

TABLE OF CONTENTS

Item-Analysis, 2

Internal analysis by d show of hallds, 2
Standards for test items: success, 8
Standards for test items: discrimination, 8
Second stage of item-analysis, 10

The Standard Error, 11

The standard error of a test score, 11
A close estimate of the standard error of a test score, 14
When are tu70 test scores "really" different?, 14
"Confidence levels", 15
Philosophic digression, 17
The standard error of an average, 20
Standard error of a difference between averages, 22
Reliability, 28
Correlation, 32
Stanines, 3~

~) Copyright 1960 by Educational Testing Servil.e

D~. M. Da~lid ~

1

SHORT-CUT STATISTICS FOR
TEACHER-MADE TESTS

For the J~on-Mathematical ~eacher

The writer is an ex-Latin-teacher with thirty years of teaching experience who was attracted to testing by the fact that so much nonsense is written and spoken about education. He wanted to find out, at least in his own classes, what worked and what did not work by means of tests of his own construction—both essay tests and objective tests. Since it took him longer than he cared to spend to analyze his test results by the precise and elegant methods favored by statisticians, he gradually learned or developed shortcuts that yielded approximately the same results.

All of these short-cuts have passed two basic tests. First, they were applied to actual data by the writer's son while he was in the eighth grade making B's in arithmetic, and he had no trouble with the mathematics

Second, they have all been discussed with competent statisticians who winced slightly but agreed that the methods are valid for the purposes for which most teachers will use them, and as precise as the data from classroom tests will ordinarily warrant.

ITEM-ANALYSIS

Item-analysis by a show of hands. One of the chief advantages of published tests over teacher-made tests is that the former are pre-tested on a large number of students like those for whom the test is intended, and then the professional test-maker gets figures on (a) the success of the group on each item (what percent got it right); (b) the discriminating power of each item (based on how many more high-scoring than low-scoring students got it right); and (c) how many high-scoring and low-scoring students chose each response to each item. The test-maker then discards items that are too hard, too easy, or non-discriminating, or else touches up items by revising some responses or substituting others. Usually at least half of the items that are pre-tested in this way are either discarded or revised, and the final form of the test contains only items that are likely to work well.

Teachers cannot pretest items for important tests on the same group that is to take the final forms of these tests, for that would show them what questions were going to be asked, and students would bone up on them. However, if teachers item-analyze each important test after it is given, they can gradually build up a file of test items that have worked well in the past or have been revised to eliminate faults that appeared in earlier forms. This file will both reduce the work of constructing tests and improve the tests. If the file is large (as it very soon will be), students seldom learn what questions to expect. Examiners report very little tendency for old items to get "easier" as the years roll on.

Unfortunately, the only way of making an item-analysis that is explained in the books on tests and measurements is so laborious and time-consuming that no teacher who has tried it once is ever likely to try it again. It consists

of preparing a form and then putting down a tally for each student's response to every question—in other words, copying all answers to all questions. If there are 40 questions in the test and 40 students, that means putting down 1,600 tallies. If one is careful, it also means checking every tally, since nothing is easier to misplace than a tally. If one skips an item, for example, all of the tallies down to the point at which one discovers the error will record the student's answers to the wrong questions. Hence there will be at least 3,200 operations to perform, not counting the correction of errors, for each forty-minute test in one class. It is not surprising, therefore, that item-analysis is almost never applied to teacher-made tests, even though it is the basic operation that all published tests have to undergo and the basic reason for whatever superiority they possess.

Fortunately, all of this work can be done by a show of hands in class in so little time that students do not resent it. It adds greatly to their understanding of the test and is the only sound basis for class discussion of items that gave trouble. If one asks students to suggest items to discuss, the first students are naturally the first to respond, and they tend to suggest items that present subtle problems of interpretation. One may never get to the point that reveals the basic weaknesses of the class.

For routine tests, the teacher may call out the numbers of their hands one by one. Each student holding a paper that got that item wrong holds up his hand. The teacher counts and announces the number of hands that are up for each item, and writes that number opposite the item on his

but it is still quite clear how much of a difference is desirable. It ought to be at least 10% of the class. In a class of 40 students, at least four more students in the bottom half should get an item right. More precisely, the difference between the highs and the lows should be at least four; if 12 highs and 8 lows in a class of 40 got an item right, it would meet this standard.

This figure was not chosen at random or by rule-of-thumb. Here we must get just a bit technical for a moment, because part of the fun is the pure swank of knowing what the experts are talking about, and knowing that one has comparable figures for one's own tests. The index of discrimination that they use is called the "biserial correlation with total test." It is a decimal that shows to what extent success on the item is related to success on the test as a whole. Putting it another way, it tells the extent to which people who did well on the whole test did better on this particular item than people who did poorly on the whole test. The professionals like to have their average biserial above .4 and are quite proud of themselves if it hits .5 or above. They look hard at items with biserials below .3 and either touch them up or get rid of them unless they can prove on other grounds that the item is a good item that is not closely related to the rest of the test. NOOT, it just happens that, for items in the middle range of difficulty (that 25% to 75% of the students answered correctly), the biserial correlation with total test is approximately equal to three times the high-low difference. This is true when the high-low difference is based on high-low halves of the class—not otherwise. If the high-low difference is four, and this is 10% of the class (of 40 students), the biserial correlation of this item with the total test will be approximately .30. If it is six, or 15% of the class, the biserial will be approximately .45. This approximation does not get seriously wrong until one reaches items that more than 80% or fewer than 20% of the class answered correctly. For these extremely easy or extremely difficult items, it is usually a serious underestimate of the true biserial. One conclusion is that, while such items may be highly discriminating; they discriminate for a very small fraction of the group. Still, one occasionally wants a very easy or a very hard item. In such cases a high-low difference of even 5% of the class may be quite acceptable, and certainly anything higher is hard to get, but the difference between high-low halves is not good in itself at these extremes. If you are willing to go along with the writer's judgment that high-low halves are most practical for item-analysis by a show of hands, the next problem is how to get papers above the middle score passed to your right, and papers below the middle score passed to your left, so that there will be a clear division between highs and lows when hands are raised. How does one find out quickly just where the middle score is? The situation is that the students have just finished scoring and checking the papers in the first ten minutes of the class period, and you propose to spend the rest of the middle score, and is not to be counted in the item-analysis. The student who did not get a paper is appointed scorekeeper and writes the figure for each item on the blackboard. The teacher now has the 22 high-scoring papers on his right, the 22 low-scoring papers on his left. Presumably the student is holding his own paper at this point, and will not be embarrassed by holding up his hand, since he will not be speaking for his own part. The four figures obtained for each item may be labeled and defined as follows:

H = the number of "highs" who got the item right

L = the number of "lows" who got the item right

H+L = "SUCCESS"—the total number who got the item right

H-L = "DISCRIMINATION" or "the high-low difference"
(how many more highs than lows got the item right)

1

period in item-analysis and discussion of items that gave trouble. You do not want to keep the class waiting five minutes while you make a distribution of scores at your desk and find the median by orthodox procedures. Let us assume that the test had forty items and was pretty hard.

One way out is to inquire, "Does anyone have a paper with a score of 35 or higher? No one? How about 34? 33? 32? 31? I see two hands. Any score higher than 31? I shall assume, then, that 31 is the highest, and that two people made it. Now let's find the lowest score. Does anyone have a paper with a score of 10 or lower? 11? I see one hand. Anything lower than 11? Then I shall assume that 11 is the lowest score, and that one person made it. Now I shall write on the blackboard all scores from 31 down to 11. Hold up your hand as soon as I come to the score of the paper that you are holding. Since I shall have my back to the room, I want you, John, to tell me how many hands are raised for each score. If none, tell me zero." The teacher calls out each score as he writes it, students whose papers have that score raise their hands; John counts the number of hands and calls it out to the teacher, who writes it after the score and calls out the next

lower score. It should take only two or three minutes to get the following distribution of scores written on the blackboard:

24	3
23	3
22	3
21	5
~n	~
17	2
16	2
15	2
14	2
13	1
12	1
11	1

Since the numbers add to 45, and 45 students are present, no one has been left out. That is unusual; usually two or three students have been distracted, and their scores have to be added at the end, but here we have the complete distribution. We then count down to the middle score—the twenty-third—and find that it falls at the score of 21, which five students made. The teacher asks all papers with scores above 21 to be handed to the students on his right, and all papers with scores below 21 to the students on his left. A few students may have to change seats to make a clear division between the high and low groups. The five papers with scores of 21 are

handed to the teacher, who passes two at random to his right and two to his left and keeps the fifth paper himself, since it is the odd paper with the

The teacher calls out the numbers of the items one by one: e.g., "Item 1." Everyone whose paper got that item right holds up his hand. The counter for the highs calls out the number of upraised hands in his section: e.g., "Fourteen." Then the counter for the lows calls out the number of upraised hands in his section: e.g., "Eight." The scorekeeper, be he teacher or student, immediately adds these two figures and calls out the total: e.g., "Twenty-two." He then subtracts the lows from the highs (in his head) and calls out the difference: e.g., "Six." Everyone copies these four figures at the bottom of item 1 on the copy of the test that he is holding: 14 8 22 6. There is no need to label them, since this is a standard sequence, before long everyone will know what it means. The rhythm of the operation is approximately as follows: Item 1. Hands. Pause for counting. 14. 8. 22. 6. Item 2.... If the teacher or a student wants to call for any of these figures again, the proper short form of the question is, "That was the high? the low? the total? the difference?"

After a little practice, the complete item-analysis for a one-period test will take between ten and twenty minutes, depending on the number of items. It would take the teacher at least two hours to do it at home, and he would make far more mistakes than will be made in class, where the most alert student will be only too happy to pounce on any mistake in counting, adding, or subtracting. Teachers in the writer's measurement classes conducted such item-analyses as far down as the fourth grade and reported that the students had no trouble understanding the procedure.

carrying it out. At the other end of the scale, even students in graduate courses do not resent it. It gives them visual, auditory, and tactile clues to the success of the class on each item, and it shows them graphically and convincingly which items separated the sheep from the goats. They get personally involved in finding out how well the class did on the test, and why they went wrong on the items that gave trouble. By contrast, if the teacher does all the work for them at home and hands them the results of his analysis on a platter, no one will understand and no one will be interested. They have to get into the act if the analysis of a test is to be a moving and enlightening experience.

Standards for test items: success. It is a common belief that most tests should start with very easy items, gradually get harder, and end with very hard items. If this sequence is hard to arrange, at least the test should cover a very wide range of item-difficulties. While many professionals share this view, it is worth knowing that practically every serious investigation of this problem since 1932 has come up with the opposite conclusion: that precision of measurement is greatest when all of the items in a test are about equally difficult for the group tested; that maximum reliability and dispersion of scores will be attained if every item in the usual sort of multiple-choice test is answered correctly by somewhere between 60% and 70% of the students tested. We do not want to insist on this point, since the advantage of a narrow range of item-difficulties is very small in relation to other sources of validity and reliability, and since it is usually almost impossible to achieve a narrow range of item-difficulties. Still, teachers should know that if they sweat hard in order to achieve a nice progression from easy to difficult, their effort has probably been wasted, and its most probable effect will be precisely the contrary of what they expect. They expect it to yield a wider spread of scores. What it actually yields is a

narrower spread of scores than if all the items were of approximately equal difficulty. Hence items that more than 90% got right should be questioned as too easy, and items that fewer than 30% got right as too hard for inclusion in a test. Questioned, mind you, not rejected—for they may be justified on other grounds.

Standards for test items: discrimination. It has already been indicated that the minimum acceptable high-low difference by professional standards is 10% of the class, and why this is so, except in very easy and very hard items. The "standard error" of this sort of high-low difference however, is so large that at least a fifth of the items that turn out to be discriminating after repeated use may fall below this standard in any administration of the test by pure chance. Let us take one such item: in which the "true" high-low difference was 6, or 15% of a class of students, which is approximately equal to a biserial of .45—quite respect discrimination. The "true" figure may be though of as the average high-low difference for this item after we had given it to 100 classes of the size and ability. Never mind now how we compute it, but the "standard error" of this difference would be 3. That means that in two-thirds of the 100 classes, the difference we actually found would be anywhere between 3 and 9; in a sixth it would be below 3, in another sixth above 9; but all obtained differences would average out after a while to 6. Hence we should be wary of rejecting an item if it falls below the suggested minimum first time it is tried if, after due consideration, we can find nothing wrong with the item. It is quite strict enough—perhaps too strict—to say not more than a fifth of the items in the final test should fall below standard, and the average high-low difference should be above 10% of class—preferably 15% or above.

A teacher who uses this method of item-analysis will soon find out that high-low differences for some of his items will be zero or negative: that the same number of students in the top and bottom halves may get the same right, or more low-scoring than high-scoring students may pick the key answer. One of the chief uses of item-analysis is to direct attention to such items. While this sort of thing can happen by pure chance, a closer look at the item will often reveal why the better students shied away from intended answer. One can touch up the ambiguity or inaccuracy and then save not only the item but the resentment of future students who would be bright enough to detect the error.

All discrimination figures look wonderful toward the end of a test that only the high-scoring students were able to finish. For example, it may appear that almost all of the high-scoring students and none of the low-scoring students answered the last item correctly—which would be ideal if it were not spurious. All the low-scoring students might have known the answer but simply did not reach the item. After a fifth of the students have dropped out, item-analysis figures are so misleading that it is well not to continue the analysis beyond this point.

Second stage of item-analysis. There may be a few items in a test that turned out to be too easy, too hard, or did not discriminate satisfactorily for no apparent reason, and class discussion does not reveal anything wrong with them. If there is time, these may be subjected to a second stage of item-analysis, which is too laborious and time-consuming to apply to more than a few items. For these few items, one asks how many in the high group, and then how many in the low group (a) omitted the item, and (b) chose each response. Results like the following may indicate what is wrong:

Omit
High 0 11
Low 0 14

i

4 2 0 0

The right answer, response 1, is indicated by a line between the highs and lows who chose it. Three more lows than highs chose it; hence its index of discrimination is -3. Why? The figures for response 2 suggest an answer. This response was too attractive to the high-scoring students. Perhaps they thought response 1 was too obvious; they suspected a trap; then they figured out some interpretation of response 2 that they could defend as the right answer. If so, discussion should reveal what interpretation they gave to response 2, and it can be revised in a way that does not permit this interpretation. At the same time, responses 4 and 5 might be made a shade more plausible, but still definitely wrong, because in their present form they were wasted; nobody chose them. Incidentally, item-analysis has probably been a factor in reducing the five-choice item, which was standard a generation ago, to the four-choice item which is more popular today except in a few item-types (such as spelling) in which the fifth response is usually "none of these." Item-writers were not very successful in framing five responses that were all sufficiently plausible to "draw blood."

THE STANDARD ERROR

The standard error of a test score. Since we have already introduced the concept of "standard error" in connection with high-low differences this may be a good time to extend the concept to test scores. The first thing to be said about it is that the standard error is not computed in the same way in these two cases and is not of anything like the same magnitude. If you look in the index of a textbook of elementary statistics, you will find at least fifteen different kinds of standard errors: of scores, averages, differences, correlations, proportions, etc. They are all computed differently and yield figures of different orders of magnitude. The standard error of an average, for example, is usually much smaller than the standard error of a single score, while the standard error of the difference between two scores is larger than the standard error of either score. They all have this basic meaning in common, however. Suppose you repeated a certain measurement operation a hundred times and kept averaging the results until no further repetitions would change that average one iota. You may think that final average as the "true" measure, no matter whether it is a score on spelling, the average of a class, the difference between two classes, the correlation between spelling and verbal intelligence, or whatnot. You might then mark off the points that would enclose the middle two-thirds of the figures you got on the various trials on your way to that final average. You would call these points one standard error above the true measure and one standard error below it. You might then go on to mark the points that would enclose the middle 95% of all the figures you got on the various trials. You would call these points two standard errors above the true measure and two standard errors below. There would still be 5% of extremely deviant figures beyond these two points, but the limits of two standard errors would enclose most of the results that you would get.

The trouble with applying this concept to testing is that we are never sure what the "true" measure is, since we do not have time in schools to measure the same attribute a hundred times, and if we did, we would change it beyond recognition. But statistical theory permits us to compute the standard error of most measurement operations on the first trial, and then we can say that the chances are two out of three that the obtained figure lies within one standard error of the true figure, and 95 out of 100 that it lies within two standard errors.

The next thing to be said about the standard error is that it is not the same as the "probable error" that was popular a generation ago, but it is based on the same idea of the limits within which measures may vary by pure chance, and either figure may be translated into the other. The chief reason why "probable error" is no longer used is that there is no way to compute it directly; one first has to compute the standard error and then take approximately two-thirds of it to get the probable error. The only point in doing so was that the early statisticians thought it would be easier for the hayseeds to grasp the idea that the chances were fifty-fifty that the obtained figure would lie within one "probable error" of the true figure, rather than that the chances were two to one that it would lie within one "standard error." On mature reflection, however, it seemed that the first idea was not really any easier to grasp than the second, and it was rather silly to keep on performing an extra operation every time one computed an error of measurement just to make the figure more appealing to the laity. The name "probable error" undeniably had more popular appeal, but the appeal was spurious on two counts. First, this kind of "error" is not "probable"; it is certain. Second, it gave the idea that someone may have made that much of a mistake in taking the measure. If any such mistakes are made, they are not included within this type of "error." It must be understood in its root sense of "variation." It assumes that all the measures have been taken and recorded accurately; even so, you are not going to get the same figure twice except by luck. The "error" indicates within what limits the obtained figures are likely to vary by pure chance.

Not all kinds of chance, however. If a teacher gets angry at the students who were absent during a crucial examination and sees to it that the make-up test is harder and marked more severely, their scores will dip in a way that could not be predicted mathematically. Mistakes in writing items, scoring, or marking unintended answers and external circumstances that may affect scores, such as sickness, noise, interruptions, hot sticky days, etc., are also beyond the pale of the standard error. The only kind of variation in scores that is standard and therefore measurable is "sampling error." Suppose you want to find out how well your students can spell. There are at least 600,000 English words that you might ask them to spell, but let us suppose that there are only 10,000 that they would ordinarily be asked to spell by the end of grade 6. If you select 100 of these words completely at random and get an accurate score on the number they were able to spell, the score will give you an estimate of the percentage of the 10,000 words that they are probably able to spell. But if you take another 100 words from the same pool of words completely at random, you know that very few students will get exactly the same score as on the first 100. Variation will be much larger if two different teachers independently try to find out how well the same class appreciates Hamlet. Here the number of valid questions that they might ask is theoretically infinite, but each time to ask only 40 questions. If we can regard each set of questions as a random sample drawn from an infinite pool of items testing the same ability, the variation in scores from one such sample to another is the sort of thing that is measured by the standard error, and in practice, the variation will be much greater than you could account for by the standard error, since I

teacher's bias will affect his selection of questions: one may be a bear character development, the other on figures of speech. They are not measuring the same attribute at all, even though both call it "appreciation Hamlet."

For these reasons, the standard error accounts for only a small part the variation in scores that may be expected in practice, but it is quite large enough to make us want to get several independent scores before making up our minds as to the degree of success of our students in attaining the objectives of the course. The standard error tells within what limits scores may be expected to vary by pure chance in the selection of items. If

add to that our own bias in the selection of items, the stupid mistakes made in writing the items and in scoring them, and external circumstances that may affect the ability of the student to answer the questions, it is obvious that the variation we may expect between two independent measures of an ability that we refer to by a single name may be quite large. Not so large, however, that we should despair of ever being able to find which of our students have been more successful than others in attaining the objectives of the course. Since we usually have them for a full year we need never rely on a single measure but can give them a long series of measures. Any one measure is like any one baseball game, in which the team that is in the cellar may clobber the team at the top. But over the whole season of 154 games, the team that is really superior will rise to the top, and the team that is really inferior will fall to the bottom. A close estimate of the standard error of a test score may be found in the following table:

The standard error is:

- 0 when the score is 0 or perfect
- 1 (a) when the score is 1 or 2 (all of these refer only to raw scores)
(b) within 2 points of a perfect score
- 2 (a) on tests of less than 24 items
(b) when score is 3 - 7
(c) within 3 - 7 points of a perfect score
- 3 (a) on tests of 24 - 47 items
(b) when score is 8- 15
(c) within 8- 15 points of a perfect score
- 4 on tests of 48 - 89 items
- 5 on tests of 90 -109 items
- 6 on tests of 110-129 items
- 7 on tests of 130-150 items

1

(except at extremes as noted above. On tests of more than 90 items, the standard error of a few scores just inside these extremes may be over-estimated by one point).

To simplify the picture still further, it is usually safe to say that the standard error of a "short" teacher-made test of up to 50 items will be three raw-score points, while the standard error of a "long" test of 100 items will be five raw-score points. The few scores that come close to a perfect score may have a slightly smaller standard error, but there are

usually not enough of them to justify separate treatment.

If your local Director of Research casts aspersions on this table, ask him to read two articles by Frederic M. Lord, "Do Tests of the Same Length Have the Same Standard Errors of Measurement?" and "Tests of the Same Length Do Have the Same Standard Error of Measurement" in *Educational and Psychological Measurement*, XVII, 4 (Winter, 1957): 510-521; and XIX, 2 (Summer, 1959): 233-239.

When are two test scores "really" different? The Cooperative Test Division of Educational Testing Service has been the first major test publisher to enforce attention to the standard error of test scores by reporting scores on its new SCAT and STEP tests as bands rather than as points. Each "band" extends from one standard error below the obtained score to one standard error above, and it is explained that the chances are two out of three that the "true" score lies somewhere within this band. Teachers are

urged not to regard two scores as "really" different unless the two bands do not overlap: i.e., unless the two scores are at least two standard errors apart. While this is a great improvement over previous practice in interpreting differences between scores, a teacher who has managed to read this far without losing his grip may want to carry this line of thinking a step further in order to get hold of the concept of "the standard error of a difference." It was indicated in passing on page 11 that the standard error of a difference between two scores is larger than the standard error of either score. Think of the difference as a rope tied between two stakes, which are the two scores. Since there is wobble in both stakes, there is bound to be more wobble in the rope than there is in either stake.

To get the standard error of the difference between two scores, square the standard error of each score, add the two squares and take the square root. For example, it was shown on page 14 that the standard error of test of 24-47 items is 3 (rounded to the nearest whole number). Three squared is nine, the square of the standard error of each score. Nine plus nine is eighteen—the sum of the squares of the standard errors of two scores. The square root of 18 is approximately 4.2. This is the standard error of the difference between the two scores. You can see at once that it is appreciably larger than the standard error of either score, which is 3. Now, if you want to be 95% sure that the two scores represent a true difference in ability, the difference between them ought to be at least twice the standard error of the difference—not twice the standard error of either score. In other words, the two scores should be at least 8.4 points apart, not just 6 points apart as the Cooperative Test recommendation implies. The Cooperative people are well aware of this point but do not use it in reporting scores because (1) it would be too complicated for teachers to square, add, and take a square root before comparing any two scores; (2) if two bands do not overlap, they usually do not touch, and the distance between them is likely to reach statistical "significance"; (3) even when they do touch, the difference between the two scores is "significant" at about the 15% confidence level, which is good enough for most classroom purposes.

"Confidence levels." When people report "findings" rather than "opinions," it is now common practice for them to tag each "finding" as
** (significant at the 1% confidence level);
* (significant at the 5% confidence level);
NS (not significant).

16

The last is professional shorthand for "not significant even at the

5% confidence level.⁹ Thus, the difference between two Cooperative Test scores whose bands touched but did not overlap would be reported as "not significant"—because it is significant only at about the 15% confidence level. That is, out of every 100 differences of exactly this size, 15 might be due to pure chance in the selection of items for the test. In any one of these cases, there is no way to tell whether the difference was "real." One can only report, after computing the "wobble" in the measure, that there are 15 chances in a hundred that it might have been a fluke. That is commonly regarded as "not significant."

It is obvious from this that a statistician is a man who, if he remains true to his principles, would never bet on horse-races. He is willing to say that a difference is "real" (i.e., not a chance difference) only if there are less than five chances in a hundred that the obtained difference could have come about by accident of sampling. Even this is considered rather a grave risk, and he is really happy only when there is less than one chance in a hundred that the difference was a fluke. Since he also has a knack for inventing names that mean the opposite of what the layman would think he meant, he calls these two points "the 5% confidence level" and "the 1% confidence level." These sound as though he had less confidence in the second than in the first, but the exact opposite is true. The first means that there are less than five chances in a hundred that the difference is a fluke, the second that there is less than one chance in a hundred. Although he would shudder at the loose language, surely we are justified as laymen in thinking of the first as "95% sure" and the second as "99% sure" that the difference is "real." We ought, however, to be sure-footed in our definitions of these looser terms. "Real," for example, here means only "non-chance." It does not necessarily mean "true," for if an experiment was set up by a very biased person, it might yield results in one direction which were the exact opposite of the truth (as it ultimately emerges from the consensus of later investigators). It would still be quite proper to say that the results obtained by the first investigator did not arise by chance—by accident of sampling. They arose from bias.

Since bias, stupidity, and carelessness seem far more likely to the layman to vitiate the results of experiments than pure chance, he wonders whether it is worth while to discount the effect of chance alone. The answer seems to be that it is worth while, chiefly because almost all educational measurements contain so large an element of pure chance that many score differences can be attributed to accidents of sampling. Then we can go on to consider whether the remaining differences are true and important, or simply the logical result of the stupid and biased way in which the experiment was conducted. But how does one establish these two levels of "significance"? First, a difference is significant at the 5% confidence level if the difference is twice as large as its own standard error (not the standard error of the test scores, but the standard error of the difference). It is significant at the 1% confidence level if the difference is 2.3 times as large as its own standard error. You divide the difference by its own standard error, and if the quotient is between 2 and 2.3, you are in the clear; if it is 2.3 or more, you are on velvet—or, as the statistician would say, "not in the challenge domain." There are many other "tests of significance," but this one is probably the most widely used in educational research, and sufficient representative to give you the basic idea.

Philosophic digression. Since it is as hard for the writer as for an equally non-mathematical reader to keep his mind on the mathematics of the testing situation, perhaps we both may be forgiven for pausing a moment to cackle over the rather odd definition of reality that has come to be accepted as a rule of the game by people who are searching for real

in the supremely important area of the growth of the mind. Such people may be visualized as primitive parents who are standing the minds of their children up against the back door and measuring the aspects of those minds that they know how to measure at all with a foot-rule that stretches contracts every time it is used. All that they feel safe in saying about their measures is that two-thirds of the time they come within an inch of true figure, but five percent of the time they are more than two inches. Therefore, before they say that the mind of Susie has grown up more like the mind of Joe toward such a goal as the appreciation of Hamlet, they that the difference between them be at least twice the amount that ruler will stretch (or contract) in measuring such differences, and probably 2~ times that amount. Since the standard error of any one measurement with this ruler is one inch, its standard error in measuring a difference will be—how much?

Square the standard error of Susie's measurement. 12

Square the standard error of Joe's measurement. 12

Add the two squares.

Take the square root.

Thus the standard error of our ruler in measuring a difference is 1.4 inches. (If you do not know how to extract square roots, any math teacher can give you a table of squares and square roots of numbers between 1 and 1,000.) Then, by the rules of the Ancient and Honorable Order of Measurers, we are allowed to certify that Susie is bigger than Joe in appreciating Hamlet only if she is at least 2~3 inches bigger on our fallible foot-rule

(twice the standard error of our instrument in measuring differences).

If other members of the tribe want to know how certain that verdict is, we can tell them that, if there were no true difference, an apparent difference as large as this would turn up less than five times in a hundred measurements of the same kind. If they have an immense prize of a ton of gold for the best appreciator of Hamlet (surely a wise investment for any community), and want to be surer than that, we can insist that Susie be at least 3.5 inches bigger on this wobbly instrument (2~ times the standard error of the instrument in measuring differences). Then we can certify that the chances are less than one in a hundred that we would get a difference as large as this if there were no true difference.

Obviously there will be a great clamor among the more ignorant members of the tribe that this is no way to go about it; the thing to do is to buy a steel foot-rule that will not stretch or squeeze on every measurement and that will yield absolutely exact results. Alas, there are no such instruments for measuring the growth of the mind, and we have to put up with those we have. Of course, there will be members of the tribe who will insist that they can ask Susie and Joe five questions about Hamlet and tell you for sure which one appreciates it best, but such people will be found to differ far more widely in their verdicts than will the measurers.

"All science," says Bertrand Russell in *The Scientific Outlook*, "is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man."

Most of philosophy, as well, has been concerned in one way or another with the problem of distinguishing appearance from reality. Like the poor educator who gets fed up with the vast amount of nonsense that is talked and written about education, and who turns to testing to find something that is real as a basis for his deductions, the philosophers have been busy since the beginning of time with the problem of separating truth from opinion—warranted assertibility from mere assertion. While they have done a great

deal to clarify the problem, there are not too many instances in which they

have come up with widely understood and accepted rules to guide the seeker of reality. Among these are the rules of logic and the canons of scientific investigation. Far down among the latter is the convention that a difference may be accepted as real (as caused by something other than the vagaries of the measuring instrument) only if it is twice as great as the standard error of the instrument in measuring differences, and preferably 2-2 times as great. That sort of ground-rule for conducting an inquiry into the truth about education would have interested Plato, and he would probably have approved of it, since he was a good mathematician himself and regarded mathematics as a basic discipline for anyone seriously interested in the search for reality.

A few disgraceful members of the teaching profession may wonder why anyone should have any trouble discovering what is real about education. What is real about it, they will tell you, is the sweat, the smell, the noise, the trouble with discipline, the overcrowded classes, the low pay, and so on. If anyone professes to find reality in education by the process of computing standard errors of differences, they will hoot with derision. We might agree that these are some of the unpleasant realities in the job of education as it is now conducted, but we are not interested in them; we want to know what is real in the process of educating: that is, in assisting the growth of the mind (not just in general but in specified dimensions, such as spelling, in arithmetic, in reading comprehension, and so on up to the appreciation of Hamlet). If we looked for such growth amid the noise and smells of the classroom of the naive realist, we might find none at all. Who, then, is overlooking the reality: the measurer who does not care about the noise, or the realist who does not care about education? Both ignore certain aspects of reality, but the part that the realist excludes from consideration seems to many level-headed people far more important. At another end of the spectrum are some very nice people who find what is real in education in the light that is in the eyes of the children, in the of their voices, in the cute things they say, and in the charm of their artistic productions. They, also, would deplore the quest for a reality that is certified by two standard errors. But they would also have to assent to the proposition that their job is not limited to keeping students happy and creative; they have to assist the growth of the mind; and it is their hypothesis that happiness and creativity assist that growth better than blood, sweat, and tears. Very well—but that hypothesis requires evidence. The evidence cannot be that the children are in fact happy and creative. It must show that they learn more than when they are unhappy and uncreative. And to show that they learn more—there you have a difference, and it is a good discipline in thinking about education to refuse to recognize it as "real" unless it is at least twice as great as the standard error in measuring such differences. ~ .

The standard error of an average. While the reader may look upon this heading gloomily as "more of the same," the proper response to it, if he only knew, is "Hope begins to dawn," or "The United States Marines are coming!"

He must have wondered how he could ever prove that any distance between two points in education was real, when the foot-rule we conjured up for measuring appreciation of Hamlet was, in common language, "accurate within one inch," yet the minimum difference we could certify as real turned out to be 3/4 inches. Also, the standard error of most classroom tests is about three raw-score points, yet the minimum difference

between two scores that we could certify as real (at the 5% confidence level) was 8.7 points. At that rate, all that we could assert about the distribution of scores that we used as an example on page 6 would be that most of the students in the top quarter of scores on this test were probably superior to most students in the bottom quarter. We could make no assertion with confidence about the scores of the middle half of the class.

All of this is sad but true; there is very little hope of proving anything in education with single measures. The real hope lies in repeated measurements: either testing many students with each single measure, or testing the same student with many different measures in the course of the year. The reason is that the standard error of an average is much smaller than the standard errors of the scores that enter into it. With each additional case or measure, the standard error gets smaller, until in practice it is really not difficult to prove that some things work better than others, or that some students are superior to others with respect to any given objective.

When we compute the standard error of an average is to divide the "standard deviation" of the scores by the square root of the number of students. If you are averaging many tests of the same ability (on the same score-scale) for a single student, you divide the "standard deviation" of his scores by the square root of the number of tests. The more general statement that takes in both of these cases is that the standard error of an average is the standard deviation of the measurer divided by the square root of the

-

20

number of measures. If the number of measures is less than thirty, you are supposed to divide by the square root of one less than the number of measures ($\sqrt{n-1}$):

Now we have to find out what the "standard deviation" is and how to compute it. This is more important than you may think, for practically every other statistic that you will ever compute has the "standard deviation" somewhere in its formula. It is like the recipe for "white sauce" in the cookbooks. You may skip it on the ground that you don't care for white sauce and want to get on to something more exotic, but you find that most of the recipes for other sauces begin, "First make some white sauce. Then...."

It is only fair to tell you right at the start that there is a very simple way to find the standard deviation, proposed by W. L. Jenkins of Lehigh University, that will work well enough when you are in a hurry and when the distribution of scores is approximately "normal"—that is, when it resembles the familiar "bell-shaped curve." Subtract the sum of the bottom sixth of scores from the sum of the top sixth and divide by half the number of students in the class:

Standard deviation =

$$\frac{\text{Sum of high sixth} - \text{sum of low sixth}}{\text{Half the number of students}}$$

We may illustrate this operation by using the distribution of scores that was printed on page 6 of this bulletin to illustrate something else. There we had 45 students. A sixth of 45 is 7.5 students. Ordinarily we would say "Forget about the half" or "Take the next higher number," but here the formula itself is an approximation; hence the numbers that go into it ought to be as nearly accurate as we can manage. While there would be no

way to take half of the eighth student from the top, we can jolly well take half of his score. Hence we add the first seven scores down from the top and then add half of the eighth score. The sum of these (as you will find if you try it on page 6) is 216. Then we add the seven scores from the bottom plus half the eighth score. The sum of these is 102. (Try it on page 6 to make sure that you know how to do it.) $216 - 102 = 114$.

Now we have to divide 114 by half the number of students, which is 22.5. In the item-analysis, we left out that half student, since it would have been impossible to get half of him to sit with the highs and half with the lows. Here there is no point in leaving him out, since it is almost as easy to divide by 22.5 as it is to divide by 22. The quotient is 5.06, which

P 21

rounds to 5 as the nearest whole number—the same as you would get in computing the standard deviation by orthodox procedures. Now, the standard error of the average score on this test is the standard deviation divided by the square root of the number of students. (Since the number is above 30, we can forget about taking one less than the number of students.) The square root of 45 students is 6.7 students. (Don't bother to compute it; look it up in a table of square roots.) The standard deviation, 5, divided by 6.7 = 50.00 divided by 67 = .75 (rounding to the nearest hundredth) .

Now you can see how the standard error of an average compares with the standard error of the scores that enter into it. Since this was a test of 40 items, the standard error of each score was approximately 3 raw-score points. The standard error of the average of the class was only three-quarters of a point. This means that the chances are two out of three that the true average of the class on exactly this sort of test at the present time lies within .75 points of the average they got on this occasion (21.06 if you want to figure it out). The chances are 20 to 1 that it lies within 1.5 points: that is, that the true average lies between 19.56 and 22.56.

This ought to show you why it is still possible to find things out about education by means of tests even though the standard error of an individual score is quite large. Most of the time you are not dealing with individuals but with classes. You have not taught Hamlet in one way to Susie and in another way to Joe, but you may well have taught it in two different ways to two different classes of approximately equal ability (for example, by using the admirable Maynard Mack film in one class but not in the other). The average scores of the two classes on the same test may very well tell you whether the film made any average difference. Remember, however, that you must take the standard error of the difference between the two averages rather than the standard error of either average. This is computed exactly as the standard error of a difference was computed on page 15: square the standard errors of the two averages, add them, and take the square root.

Standard error of a difference between averages. We should like to run through this process once more using more orthodox procedures, since there are many situations in which the simple Jenkins formula will not work. Chief among these is the situation in which the distribution of scores does not look anything like the normal bell-shaped curve, as on a mastery test in which most of the scores are within a few points of a perfect score. Again, it is hard to apply to letter-grades, where the spread in scores is very small. Third, it may be difficult to apply, and entail large random errors, when the number of measures to be averaged is very small. We shall

take up this last case in the example below, since it will serve to illustrate the standard procedure with a minimum of numbers.

The problem arose when the writer and his friends were living in Chicago and had a choice between the Pennsylvania and the New York Central in getting to New York. Most of the men preferred the Central on the ground that it was smoother. Just to be ornery, the writer argued that they were the victims of propaganda: they had been reading the slogan "The Water-Level Route—You Can Sleep" for so many years that they had come to believe it. The writer argued that there was no true difference in bumpiness at all.

. Since these were measurement men, they naturally cast about for some means of measuring bumpiness. One of them found an empty bottle that had contained Aqua Velva Shaving Lotion. It was admirable for the purpose, since it was a square bottle that could be held precisely in one position on its side, and it had a narrow mouth through which water would squirt rather than pour at every bump. They filled it half full of water—up to the point at which just no more water would spill out when the bottle was laid on its side. Then some tidy soul objected that they ought not to let the water spurt out on the floor of the car, or the porter might interrupt the experiment. This problem was solved when they got to Cleveland, which has a toy shop in the terminal. They bought a toy balloon and slipped it over the mouth of the bottle to catch the spilled water.

Then, when they all agreed that the train was going full speed, they laid the bottle on its side on the window-ledge of the car, pointing toward the aisle, so that it would make no difference whether the train was going uphill or downhill. They left it there five minutes and then took a reading to find out how much water had been displaced. (They had marked off a scale in millimeters on the side of the label.) After half an hour, when the train again was going full speed, they took another reading. There was time for only five readings before they went to bed. On the return trip, they changed tickets to the Pennsylvania and took five readings under exactly the same conditions. Since five dollars was riding on the outcome, they all checked every measure to make sure that there was no mistake and nothing unfair about the reading.

f-22~ 23

~)

It turned out that the Central displaced an average of 9 millimeters of

water per reading while the Pennsy displaced 14. This would have been enough for the average bet, but these were measurement men, so they insisted that the difference be significant at the 5% confidence level or better before the bet would be paid. They quickly performed the necessary calculations on the back of an envelope and found that the difference was significant far beyond the 1% confidence level. Hence there was less than one chance in a hundred that further readings, no matter how many times repeated, would finally average out to a verdict of "no real difference." How did they figure it?

The back of the envelope looked more or less like this:

Central

Score f d fd fd2

12 1 3 3 9

11 0 2 0 0

10 1

9 1 0 0 0

8 1 ~

7 0 -2 0 0

6 1 -3 -3 9

N = 5. 20, 2fd2

$\bar{x} = 4.4$; $s = 2$, S.D. or $\sim A$

Pennsy

Score f d fd fd2

17 1 3 3 9

16 0 2 0 0

15 1 1 1 1

14 1 0 0 0

13 1 -1 -1

12 0 -2 0 0

11 1 -3 -3 9

$\bar{x} = 12.4$; $s = 2$, S.D. or $\sim B$

S.D. $\frac{2}{\sqrt{5-1}} = \frac{2}{2} = 1$, standard error of each average

S.E.djfl = $\frac{2}{\sqrt{12+12}} = \frac{2}{\sqrt{24}} = 1.4$, the standard error of the difference.

Of course, there is quite a lot to explain here, but the actual operations are as simple as falling off a log. After each score, you put down how many times it occurred under f (frequency). Here none of the scores occurred more than once, and scores of 11 and 7 on the Central, and of 16 and 12 on the Pennsy, did not occur at all, but we have entered them as 0 to make it clearer what we are doing in the column headed "d". Notice the numbers under d: in both railroads they go 3, 2, 1, 0, -1, -2, -3. What does that look like? It looks like these numbers tell how far away each score is from the middle score. That is why the column is headed d, standing for "deviations." The middle score does not deviate at all from itself, so its deviation is 0, and is so entered. You can always fill out the "d" column quite automatically, simply numbering up and down from the middle score. The next column is headed "fd," and what does that suggest from your memories of algebra?

It suggests that you multiply each f by the corresponding d to get fd ; and that is precisely what you do: you multiply the second column by the third to get the fourth. Then what does fd^2 suggest? It suggests that if you multiply the third column by the fourth, you will get the fifth—since $d \times fd = fd^2$. Notice that wherever a zero enters into the multiplication, the product is zero, and notice that when you multiply two negative numbers together, as in columns three and four, the product is positive, as in column five. You add all those products in column five and write the sum at the bottom of the column. The rather odd symbol annexed to it, Σ , is the Greek capital S , and simply means "sum of." You divide this sum, 20, by the number of measures, 5, and get 4, the average squared deviation. The square root of 4=2, which is the "standard deviation" of the scores for both the Central and the Pennsy, computed by orthodox and standard procedures that you can apply (with a little practice) to any distribution of test scores. For practice, you might apply it to the distribution of scores on page 6. The sum of the squared deviations (Σfd^2) in that case should come out to 1129. Dividing by N , 45, you get 25, and the square root of that is 5—the same as in the shorter Jenkins method.

The two lines of figures below the point at which we found the "standard deviations" of the two railroads should by now be familiar territory that we have traversed on foot. It will be good discipline for you to read every symbol in these two lines and make sure that you know why it is there. In the first of these lines, beginning S.E., what does the S.E. stand for? "Standard error," of course, as is written out at the end of the line. What kind of standard error is it? The standard error of an average of five scores, which means that we can use the formula: standard deviation of the measures divided by the square root of one less than the number of measures (pages 20-21). We have found that the standard deviation of these measures is 2 (for both railroads). The number of measures in each case is 5. One less than this number is 4. The square root of 4 is 2. Hence the standard error of each average is 2 over 2, which is 1. See whether you can read all this in the single line of figures that begins "S.E."

Then, in the last line, S.E. difference pretty obviously stands for the standard error of the difference between these two averages: the square root of the

24 25

sum of squares of the two separate standard errors. Since both have a standard error of 1, the square is also 1, and the sum of the two squares is 2. The square root of 2 (look it up!) is 1.4. The least that the two averages can differ, therefore, and have us certify it as a real difference at the 5% confidence level, is 2.8 points (millimeters). If they differ by more than 3.5 points (2.5 times the standard error of the difference), we can certify it as "significant at the 1% confidence level." Since the actual difference between the two averages was 5 points, it is obviously far and away beyond the 1% confidence level: there is far less than one chance in a hundred that the obtained difference between the two averages was a fluke. Hence the measurers felt no compulsion to stay up all night on all subsequent trips between Chicago and New York, measuring the bumpiness of the two roads over every mile of roadbed. They had enough confidence in their statistical theory to realize that such effort would be wasted. There was considerably less than one chance in a thousand that any subsequent measurement of the same sort would ever upset the general verdict that "the Pennsy is bumpier than the Central between Chicago and New York."

Obviously such a conclusion would make the public relations officers of the Pennsy apoplectic with rage, and they might be tempted to spend fifty thousand dollars building some kind of go-cart to trail behind their train

in order to measure bumpiness with greater precision. But the whole theory of measurement suggests that such an investment would be unwise. When a difference gets out beyond the .001 level of confidence with even crude but fair measures, it is highly unlikely that refinement of the measures will show a true difference in the opposite direction. We are now in a better position to appreciate what a "standard deviation" means. It is the spread out from the middle score, or mean. One standard deviation above and one standard deviation below the mean will enclose two thirds of the scores if the distribution is normal. Two above and two below will enclose 95% of the scores. This sounds exactly like the standard error—and, in fact, the two have the same basis in statistical theory. But notice that the standard error enclosed hypothetical scores: the limits within which scores might fall by pure chance in the selection of items if the same student were given an infinite number of parallel forms (without learning anything or forgetting anything). The standard deviation encloses the actual scores made by a given class in any one administration of the test or, in this case, the actual scores made by two different subjects in five administrations of the same test.

It is worth remembering that the standard deviation will usually lie between 10% and 20% of the number of items in the test, except in mastery tests in which most students come close to a perfect score, when it will be smaller. If you have to make a quick guess, probably the safest guess for most teacher-made tests (except mastery tests) is that the standard deviation will be 15% of the number of items in the test. Since the actual scores made by a class will ordinarily spread out farther than the hypothetical scores that any individual might make on parallel forms, we must expect the standard deviation to be larger than the standard error of an individual test score. That is, in fact, what we found for the distribution of scores printed on page 6. The standard deviation of these scores was 5 raw-score points; the standard error of any individual score within this distribution was approximately 3 raw-score points; the standard error of a difference between any two of these scores was 4 points; and the standard error of the class average on this test was only .75 of one raw-score point. These figures will give you an idea of the relative order of size of the quantities we have been talking about up to this point.

26 27

RELIABILITY

Test reliability. We are now in a position to compute the reliability of objective tests in which all items are given equal weight. It will take approximately two minutes after you know the standard deviation. (If the shortcut formula for the standard deviation escapes your memory, you will find it on page 21—but that is one you ought to learn by heart.) The reliability of the test depends on just three quantities: the number of items, the standard deviation, and the mean (average). If we use n for number of items (not number of students, remember!), s for the standard deviation, and A for the mean, the recipe for computing the reliability of a test is the following:

$$r_{kl} = \frac{ns}{M(2n - M)} \quad (\text{Kuder-Richardson Formula 21})$$

In the illustrative test on page 6, n was 40, s^2 was 25. The product of these two numbers is 1,000. We then multiply the mean, 21, by number-of-items-minus-the-mean, 19, and get 399. We subtract this from 1,000 and get 601. We divide 601 by the same quantity we started with, 1,000, and get .60, the test reliability. If the number of items is less than 30, the denominator in the formula should be $(n-1) s^2$, and it is legitimate to do this at any time. It may add a point or two to the reliability, but the formula as given above saves one extra computation and is close enough for most classroom purposes. If even this much computation leaves you cold, you can find the approximate reliability of most of your tests in one of the two tables below. If the average score on your test is between 70% and 90% correct, use the first table. If it is between 50% and 70% correct, use the second table. Then compute the standard deviation of your test by the shortcut formula on page 21. If the standard deviation (labeled S.D. in the tables) is nearest to 10% of the items, use line 1; if 15%, use line 2; if 20% (which happens very rarely), use line 3. If you have to guess, use line 2. Then choose the column that is nearest to the number of items in your test. The figure at the intersection of this row and column will be the approximate reliability of your test.

Approximate Reliability of Easy Tests (average score 70% to 90% correct)

Number of items (n)	20	30	40	50	60	70	80	90	100
If S.D. is .10n	.21	.48	.62	.69	.75	.78	.81	.83	.85
If S.D. is .15n	.68	.80	.84	.88	.90	.91	.92	.93	.94
If S.D. is .20n	.84	.90	.92	.94	.95	.96	.96	.97	.97

Approximate Reliability of Hard Tests (average score 50% to 70% correct)

Number of items (n)	20	30	40
If S.D. is .10n	.21	.41	
If S.D. is .15n	.49	.67	.75
If S.D. is .20n	.74	.83	.88

These reliability coefficients are conservative estimates of the correlation you would get if you administered two parallel forms of the test so closely together that no learning took place between them and computed the correlation between the two sets of scores. In simpler terms, test reliability is an estimate of how close you would come to the set of scores if you administered a second form of the test. It is not a percent and should never be referred to as "a reliability of 60%," or "60% reliable." There is no simple way of saying what it means on a percentage basis, since it is affected both by the number of students who would get identical scores on another form of the test and by how close the remainder would come to their original scores. You could set a reliability of .60 (conceivably) if nobody got the same score the second time as the first, but everybody came close. We are often asked what level of reliability is satisfactory. The answer has to be "whatever you can get in a given time limits." Test publishers have traditionally not been satisfied with reliabilities less than .90, but teacher-made tests must usually settle for less. During the past three years, over 300 teachers have attended the writer's classes in measurement and most of these have produced tests and tried them out in their own classes. Most of those that the writer regarded as good, usable tests achieved reliabilities between .60 and .80. If we wanted a test to be

highly reliable to serve as a final examination, we usually found that it took two class periods and had to be administered on two successive days: Part I on Thursday, for example, and Part II on Friday.

It is good to compute these reliabilities routinely because they take only about two minutes apiece and flash a warning signal when the reliability dips so low (as a rough rule-of-thumb, below 60) that the scores are hardly worth recording. They will also set you up in the eyes of your colleagues as a man of science, since one of the few terms in testing they have heard about is "reliability." They vaguely believe that it takes vast erudition and possibly an electronic computer to compute reliability, and they will be greatly impressed if you can do it in two minutes for any of your tests on the back of an envelope. Still, you must not let them go away with the idea that reliability is the only virtue in a test. The easiest way to achieve it would

be to ask a large number of petty factual questions in a form that could be answered very rapidly, so that you might get 100 answers from each student within one class period. They would probably hit a reliability of .90, and since the brighter and better students would probably get higher scores than the dull and lazy, the scores might have quite a respectable correlation with your grades. Still, you would know, your colleagues would know, and your students would know that it was a lousy test. The thing to do, therefore, is to make the best test you can within the time-limits you have available, and then compute the reliability. If it is unsatisfactory, it only means that you need more items to work to a stable score; hence make another test. The following formula will tell you how many times to C~ lengthen the test to get up to any desired reliability:

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

If you want .90 and got .60 with your first test, this becomes:

$$\frac{.90 \times (1 - .60)}{1 - .60} = \frac{.90 \times .40}{.40} = 2.25$$

•60 X (1-.90) .60 X .10 0600 = 6 (times longer)

Thus, it takes 6 tests with a reliability of .60 to work up to a reliability of .90. Also, it takes 3 tests with a reliability of .75 to work up to a reliability

of .90. Either of these is entirely feasible if you have the students for a semester or for a year. Simply make up more tests of the same ability. This formula seems inconsistent with the effect of the standard deviation—the spread of scores—on reliability, and to make reliability entirely a function of the number of items in the test. The supposed inconsistency can be straightened out as follows. Suppose you have just given a test on appreciation of Hamlet to your Advanced Placement Class of superior students, and its reliability with this class turns out to be .60. That means that if you gave another test of the same kind to the same class tomorrow, quite a few students would change position enough to affect their grade. There are two ways in which you could increase this reliability. One would be to go across the hall and administer the same test to a regular, unselected class that had everybody in it from geniuses to morons. The reliability over there might well go up to .90, since these people differed so widely in

ability that another test of the same kind would not shift the rank-order of very many students. This one test would be sufficient to give that class reliable grades on Hamlet. "But," you would properly argue, "I am not responsible for the grades of the class across the hall; I am responsible for the grades of this particular class; and I want them to be sufficiently reliable so that one more test would not shift them in very many instances." Hence you apply the foregoing formula and find out that you would have to give six tests of this kind to this particular class during the unit on Hamlet to get their scores up to a reliability of .90. The formula applies only to the sort of group that you have just tested, and it assumes that the range of ability within this group is not going to change appreciably during those six tests. For this reason, the reliability can be predicted on the basis of number of items alone, assuming that the true standard deviation within this group is going to remain constant.

~CORRELATION

Correlation. This is the other magic word from the art and mystery of testing. If you can do both reliabilities and correlations, and come up with results within five minutes, your colleagues will regard you as another Einstein. Actually, any moderately bright eighth grader who has been getting B's in arithmetic can learn how to do the simpler kind of correlation in about fifteen minutes, and it should not take him longer than five minutes to compute one for a class of average size. Here's how to do it: find the percentage of students who stood in the top half of the group on both measures you are correlating and look up the correlation (r) corresponding to this percentage in the following table:

% r
37 .69
36 .65
35 .60
34 .55
33 .49
32 .43
31 .3
30 .31

29 .25
28 .19
27 .13
26 .07
25 .00
24 -.07
23 -.13
22 -.19

% r	% r
21 -.25	13 -.69
20 -.31	12 -.73

11 -.77
 10 -.81
 9 -.85
 8 -.88
 7 -.91
 6 -.93

% r
 45 .95
 44 .93
 43 .91
 42 .88
 41 .85
 40 .81
 39 .77
 38 .73

These are called "tetrachoric correlations," while the more common but more difficult kind are called "product-moment correlations." They mean the same thing, in the sense that the tetrachoric yields a fairly accurate estimate of the correlation that you would get by the product-moment method. Tetrachorics are perfectly respectable and are often used in educational research, but you can see that they are not very precise, since a difference of 1% can make a difference as great as .07 in the correlation. However, the reliability of the data that teachers usually have to work with and the relatively small numbers of students involved usually do not justify more precise methods of computation. The best you can hope to get by any method is a rough idea of the general order of magnitude of the relationship.

Since even 1/70 of the students can make so much difference in the correlation, it is important to use a standard, uniform method of counting how

19 -.37
 18 -.43
 17 -.49
 16 -.55
 15 -.60
 14 -.65

many students stood in the top half on each measure. We trust that you know how to find the middle score on each measure. List the scores on each measure from highest to lowest and put a tally after each score for each student who made it. After all the scores have been tallied, count down the tallies to half the number of students in the group. The score at which this middle tally falls is the middle score.

Suppose, now, that seven students made this middle score on one of the measures. There are 81 students in the group you are studying, and you need just 2% of the students with middle scores to get down to exactly half the number of students. In all such cases, the next higher number (in this case, 3) of the middle scores are to be counted as standing in the top half on that measure; the other four students who made the same score are to be counted as standing in the bottom half.

Obviously you are not at liberty to pick whichever three out of the middle seven students will do your correlation the most good (i.e., the three who also stood in the top half on the other measure). You have to choose them at random. Taking the first three in alphabetical order is as good a way to insure random selection as any.

You will ordinarily have the whole group of 81 listed in alphabetical order anyway, and after each name you will have the two scores that you are correlating. After you have found the middle score on each measure, go down the list and put a check after each score that stands above the middle score on that measure; a straight line after each score that stands at the middle score. Do this separately for each of the two measures. Then, if you need three more students with middle scores on Measure A to take in half of the group, put a check through the first three straight lines on Measure A that you come to in alphabetical order. If you need five more students with middle scores on Measure B, put a check through the first five straight lines after the scores on that measure. Then count how many students have two checks after their names. Turn this number into a percent by dividing it by the total number of students (81 in this case). (Incidentally, this should be the total number for which you have both measures. Students for whom one measure or the other is missing should have been discarded before you even began finding the middle score on each measure.) Look up this percent in the foregoing table. The decimal corresponding to it will be the correlation between the two measures. It is not necessary for the two measures to be on anything like the same scale. It is perfectly valid, for example, to correlate height in inches with

weight in pounds; or scores on an objective test that run from 200 to 800 with scores on an essay that run from 1 to 9. All that is necessary is to count how many students stood in the top half of this same group on both measures .

It is impossible and meaningless, however, to correlate the scores of two different groups on the same measure: for example, to correlate the scores of the boys with those of the girls (except in a way to be explained hereafter, for a purpose that would probably never occur to you). You ordinarily start with a single list of names, each of which has two scores after it. Then you can correlate the first set of scores with the second set of scores. But if you have two separate lists of names, each with a single score after it, there is no way to count how many students who stood high on the first measure also stood high on the second. There is only one measure.

Teachers often speak loosely of "correlating" one class with another when they really mean "comparing." They use the longer term only because it sounds more scientific to them; but to anyone who knows what a correlation means, it is the most flagrant of boners. Except in the unusual sort of situation to be explained below, there is no conceivable way to correlate two groups of students on the same measure; one can only correlate two sets of measures on the same students. To compare the performance of two different groups of students on the same test or other measure, you compare their averages, and if you want to find out whether the averages were "really" different, you compute the standard errors of these averages and then the standard error of the difference, as we explained on pages 20-26. The general meaning of correlation may be remembered this way. A positive correlation means that the higher a student stood on one measure, the higher he stood on the other. A negative correlation means that the higher he stood on one measure, the lower he stood on the other. (We often get such correlations: for example, between number of errors in a composition and teachers' grades on those compositions.) A zero or near-zero correlation (roughly from .25 to -.25) means that a student who stood high on one measure might stand anywhere at all on the other (for example, the correlation between height and I.Q.). A correlation indicates the degree of closeness between two sets of scores or other measures—almost never the

degree of closeness between two sets of students.

Just for the sake of completeness (to keep someone from citing a case in which such a correlation might occur, and thus upsetting your faith in the writer's omniscience), we shall add a case in which one can compare the performance of two different groups by means of a correlation—even though you will probably never have occasion to compute such a correlation. You might have the hypothesis that there are certain words that rich students are likely to know more often than poor students—to a greater degree than the normal difference between these two groups on vocabulary tests. You test 100 rich students and 100 poor students in grade 6 on 50 of these words, and for each word you put down the number of rich students and then the number of poor students who were able to pick the correct definition, like this:

	Rich
Word 1	80
2	40
3	60
4	50
5	90

	Poor
60	(got it right)
20	
40	
30	
70,	etc.

If the figures kept on going like this, you could conclude two things: first, that the rich students were about 20% better in vocabulary than the poor students; second, that the correlation of word-difficulties for the two groups was almost perfect. Every word that was relatively hard for the rich students was relatively hard for the poor students, and by the same amount; while every word that was relatively easy for the rich students was also relatively easy for the poor students, and by the same amount. Thus, you are apparently comparing the performance of two different groups on the same test by means of a correlation. Actually, you are correlating item-difficulties in groups A and B. The items have here assumed the normal position of students in a correlation; the two different groups have assumed the position of the two tests. We include this one example, even though you will probably never do one like it, just to head off what is probably the most common and most revealing mistake that teachers make in talking about correlations. They often speak of correlating the boys with the girls, or the first period class with the second period class. Except in this very unusual sort of situation, there is no possibility whatever of doing so. What they mean is "comparing."

The topic of correlation is closely related to the preceding topic of reliability, because often the only way of computing the reliability of a test is to give two tests of the same ability and correlate the two sets of

scores. This is true of (a) essay tests and (b) tests in which the items receive different numbers of points. The Kuder-Richardson Formula 21 given on page 28 will work only for objective tests in which all items are scored either 1 or 0: that is, as either right or not-right (wrongs and omits counting equally as not-right). It is also true (although this principle is often violated) of tests in which more than 20% of the students were unable to

finish: that is, of speeded tests. Speed spuriously increases reliability to an extent that, if the less able students were able to finish only half the test, it

would be almost impossible to get a reliability less than .90 (with the large number of items that such tests naturally involve). Yet sometimes it is appropriate and necessary to give a speeded test. In such cases, the only fair, acceptable way to estimate reliability is to give two tests of the same sort and compute the correlation between the two sets of scores.

Sometimes teachers cheat themselves by securing two essays, each graded independently, for their final examination; by correlating grades on the first set of essays with grades on the second set; and by calling that correlation the reliability of the examination. It is not; it is the reliability of one essay.

If you use the sum or average of both essay grades as the grade for the examination, its reliability is twice the correlation divided by one plus the correlation. For example, if the correlation is .60,

$$1 - 2 \times .60 = 1.20 = 123$$

$$1 + .60 = 1.60 = 164$$

This is called the "Spearman-Brown Prophecy Formula." Another form of it appears on page 30. It should also be used whenever you are computing reliabilities by the old method of correlating scores on even-numbered items with scores on odd-numbered items. The correlation you get is the reliability of half the test. To get the reliability of the whole test, do as above: double the correlation and divide by one plus that correlation.

STAN INES

Stanine scores. Academic bookkeeping is fearfully complicated. Almost no two institutions use the same score scale, or mean the same thing by it even when they use the same numbers or letters. It is as though all our business transactions had to be conducted in all the currencies in the world—dollars, pounds, francs, marks, drachmas, piasters, pesos, and the like—with all their shifting values from day to day. The difficulty would be further compounded if the bookkeepers handling all these currencies did not realize very clearly that one cannot add one pound, one mark, one drachma, and one piaster and come out with four dollars. That is approximately what happens when teachers add together marks on recitations, quizzes, short papers, formal writing assignments, laboratory reports, book reports, ratings, and the final examination (which is supposed to "count one-third") and come out with a grade of B, or 88%. The quantities that enter into an academic grade usually cannot be added or averaged on any rational, mathematical basis.

Meanwhile, millions of Americans have become familiar with a simpler and more rational score scale that was developed by the Air Force during World War II and has since been widely adopted. It is a scale of nine points called "stanines" (a contraction of "standard-nine") in which each point covers half a standard deviation. When a distribution is "normal," each stanine includes the following percentage of scores:

	Low	Average							High
Stanine scores	1	2	3	4	5	6	7	8	9
True proportions	4%	7%	12%	17%	20%	17%	12%	7%	4%
Rounded proportions	4%	8%	12%	16%	20%	16%	12%	8%	4%
Relative numbers	1	2	3	4	5	4	3	2	

The slight distortion of the true stanine proportions indicated above as "Rounded proportions" is the fruit of many years' experience in trying to get readers to use stanine scores in grading papers, so as to reduce the variability in their grades. They made so many mistakes in trying to use the exact proportions that the writer deliberately introduced a "rounding error" at two points as shown above. Each "rounding error" makes a difference

36 37

of only 1% in the number of papers to be placed in the piles that are affected, and if you think that a reader can detect a difference of 1% in the quality of a paper, you are crazy. You are not crazy, however, if you think that the average reader will have trouble remembering the true stanine proportions, and then in computing something like 17% of 43 papers. The best he can do is to remember that the number of papers to be placed in the various piles goes 1,2,3,4,5,4,3,2,1—as in the last line of the preceding table. This accounts for 25 papers. He then multiplies each of these numbers by any number he needs to get up to the nearest multiple of 25 and places the remaining papers in the piles they most nearly resemble in quality. More exactly, what he ought to do is this. He first ought to read about fifty papers, selected at random, and sort them out into three piles: roughly, the top quarter, the middle half, and the bottom quarter. The first crude criteria for these piles are: (1) This interests me. I like it. (2) This does not interest me. It leaves me cold. (3) This nauseates me. I dislike it. The main point to remember in this first crude division is that the papers that leave him cold are not failures, they are average. A paper has to annoy or disgust him in some way in order to get into the bottom quarter. Since it is impossible to divide fifty papers into quarters, and your reader may wonder what to do with the two extra papers, tell him to place 12 in the low pile, 26 in the middle pile, and 12 in the top pile. On the second reading, he separates each pile into three smaller piles. From the lowest pile of 12 papers, he sorts out two that are completely unacceptable, four that are unsatisfactory but have some redeeming quality, and six that are poor but passing. These might be thought of as F, D-, and D. From the middle pile of 26 papers, he sorts out ten that are dead-center average, eight that are somewhat better, and eight that are somewhat worse. From the top pile of 12 papers, he sorts out two that are straight A, four that are A- (would have been A, but something went wrong), and six that are straight B's. Of course, it may be impossible to hit exactly these proportions in any one set of fifty papers, but he can come pretty close, and in doing so he can set standards for the various piles. After the first fifty, he can simply match papers with the pile that they most nearly resemble

in quality and put them in that pile. Have him check after every 25 papers, however, to make sure that he is not straying too far from the intended proportions. This way of sorting papers sets up no absolute requirements for the various piles. It merely asserts that the top pile contains the best 4% of the papers, the next contains the next-best 8%, etc.

While many fine teachers would rather be found dead in a ditch than doing anything so mechanical as "grading on a curve," something pretty close to this curve makes very good sense when you are really trying to measure something by means of essay grades rather than using the grades as tokens of praise or blame. For example, suppose you wanted to find out how much writing ability had improved in your school in grades 10-11-12, and how the teachers did it who produced much better than average improvement two years in succession. For this purpose you might mimeograph ten essay topics requiring the same mode of writing from which each student would choose one topic at the beginning of the year and another topic at the end. (No teacher should dictate the choice of these topics, so that no reader could guess that a paper written on topic 4 was probably an initial paper, while a paper written on topic 6 was probably a final paper. Students should be left completely free to choose their own topics on both occasions.)

It has been found that if students choose a topic and write the paper entirely within one class period, the papers do not fairly represent the kinds of papers they normally hand in. The following has been found the best compromise between giving them time enough to write as well as they really can, and making sure that the writing is their own unaided work. They choose a topic and write an outline and rough draft on Day 1. The teacher keeps these overnight. On Day 2 she returns them, confiscates any papers written by Mom or Dad that the student was intending to copy, and has them write their finished paper in class, preferably in a standard essay booklet in ink. This same performance is repeated during the first and last months in school every year. Each student writes his name, the name of his teacher and class, the date, and any number of six digits, chosen by himself, in the top right-hand corner of page 1 of his essay booklet, where it can easily be torn off. He copies that same number of six digits in the top left-hand corner of page 1 of his booklet. Then, when the "name-slip" is torn off, his booklet will be identified only by a self-chosen number of six digits, so that the reader cannot tell who wrote it or in what class, or whether it was written at the beginning or end of the year. The choice of topic will furnish no clue, since the same topics have been available on both occasions. But after all the essay grades have been recorded, you can find out who wrote each paper, when, and where, by reference to the numbered name-slips which remain under lock and key in the possession of the head of department or the principal.

The point of having each student number his own booklet is to save time and make the numbers truly random. If teachers number them, they are likely to fall into some pattern that will tip off readers as to which are initial and which are final papers. Each student should choose a different number for his initial and final papers. Warn them against choosing numbers in sequence (like 123,456) or the same number repeated more than four times (like 222,222). Then you will find very few duplicate numbers, and these can be easily distinguished: first, by matching handwriting on the name-slip with handwriting in the booklet; second (if there is any doubt) by fitting the name-slip into the torn-out portion of the essay booklet. Name-slips may be torn off with an angle-iron in the shape of a carpenter's square (the sort used in firming up the corners of window-screens), and

should be torn off quickly and rather carelessly, so that, if any question arises, the slip can be matched with the corner from which it was torn. If duplicate numbers are found, letter the first one A and the second B, after matching the name-slips with the booklets.

This sort of reading for purposes of measuring the amount of improvement that has come about in each class is a very considerable chore, and it usually has to be done during the summer, after both sets of papers have been received and have been arranged in the serial order of their self-chosen numbers by the person conducting the experiment (usually the head of department). One way to reduce the chore to reasonable limits without losing much in accuracy is to select just ten initial and ten final papers (written by the same students) from each class. If any such selection is made, it has to be made completely at random. After the initial papers in each class have been arranged in the order of the self-chosen numbers, take every third paper until you have ten. Then see whether the same students have also written final papers. If any did not, choose more initial papers in the same fashion until you have ten initial and ten final papers written by the same students in each class.

If each English teacher has five classes, you will thus have 100 papers for each teacher: twenty for each of his five classes. After you have selected them, arrange all the selected papers from all classes in the order of their self-chosen numbers and tear off the name-slips, so that both papers and name-slips will be arranged in the serial order of the numbers written on them by the students. It should be obvious to everyone that this order is completely random.

Now hand 100 papers (in the order of these numbers) to each English teacher and ask him to arrange them in nine piles in order of merit in the way explained on page 38. The worst four papers are to receive a stanine score of 1, the next eight a stanine score of 2, the next twelve a stanine score

of 3, and so on. This is the point at which you can begin to see some advantage in using nine piles with a fixed percentage of papers in each pile rather than leaving it up to each reader to make his own definition of what constitutes an A paper, a B, a C, and so on. If you do the latter, you are certain to find some characters who give only A's and B's, with a few C's (like the professors in graduate courses in education), while others do not recognize the existence of the letter A, and are chiefly concerned with the number they will be allowed to flunk. If left to themselves, they will flunk about a third of the papers. You can argue about these tendencies as much as you like, but you can't change them very much. On the other hand if each reader is required to pick out the best four papers in his hundred and give them a stanine score of 9 (regardless of whether he thinks they are "really" A's or B's), and the worst four papers in his hundred and give them a stanine score of 1 (regardless of whether he thinks they should pass or not), you have some chance of getting grades that are comparable from one reader to another.

Here you have enough at stake to want to make the grades comparable. You are not grading your own students, but those of everyone else in the department; and everyone else is grading your students. You have sublime but misplaced confidence in your own grading standards, but just as little confidence in the others' standards as the others have in yours. In these circumstances teachers usually settle for the curve, and like it. It is more likely to be right than any idiosyncratic departure from that curve. One of the best writers on scientific method ventured the statement that any characteristic that is the product of more than a few independent causes will assume the shape of a "normal" distribution. It may be possible to think

of exceptions, but writing is certainly the product of more than four independent causes, and certainly does fall into a "normal" distribution whenever the number of papers is large. You Will make fewer mistakes by sticking rather closely to that sort of curve than by trusting to the diverse standards of your colleagues as to what constitutes an A, a B, a C, and so on. If you do the latter, and an undue proportion of the initial papers of your students happen to fall into the hands of an "easy-marker," while an undue proportion of the final papers fall into the hands of a "hard-marker," it will look as though they had forgotten everything they knew about writing

40 1 41

in the course of the year. When this is a chance that all the teachers are taking, they are usually more willing to settle for a curve than for the possibility that every member of the staff really "knows" what an A is, a B, and so on.

When you get the papers back, write their scores on the corresponding "name-slips," sort the name-slips into piles for each teacher, then into five more piles for the separate classes taught by each teacher, and finally alphabetize the slips within each pile so that the initial and final name-slips of each student will be brought together. Subtract each student's initial score from his final score and call the difference his "gain." Some of these "gains" will be zero or negative, due partly to the unreliability of the grading and partly to the unreliability of the student, and they should be indicated as 0, ~ 2, -3, and the like; but most of the "gains" will be positive, partly because we teachers are better at our jobs than people give us credit for, and partly because all the forces of growth are on our side. Then you ought to do something that you will not expect. You ought to take all students who started with scores of 1, 2, and 3 and compute their average gain. Then take all students who started with scores of 4, 5, and 6 and compute their average gain; and finally all those who started with scores of 7, 8, and 9 and compute their average gain. You will see why as soon as you have computed the three average gains. Those who started low will have gained by far the most; those who started in the middle will have gained about half as much; and those who started high will have gained little or nothing. This is due partly to "ceiling effect": to the fact that those who started low in this situation have nowhere to go but up, while those who started high have nowhere to go but down. It is principally due, however, to the phenomenon of "regression." Scores on a second test of any kind will always "regress" toward the mean by a spurious amount proportional to the unreliability of the tests. Grades on essays are bound to be highly unreliable; hence the "regression effect" will be quite substantial and grossly unfair to teachers whose students started high. If you made a straight comparison of average gains across the board, the teacher who would Will the prize for good teaching every year on this basis would be whoever happened to be teaching the students who scored lowest on the initial test. Their gains are always the largest, wherever they are tested, and no matter who is teaching them, but only about half of that gain is real; the other half is due entirely to the unreliability of the tests. At the other end of the scale, about half of what the good students really learn is cancelled out by the effects of regression, so that if they really learned twice as much as the poor students, it would look as though they had learned only half as much. Let us make that point clearer by a numerical example. If the low-scoring students really gained two points, they would get another two points by regression, making a total observed gain of four points. If the high-scoring students really gained four points, they would lose two points by regression, making a total observed gain of only two points. Thus they actually gained

average gains across the board. The latter would give the prize by a wide margin to Teacher D, whose thirty lowest students made average raw-score gains of 5.80 points (if you add their "comparative gain" to the average gain reported in column 2). Such gains are absolutely impossible at higher levels of initial scores, and would quite overshadow the equally remarkable gains shown by Teacher C, with his superior students who are always handicapped in this sort of "before and after" comparison. There is no way that is both good and simple to reduce gains for each teacher to a single, comparable figure.

You must not begin drawing conclusions from tables like these until you have obtained such measures at least two years in succession and can begin to sort out trends from consistent performance. Then you can begin compare gains of teachers who have required a paper a week with those of teachers who have required less than one a month; those of teachers who have proved themselves on tests that they are good writers with those who have proved they are poor writers; those who can tell a good paper from a poor paper with those who call it; and those who can detect flaws in papers and correct them acceptably with those who miss half the flaws, see flaws that are not there, and correct them in ways that are worse than what the student wrote. In other words, you can begin finding things out by means of your essays grades rather than using them only as tokens of praise or blame. But such discoveries require the comparability of standards from one reader to another that stanine scoring fosters and encourages. If such comparability were extended by expressing all other measures as stanine scores, we might at long last begin to use academic bookkeeping as an instrument of research.

PAUL B. DIEDERICH / ~ ~ t ~C
~ ti l c j ~ ~ ?~

C~ (. c~ c~ J ~

D103R40X It~m 279200

Current Titles in the Evaluation and Advisory Service Series

No. 1. Analyzing the Classroom Test a Guide for Teachers

No. 2. Short-cut Statistics for Teacher-made Tests

In ETS Developments The Evaluation Series from time to time and will be distributed by the publication of the Evaluation Service a division of ETS standing and use of measurement large to assist educators in their work.

Martin Katz
Series Editor