DOCUMENT RESUME

ED 081 785 TM 003 145

AUTHOR Diederich, Paul B.
TITLE Short-Cut Statistics for Teacher-Made Tests.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 73
NOTE 12p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Correlation; Guides; *Item Analysis; *Standard Error of Measurement; Statistics; Teacher Developed Materials; *Test Reliability; *Test Results

ABSTRACT

Written by an ex-Latin teacher, short-cuts to analyzing test results for the non-mathematical teacher are provided. Discussions are given of item analysis (item analysis by a show of hands, standards for test items: success, standards for test items: discrimination, and the second stage of item analysis. The standard error is then presented (the standard error of a test score, estimated standard errors of test scores, when two test scores are "really" different, levels of significance, philosophic digression, the standard error of an average, standard error of a difference between averages). Test reliability computation is then described, and t'e way to compute a simple type of correlation for a class of average size is presented. (DB)

SHORT-CUT STATISTICS FOR TEACHER-MADE TESTS

by Paul B. Diederich

ETS EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY

For the Non-mathematical Teacher

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 short-cuts that yielded approximately the same results.

All of these short-cuts have passed two basic tests. First, they were all 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 pretested 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 pretested 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.

Yet 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 a better basis for class discussion of items that gave trouble than having students suggest items to discuss. The bright 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 items that reveal the basic weaknesses of the class.

For routine tests, the teacher may call out the numbers of the items 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 he sees for each item, and writes that number opposite the item on his own copy of the test, encircling items that call for discussion. It goes like this:

"Item 1. How many of you are holding a paper than got item 1 wrong? Hold up your hands. I see three hands. Anyone else? Let me repeat my question to make sure that you have this straight. Look at item 1. Is it marked right or wrong? If it is marked *wrong*, hold up your hands. I now see four hands. Larry, what was the trouble? You thought I meant *right*? No, that is the other kind of item-analysis; here I just want to find out which items gave us the most trouble. Now go on to item 2. Hands? I see two hands. Item 3? No hands. Did nobody get it wrong? Very good. Item 4. I see fourteen hands; we'll have to discuss that one. Item 5, zero. Item 6, two."

And so on. Remember that the teacher records the number of errors opposite each item on his own copy of the test, and encircles questions that enough students missed to warrant discussion.

For more important tests, the teacher may want the "high-low" type of item-analysis that will also reveal the discriminating power of each item, as shown by the fact that more high-scoring than low-scoring students got it right. Professionals use the top 27% in total scores on the test as the "high" group, the bottom 27% as the "low." If the teacher uses these proportions, he can use the *Item Analysis Table* prepared by Chung-Teh Fan in 1952 (available through Office of Information Services, Educational Testing Service, at two dollars per copy) to look up all the item-statistics discussed in the following section.

The writer, who has been conducting these item-counts by a show of hands for years in his own classes, prefers using top and bottom *halves* for the reason that otherwise the whole middle half of the class has nothing to do during the item-analysis and feels left out and gets into mischief. One must expect smaller differences in percent correct than one would get between the top and bottom 27%, 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 top half than in the bottom half should get an item right.

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.

Now, 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, expressed as a percent of the class. 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 consolation 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 a good index at these extremes.

Do not fear that you will have to compute these percents for every item. When you begin each item-analysis, divide the number of students who are present by 10 and round to the nearest whole number. If 38 are present, the minimum acceptable high-low difference will be 4. If an item exceeds this number, its discrimination is satisfactory; if not, you will have to look at it to see whether anything is wrong.

After you or the students have finished scoring the test, arrange the papers in descending order of total scores and count down to the middle score. Suppose this score is 21, and five students made it. All papers above this score obviously go into the high group; those below go into the low. But what about the five middle papers? Put them at random into the high and low piles until the numbers in each pile are equal. If you have an odd number of students, hold out one middle paper and do not count it in the item-analysis. The student who does not get a paper will be the score-keeper and will write the figures for each item on the blackboard; otherwise the teacher will do it.

Now it is necessary to have a clear separation between the "highs" and the "lows" in the classroom. To avoid shifting the students, those on the right may get the high papers, those on the left the low; or those in front may get the highs, those in back the lows. The teacher appoints a counter for each group, to call out the number of hands raised in his part of the room for each item.

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)

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 score-keeper, 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, and 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, "What 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 every alert student will be only too happy to pounce on any mistake in counting, adding, or subtracting. Teachers in the writer's measurement classes have conducted such item-analyses as far down as the fourth grade and have reported that the students had no trouble understanding the procedure or 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 every hard items. If this sequence is hard to arrange, at least the test should cover a 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 quite discriminating after repeated use may fall below this standard in any one administration of the test by pure chance. Hence we should be wary of rejecting an item if it falls below the suggested minimum the first time it is tried if, after due consideration, we can find nothing wrong with the item. It is quite strict enough to say that not more than a fifth of the items in the final test should fall below this standard, and the *average* high-low difference should be above 10% of the class—preferably 15% or above. High discrimination spreads out the scores as widely as possible and hence increases the reliability of the test.

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 is, the same number of students in the top and bottom halves may get them right, or more low-scoring than high-scoring students may pick the keyed 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 the intended answer. One can touch up the ambiguity or inaccuracy and thereby 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:

		Responses				
	Omit	1	2	3	4	5
High	0	11	9	0	0	0
Low	0	14	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 the 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 concepts 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 the 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 of 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 figures 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 figures 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 the "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. This variation, due to the sample that happens to be chosen, is what the standard error means.

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

For these reasons, the standard error accounts for only a small part of 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 we make 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 we add to that our own *bias* in the selection of items, the *stupid mistakes* we make in writing the items and in scoring them, and *external circumstances* that may affect the ability of the students to answer the questions, it is obvious that the variations we may expect between two independent measures of an ability that we refer to by a single name may be quite large. It is not so large, however, that we should despair of ever being able to find out 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, the team that is really superior will rise to the top, and the team that is really inferior will fall to the bottom.

Estimated Standard Errors of Test Scores:

NUMBER OF ITEMS	STANDARD ERROR
< 24	2
24-47	3
48-89	4
90-109	5
110-129	6
130-150	7

EXCEPTIONS: Regardless of the length of test, the standard error is:
0 when the score is zero or perfect;
1 when 1 or 2 points from 0 or from 100%;
2 when 3 to 7 points from 0 or from 100%;
3 when 8 to 15 points from 0 or from 100%.

This table may be interpreted as follows: In an objective test of 50 items. two scores out of three will lie within 4 raw-score points (one standard error) of the "true score" these students would attain if you continued testing with repeated random samples from the universe of items testing the same ability, and 95% of the scores will lie within 8 raw-score points (two standard errors) of "true scores." The relatively few scores at the extremes will have slightly smaller standard errors, as indicated under "Exceptions," but there are usually not enough of these 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? Cooperative Tests and Services, 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 4 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 above that the standard error of a 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 such scores. The square root of 18 is approximately 4¼. 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 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 8½ 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% level, which is good enough for most classroom purposes.

Levels of significance. When people report "findings" rather than "opinions," it is common practice for them to tag each "finding" as

**(significant at the 1% level);
*(significant at the 5% level);
NS(not significant).

The last is professional shorthand for "not significant even at the 5% level." 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 the 15% 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% level" and "the 1% level." These sound as though the second was less significant than the first, but the 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 that were the opposite of the truth (as it ultimately emerges from the consensus of later investigators). It would still be 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. The critic 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% level if the difference is twice as large as its own standard error (not the standard error of the two scores, but the standard error of the difference). It is significant at the 1% level if the difference is 2.6 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 2.6, you are in the clear; if it is 2.6 or more, you are on velvet—or, as the statistician would say, "not in the chance domain." There is, of course, no reason to set any particular limit as the boundary between reality and chance, but the 5% and 1% levels of significance are most commonly reported for the sake of simplicity. There are many other "tests of significance," but this one is probably the most widely used in educational research, and sufficiently 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 reality 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 or 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 the true figure, but five percent of the time they are more than two inches off. Therefore, before they say that the mind of Susie has grown up more than the mind of Joe toward such a goal as the appreciation of *Hamlet*, they ask that the difference between them be at least twice the amount that the ruler will stretch (or contract) in measuring such differences. and preferably 2.6 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.	$1^2 = 1$.
Square the standard error of Joe's measurement.	$1^2 = 1$.
Add the two squares.	2
Take the square root.	1.4

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 Measures, we are allowed to certify that Susie is bigger than Joe in appreciating *Hamlet* only if she is at least 2.8 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 3.6 inches bigger on this wobbly instrument (2.6 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 exact 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 2.6 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 educating as it is now conducted, but we are not interested in them; we want to find out 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 in 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 lilt 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 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 fo t-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 2.8 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% level) was 8½ points. At that rate, all that we could assert about the distribution of scores on most classroom tests 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 wi h respect to any given objective.

The way to compute the standard error of a class 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 measures divided by the square root of the 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. . . ."

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 tested:

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

Let us try this formula on the following distribution of scores on a test of 40 items:

31	2	24	3	17	2
30	1	23	3	16	2
29	1	22	3	15	2
28	1	21	5	14	2
27	2	20	3	13	1
26	2	19	3	12	1
25	2	18	3	11	1

There are 45 students. A sixth of 45 is 7½ 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 is 216. Then we add the seven scores from the bottom plus half the eighth score. The sum of these is 102. Subtracting 102 from 216 gives us 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 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 now turns out to be 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 5: 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.

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% 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% 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	1	1	1
9	1	0	0	0
8	1	−1	−1	1
7	0	−2	0	0
6	1	−3	−3	9
N = 5.				20. Σfd²

$\frac{20}{5} = 1.\ \sqrt{4} = 2$, S.D. or σ

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	1
12	0	−2	0	0
11	1	−3	−3	9
N = 5.				20, Σfd²

$\frac{20}{5} = 4.\ \sqrt{4} = 2$, S.D. or σ

$$\text{S.E.} = \frac{\text{S.D.}}{\sqrt{N-1}} = \frac{2}{\sqrt{5-1}} = \frac{2}{\sqrt{4}} = \frac{2}{2} = 1,$$ standard error of each average

$$\text{S.E.}_{\text{diff.}} = \sqrt{1^2+1^2} = \sqrt{2} = 1.4,$$ 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 7. 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 (page 7). 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.$_{diff.}$ pretty obviously stands for the standard error of the *difference* between these two averages: the square root of the 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% level, is 2.8 points (millimeters). If they differ by more than 3.6 points (2.6 times the standard error of the difference), we can certify it as "significant at the 1% level." Since the actual difference between the two averages was 5 points, it is obviously far and away beyond the 1% 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 hundred 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 trains 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 1% level 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" is. It is a kind of average of how far the scores are 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. This is, in fact, what we found for the distribution of scores printed on page 7. 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 4¼ 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.

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 7—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 *M* for the mean, the formula for computing the reliability of a test is the following:

$$\text{rel.} = 1 - \frac{M(n-M)}{ns^2} \text{ (Kuder-Richardson Formula 21)}$$

In the scores printed on page 7, the mean was 21, the number of items was 40, and so the number of items minus the mean was 19. 21×19=399. In the denominator, n was 40 and the square of the standard deviation was 25. 40×25 = 1,000. Rounding a bit, we get 400 over 1,000 or .4. Then —do not forget this—we subtract .4 from 1 and get .6 (or .60, if that looks more familiar) as the reliability of the test.

If even this much computation leaves you cold, you can find the approximate reliability of most of your tests in one of the following tables. 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 7. 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 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 50% to 70% correct)

Number of items (n)	20	30	40	50	60	70	80	90	100
If S.D. is .10n	—	.21	.41	.53	.61	.66	.71	.74	.77
If S.D. is .15n	.49	.67	.75	.80	.84	.86	.88	.89	.90
If S.D. is .20n	.74	.83	.87	.90	.92	.93	.94	.94	.95

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 same set of scores if you gave a parallel form of the test. It is not a percent and should never be referred to as "a reliability of 60%," or "60% reliable."

Note the decisive effect of the standard deviation—because it is in the denominator of the reliability formula and squared. A large number in the denominator at this point will make a smaller quantity to be subtracted from 1 and hence leave a larger reliability. The number of items, *n*, also in the denominator, has a similar effect. The location of the mean, *M*, in the numerator may seem to give an advantage to easy tests, but this is more than offset by the fact that such tests generally have a smaller standard deviation.

We are often asked what level of reliability is satisfactory. The answer has to be "whatever you can get in a given field within given time limits." Test publishers have traditionally not been satisfied with reliabilities less than .90, but teacher-made tests must usually settle for less. 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 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 up to a stable score; hence make another test. The following formula will tell you how many times to lengthen the test to get up to any desired reliability:

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

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

$$\frac{.90 \times (1-.60)}{.60 \times (1-.90)} = \frac{.90 \times .40}{.60 \times .10} = \frac{.3600}{.0600} = 6 \text{ (times longer)}$$

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.

We must not forget the lesson of our first section: that reliability can be increased (per unit of testing time) by dropping or touching up items that proved to be too hard, too easy, or non-discriminating. This, also, is not inconsistent with the formula for lengthening the test. That formula merely says, "Given the kinds of items you have now, it will take X times more items to boost reliability to .90." But if you drop hopeless items and improve others, the desired reliability may well be attained with fewer items than the formula predicts.

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	%	r	%	r	%	r	%	r
45	.95	37	.69	29	.25	21	-.25	13	-.69
~~44~~	~~.93~~	~~36~~	~~.65~~	28	.19	20	-.31	12	-.73
43	.91	35	.60	27	.13	19	-.37	11	-.77
42	.88	34	.55	26	.07	18	-.43	10	-.81
41	.85	33	.49	25	.00	17	-.49	9	-.85
40	.81	32	.43	24	-.07	16	-.55	8	-.88
39	.77	31	.37	23	-.13	15	-.60	7	-.91
38	.73	30	.31	22	-.19	14	-.65	6	-.93

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% of the students can make so much difference in the correlation, it is important to use a standard, uniform method of counting how 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.

You will ordinarily have the students listed in alphabetical order, 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 (not by the number in the top half). 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. You 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 anoth 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. There is no 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 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 7-9.

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.).

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 10 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 low reliability. 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,

$$\text{rel.} = \frac{2 \times .60}{1 + .60} = \frac{1.20}{1.60} = \frac{12}{16} = \frac{4}{3} = .75$$

This is called the "Spearman-Brown Prophecy Formula." Another form of it appears on page 10. 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.