking everyone the same questions, but practical considerations prevent us from using the same control for the readers. Yet, if one reader has more stringent criteria than the others, those examinees who were unfortunate enough to have their exams assigned to this reader (analogous to being assigned a more difficult test form) are at a disadvantage. Fairness requires that these differences be removed (transforming/curving the ratings of the readers so that they are comparable). Readers' criteria may also shift through time; they might be more lenient on Monday than on Friday. If such variability exists, fairness requires that these day-to-day differences also be removed.

A desirable goal, then, is to develop a methodology for scoring essays so that the final grades are less affected by when or by whom each essay was read. It seems sensible to derive such grades by somehow adjusting the ratings originally given by each reader. The rest of this essay describes one solution that relies on statistical adjustment. The solution is described in the context of a testing program that includes an important essay component, the College Board's Advance Placement (AP) Program.

THE ADVANCED PLACEMENT PROGRAM

The AP Program offers specialized curricula in a wide variety of subjects, including English, American history, European history, mathematics, biology, chemistry, French, and German. High school students who participate in the program and who do well in the final examination are eligible to receive college credit for their work. In each subject, the same final examination is given all across the United States on a particular Saturday in May. Each examination has a section of multiple-choice questions and a section of free-response questions. In mathematics or chemistry, free-response questions require the student to work out solutions to problems, while in English or American history they require the student to write essays.

The answers to the free-response questions must be scored by human raters since computer programs are not yet intelligent enough to read students' handwriting and to assign values to the material. Because tens of thousands of students may write an essay on a given topic, the grading process involves bringing together as many as a hundred readers to grade papers continuously for four or five days. The readers include both high school teachers of AP courses and college teachers of those subjects. Each essay (or problem) is read by only one reader, chosen at random from the pool of available readers. He or she assigns a grade that becomes part of the total score.

PROBLEMS WITH SCORING ESSAYS

The question of whether a student is unfairly advantaged (or disadvantaged) by having his or her essay read by one particular reader rather than another, is a critical issue. Readers, being human, will differ in their judgments of the quality of a particular essay and so the score assigned to that essay will depend to some extent on the "luck of the draw." This dependence on chance is undesirable and should be eliminated to the extent feasible.

Before we can act to eliminate this variability we have to understand how it can arise. First, different readers may have different scales for scoring. That is, two readers may agree on how to rank a set of papers but one might systematically assign higher grades than the other. Second, two readers may assign the same scores on average but generally disagree on which essays deserve high grades and which low. In practice, both kinds of discrepancies, as well as others, will occur to some extent.

Because the grading process extends over a number of days, the score assigned to an essay may also depend on when it is graded. There may be, for example, a general trend to grade more leniently (or more stringently) over the course of the week. Beyond this general trend, individual readers will exhibit their own trends through time. Such global patterns in assigned scores have nothing to do with the quality of the essays. If these patterns exist, they also contribute to the role that chance plays in the grade assigned to a student's essay.

Nonstatistical provisions are currently in place to minimize the potential impact of these factors on the grades. The AP Program carefully trains readers before the scoring sessions begin and continuously monitors them during the sessions. For each subject, a chief reader with several years experience in the program is appointed to take responsibility for the integrity of the scoring process. Soon after the answer booklets are returned, the chief reader selects a number of essays to illustrate different levels of the score scale. After extensive discussions with the senior readers and, eventually, with all the readers in the pool, the chief reader constructs a detailed list of criteria. Adherence to this "rubric" is monitored by periodically asking all the readers to grade the same paper. If substantial discrepancies occur, the readers undergo further training. This approach seems to work reasonably well but, as we shall see, there remains room for improvement.

Before we go on to discuss how statistical thinking can help, we must have some way of measuring how well a suggested approach succeeds in reducing the role of chance in grade assignment. This will provide us with a yardstick by which to judge the effectiveness of a new method.

HOW WELL ARE WE DOING?

Unfortunately there is no simple way of getting a "true score" for an essay, so we cannot simply compare the assigned score with "truth" and use the difference as an indication of the influence of chance. If an essay were read by all the readers in the pool, then the average of these scores could be used in place of a true score. It would be impractical, however, to obtain so m‹ny readings except for a very few essays.

Scientists who study test scores have followed a rather different strategy. They judge the merit of a scoring procedure by applying it twice to a large sample of essays and assessing the agreement between the two sets of results. In the case of AP, they might select a sample of 500 essays for the "experiment." Each essay would then be scored twice—each time on a day and by a reader chosen at random. Using the first set of scores, the essays would be listed from high to low. A second ranking would be obtained from the second set of scores. If the role of chance is relatively small, then an essay should fall at about the same place in the two lists. But if chance makes a large contribution, then the two rankings will differ considerably.

The level of agreement between rankings is usually measured by a quantity called the *reliability coefficient.* The reliability coefficient is a number that is calculated from the numerical information contained in the two lists. In this setting, it can range from near 0 to near 1. If there is little agreement between the two lists, the coefficient will take on a value near 0, indicating that chance is playing a substantial role in the grading process. On the other hand, if there is substantial agreement between the lists, the coefficient will take on a value near 1, indicating that chance is playing a minor role.

Typical values of the reliability coefficient for essay scores in the humanities are between .3 and .6. For problems in chemistry, the reliability coefficient usually lies between .6 and .8. To get some feeling for what these numbers mean, consider the following findings. If a group of boys are ranked by height at age six and then again at age ten the reliability coefficient for the two lists is greater than .8. If the boys are ranked by performance on an objectively scored intelligence test at two different ages, the reliability coefficient is usually greater than .6. Finally, if boys of the *same* age are ranked once by height and again by performance on an intelligence test, the reliability coefficient for the two relatively unrelated lists is usually only about .2 or .3.

It is not unusual (such as in the study described below) to have more than just two rankings of a set of essays. In situations like this we can reduce the many rankings to just two by simply choosing any two at random and calculating the reliability as before. Later, when we talk of the reliability of a particular scoring procedure, we will be referring to a measure that is closely akin to a pairwise reliability averaged over all pairs of judges.

CALIBRATING READERS AND DAYS

We are now ready to see how statistical thinking can help. The basic idea is to reduce the effect on scoring reliability of some of the sources of variability we have mentioned: systematic differences between readers or between days. By that we mean the following: if we knew, for the same set of papers, that one reader would assign scores that were on average 10 points higher than another reader's, we could adjust the first set of scores by subtracting 10 points from each of them (or by adding 10 points to each of the scores in the second set). The two sets of scores would thus have the same average. This is as it should be since they refer to the same set of papers.

Exactly the same sort of adjustments could be used to deal with systematic differences between days. If a set of papers graded on one day received scores that were on average, say, 5 points lower than they would receive on another day, we could add 5 points to the first set of scores to make the averages equal. The process of making averages equal is called *calibration*. In the context of essay scoring, calibrating both readers and days would improve the reliability of the scores by eliminating two sources of chance variation. The degree of improvement would depend on, among other things, how large these differences were in the first place.

COLLECTING DATA

Where are we to obtain the information that we need to carry out the calibration? In the operational grading, each paper is only read once—by a particular reader on a particular day. If readers assign different average grades over the course of the five-day grading period, we do not know whether to attribute those differences to real and consistent differences among the readers, or to differences in the quality of the essays they happened to read, or both. To make some progress, we will have to collect specialized data that will give us the information we need.

Statistical theory can guide us to the design of an experiment that will efficiently collect those data and tell us how to use them appropriately. Consider the following experiment. Suppose we choose a small sample of essays at random from among the pool of tens of thousands available and arrange to have each essay read by each reader on each day. The data thus obtained would allow us to estimate average differences among readers as well as average differences among days. (We use the term estimate because we would have observed the grading behavior of the readers only for the sample and not for all the essays.)

We could use these estimates, obtained from this small sample of essays, to calibrate readers and days. That means we could adjust the scores for the entire pool of essays, by whomever and whenever they were graded, based on the information collected in the experiment. But before we do that we have to consider carefully the quality of the information we would be using.

This experiment presents at least two problems. Because of the enormous number of readings that have to be carried out, there is a severe restriction on the number of extra readings that can be added for the experiment. Since each reader is to read each essay on each day, the number of essays has to be kept very small—say, five to ten. This raises questions about the representativeness of the results: Would we get substantially different estimates if we chose another set of five essays? A second issue arises from the repeated readings of the essays. To the extent that readers remember the score they assigned to an essay on the previous day and just copy it, we are not collecting bona fide information. Such distortions in the estimates could result in our making adjustments in the wrong direction, so that calibrations would lower reliability rather than raise it!

STATISTICS TO THE RESCUE

Our aim is to estimate the relative stringency of the different readers as well as the scoring trends across time without encountering the pitfalls mentioned above. Fortunately statisticians have devoted a lot of effort to solving problems of this sort. They have developed special methods for efficiently collecting data called *experimental designs*. An example of a design that meets our needs is contained in Table 1. The table represents a set of instructions for allocating readings for a four-day experiment involving 12 readers and 32 essays chosen at random from the pool. (One of the reasons that the numbers 12 and 32 were chosen is that they are both divisible by 4, the length of this particular experiment; other combinations are possible.) Each of the 32 rows corresponds to

Table 1 Allocation plan of essays to readers

	Readers											
Essays	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	1[*]	1	1	4	4	4	3	3	3	2	2	2
2	3	3	4	2	2	3	1	1	2	4	4	1
3	4	2	3	3	1	2	2	4	1	1	3	4
4	2	4	2	1	3	1	4	2	4	3	1	3
5	1	4	4	4	3	3	3	2	2	2	1	1
6	3	2	1	2	1	4	1	4	3	4	3	2
7	4	3	2	3	2	1	2	1	4	1	4	3
8	2	1	3	1	4	2	4	3	1	3	2	4
9	1	2	2	4	1	1	3	4	4	2	3	3
10	2	3	1	1	2	4	4	1	3	3	4	2
11	3	4	3	2	3	2	1	2	1	4	1	4
12	4	1	4	3	4	3	2	3	2	1	2	1
13	1	3	3	4	2	2	3	1	1	2	4	4
14	3	1	2	2	4	1	1	3	4	4	2	3
15	4	4	1	3	3	4	2	2	3	1	1	2
16	2	2	4	1	1	3	4	4	2	3	3	1
17[†]	1	1	1	4	4	4	3	3	3	2	2	2
18	3	3	4	2	2	3	1	1	2	4	4	1
19	4	2	3	3	1	2	2	4	1	1	3	4
20	2	4	2	1	3	1	4	2	4	3	1	3
21	1	4	4	4	3	3	3	2	2	2	1	1
22	3	2	1	2	1	4	1	4	3	4	3	2
23	4	3	2	3	2	1	2	1	4	1	4	3
24	2	1	3	1	4	2	4	3	1	3	2	4
25	1	2	2	4	1	1	3	4	4	2	3	3
26	2	3	1	1	2	4	4	1	3	3	4	2
27	3	4	3	2	3	2	1	2	1	4	1	4
28	4	1	4	3	4	3	2	3	2	1	2	1
29	1	3	3	4	2	2	3	1	1	2	4	4
30	3	1	2	2	4	1	1	3	4	4	2	3
31	4	4	1	3	3	4	2	2	3	1	1	2
32	2	2	4	1	1	3	4	4	2	3	3	1

[*] The entries in the table indicate the day that reader scored that essay.

[†] Rows 17–32 are just duplicates of rows 1–16.

a different essay, while each of the 12 columns corresponds to a different reader. The numbers in each row of the table indicate which readers are assigned to read that essay on that day. For example, reader 1 grades essay 16 on day 2.

This design calls for each of the 32 essays to be scored three times each day, for 96 readings altogether. (Note that if each reader were required to score every essay each day within an overall limit of 96 readings, only 8 essays could be included in the experiment. By relaxing this requirement, we are able to employ four times as many essays.) The allocation of readers to essays is not done in a haphazard way. In fact, there is a delicate choice of reader-essay combinations that enables us to obtain estimates of systematic differences among readers, even though no two readers read exactly the same set of papers.

Over the course of this four-day experiment, each reader will read each of the 32 essays exactly once. Consequently, there are no repeat reading, or carryover, effects to worry about. Since each essay is read three times each day and each reader reads eight essays each day, we can also obtain estimates of systematic differences between days. Because our estimates are based on a sample of 32 essays, rather than the eight essays that would be the limit with a complete design involving the 96 readings, they should be more representative as well. With the particular design we have chosen, it is even possible to make useful comparisons between readers on a day-by-day basis.

SOME RESULTS

To get a flavor of the results, we present the findings of one such experiment carried out for an essay question in English Literature and Composition for which scores were on a scale of 100 (low) to 900 (high). Table 2 shows for each day the average scores assigned to essays graded on that day as well as the differences between these day averages and the overall average for the entire experiment. On day 1, for example, the average score was 490, which is 7 points higher than the overall average of 483.

Ideally the day averages should be very similar and indeed they are in this case. (The largest difference among days is 12 points. This is less than 3% of the average score in the experiment.) But this means that there is very little to be gained in trying to adjust for systematic differences among days—there just aren't any!

Table 2 Daily averages and their deviations from the mean

Day	*Day Average*	*Day Average Minus Experiment Average*
1	490	7
2	479	-4
3	478	-5
4	485	2
Experiment Average	483	

On the other hand, Table 3 presents the average score assigned by each reader over the course of the experiment. To see the substantial differences more clearly, we also show the differences between these reader averages and the overall average of 483. Reader A, the most lenient reader, typically scored essays 82 points higher than the average while reader L, the most stringent reader, typically scored essays 58 points lower than the average. Remember these are just estimates of the differences in scoring levels between readers based on 32 readings. Nonetheless, they have considerable credibility because through our design we have been able to balance out sources of variation that could otherwise degrade the estimates.

It certainly appears as if, for this question at least, days don't matter much but readers do. We have to remember, though, that there are three times as many scores contributing to a day average as there are contributing to a reader average. Accordingly, some proportion of the greater variability we observe among reader averages (as compared to day averages) may be due to the vagaries of chance. However, we can capitalize on the features of this particular design and the methods of statistical hypothesis testing to properly compare the relative variation of readers versus days. When we do, we find that our first, naive impressions are justified: readers matter much more than days.

To carry out the calibration, then, we subtract 82 from all the essays graded by A, subtract 61 points from all the essays graded by B, and so on. We can judge the effectiveness of the procedure by comparing the reliability of the original scores with that of the adjusted scores. The former is .57 and the latter is .61, a difference of .04. That doesn't sound like a great improvement for all that effort. The following calculations may help put the gain in some perspective.

By using some mathematical analysis it is possible to show that if each essay had been read independently by two readers and the average of the two scores

Table 3 Reader averages and their deviations from the mean

Reader	*Reader Average*	*Reader Average Minus Experiment Average*
A	565	82
B	544	61
C	517	34
D	506	23
E	487	4
F	484	1
G	476	-7
H	473	-10
I	454	-29
J	432	-51
K	432	-51
L	425	-58
Experiment Average	483	

To ease comprehension of this table, readers have been ordered by the average grade that they assigned.

used as the final score, then the reliability of these averaged scores would be about .73. (Obtaining multiple readings is the standard way of improving reliability.) Our gain of .04 is 25% of .16 = .73 − .57, the gain in reliability possible with double reading.

Remember that with the information gleaned from this little experiment we can adjust the scores of the entire collection of essays submitted. Our data have been obtained at a small fraction of the cost of hiring enough extra readers to double read the tens of thousands of essays on hand. We estimate the cost factor will typically be about one-thirtieth. Since we have achieved one-quarter of the gain at one-thirtieth the cost, a cost/benefit analysis would yield a factor of seven or eight in favor of the calibration approach. This means that if it cost, say, $5,000 to run the experiment, it would have cost about $150,000 to hire enough readers for a complete double reading, and so one-quarter of that amount ($37,000) would be required to achieve the same gain in reliability. This suggests that using calibration should be seriously considered.

SHOULD CALIBRATION BE USED?

Calibration experiments have now been carried out on five different AP examinations. In general, calibrated scores exhibit enhanced reliability—especially when the reliability of the original scores is on the low side to begin with. In one case the estimated reliability of the calibrated scores actually exceeded the projected reliability of double reading! The obvious success of such an experiment, however, is not sufficient to guarantee the operational implementation of the procedure. There are many other issues to be addressed.

One such issue arises because the experiment we have described requires considerable planning and analysis. We have also investigated another calibration procedure that can better be adapted to the tight time schedules that must be met in reporting scores to candidates. It is one we previously mentioned; namely, to make the adjustments on the basis of the operational grades. For example, if the average grade assigned by a reader over the entire grading period was 10 points higher than the average grade for all readers, we would then subtract 10 points from all the scores that grader assigned. A potential difficulty with this approach is that this reader, by chance, may have been assigned essays that were typically better than average and deserved the higher scores. In that case, an adjustment of 10 points would be too large.

In practice, the essay booklets undergo various stages of haphazard shuffling before landing on a reader's table. Unfortunately we have no direct way of determining whether readers typically receive representative (truly random) samples of essays. But this is precisely where our experiment plays an important role. We can compare the calibration using the operational scores and the calibration using the results of the experiment (in which the sample of essays is controlled and the randomization is carefully executed). When we do, we find the results are very much the same. This gives us confidence in the simpler, cheaper method.

It is worth pointing out that the data collected in an experiment such as the one we have described can lead to insights that are just not available from the operational data. Using methods of analysis that are too technical to be described here, we can learn more about the relative contributions of readers and days to score unreliability—and do it in a way that facilitates comparisons across different tests. We can also estimate the upper limit of reliability that can be achieved through calibration. This gives us a meaningful target to shoot for.

In addition to considerations of feasibility, we also have to take into account the possible reactions of both students and schools to the notion of statistical adjustment of scores. Since the first phase of this research has clearly established that a statistically designed experiment can make the process of grading essays more fair, it only remains to iron out these other aspects before adopting its use widely. As this essay is being written (December 1987) this decision process is under way and may be operationally in place as you read about it.

PROBLEMS

1. Does training of essay raters yield the result that all readers will score the same essay identically? Why or why not?
2. Would you expect essay readers to change their scoring scale over the course of the week?
3. Why is calibration of essay readers necessary?
4. Why can't we just have all essays read by several readers?
5. What is the advantage of using the complex experimental design in Table 1 rather than just having all experimental essays read by each of the readers on each of the days?
6. How much accuracy is gained by adjusting for differences in reader performance?

REFERENCES

H. I. Braun. 1988. "Understanding Score Reliability: Experiments in Calibrating Essay Readers." *Journal of Educational Statistics* 13:1–18. This contains a full description of the experiments and procedures summarized in this essay.

W. G. Cochran and G. M. Cox. 1957. *Experimental Designs*, 2nd ed. New York: Wiley, Chapter 9. A classical work on experimental design.

H. O. Gulliksen. 1987. *Theory of Mental Tests*. Hillsdale, N.J.: Erlbaum. (Originally published in 1950 by John Wiley & Sons.) This was the first (and perhaps still the most readable) comprehensive account of mental test theory—see especially pages 211–214 for a description of how to grade essays.