DOCUMENT RESUME

ED 068 486 TM 001 822

AUTHOR Wright, E. N.
TITLE Examinations, Marks, Grades and Scales: A Working Paper.
INSTITUTION Toronto Board of Education (Ontario). Research Dept.
NOTE 15p.; Working paper no. 19

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Performance; Essay Tests; Evaluation Criteria; *Evaluation Techniques; Grades (Scholastic); Measurement Techniques; Objective Tests; *Standardized Tests; Statistical Analysis; Statistics; *Student Evaluation; Testing; *Test Interpretation; Test Reliability; Test Results; Tests; Test Validity; *Transformations (Mathematics)

ABSTRACT

Comparisons of students' educational performances are usually based on test and examination results. However, for such comparisons to be valid, it is suggested that evaluations must be made on some common basis since many educational and employment decisions are based on these evaluations. Standardized tests, often used when comparisons are to be made, offer some guarantee of a common evaluative basis, even though they may contain measurement errors. Among the many transforming and scaling work techniques that have been suggested, two measurements are considered useful in terms of groups of scores, that is, central tendency (mean and median) and dispersion (standard deviation and interquartile range) measures. It is suggested that while transformations are required in order to validate results from different examinations, one should consider what purpose the transformation is to serve. Also, discussed are transformations that convert marks to T scores and the differences among percentiles, ranks, and standard scores. It is concluded that certain basic educational problems are revealed when scaling or transformation techniques are used. (JS)

EXAMINATIONS, MARKS, GRADES AND SCALES:

A WORKING PAPER

E. N. WRIGHT
Research Associate
The Board of Education
for the City of Toronto

RESEARCH SERVICE

issued by the
Research Department

THE BOARD OF EDUCATION FOR THE CITY OF TORONTO

TABLE OF CONTENTS

EXAMINATIONS, MARKS, GRADES AND SCALES: A WORKING PAPER

Many recorded evaluations are made of students' performance and learning. These evaluations, generally based on tests and examinations, are used in making some of the following comparisons:

between an individual's present and previous performance;

among students in a class or among classes in a school;

among schools in a school system or systems.

The educational importance of such comparisons, the whole question of when to evaluate and what to evaluate are matters beyond the scope of this discussion. Here we are concerned with the problem that if the comparisons are to be valid, the evaluations must be made on some common basis. Is a mark of 50 in English Composition equivalent to a mark of 50 in Algebra? Is Mr. Brown's mark of 80 in History equivalent to Mr. White's mark of 80 in History? Is a mark of 60 on a Grade Nine examination equivalent to a mark of 60 on a Grade Twelve examination? Is a mark of 70 at Christmas equivalent to a mark of 70 at Easter? Attempting to deal with such questions is not a mere academic exercise when many decisions, both in education and employment, are based on these results.

Examinations, Tests and Marks

A test or examination is commonly evaluated in terms of some numerical score. It is these numbers that are the centre of so much attention. In an objective test (multiple-choice, true or false, or matching) any clerk can score the paper and any two people will give the student the same score. An essay-type test is a different matter. There are many justifications for both types of examinations and the

objectives of the course and the tests will determine which is used. Research evidence is unequivocal; when a group of highly expert teachers all mark the same student's answer, they will not all arrive at the same numerical rating for it. The differences between them will of course be diminished if they are working from a common marking scheme. It has been shown that the same teacher given the same paper at a later date will assign it a different mark. This does not cast any aspersions on either the teacher or the essay-type examinations, rather it indicates some of the problems in trying to deal mathematically with evaluation. The objective test only standardizes the problems, that is, it makes them consistent, it does not eliminate them.

There are many purposes for examinations and many philosophies of marking and grading. Different systems of marking and grading are usually quite justifiable, for teachers have different objectives, the content of subjects varies, and the kind of performances demanded of the students on an examination will therefore vary. These variations, however, create serious problems when comparisons are attempted.

Two solutions have been commonly attempted. Standardized tests are frequently used when comparisons are to be made. These tests provide some guarantee that the questions and marking will be common for all students. External examinations represent another solution that is found in Ontario and in New York State. The problem of the marks on local examinations still remains.

In an effort to make comparisons more legitimate, a variety of techniques for transforming and scaling marks have been suggested. None of the techniques can eliminate certain basic problems. No perfect test of any kind has yet been devised. Even the most sophisticated of standardized tests have errors of measurement. There is no absolute standard,

such as an inch or a pound. The problem might be compared to that of measuring with a rubber ruler. It would be impossible to keep a rubber ruler at exactly the same tension for all measurements, sometimes it would stretch a little more than other times, sometimes it would contract a little more. While, in the long run, the errors would tend to cancel out each other, the individual measurements could never be considered absolute, only approximations. No scaling system or method of transforming scores can compensate for measurement weaknesses of the tests themselves. All transformations must be applied on the assumption that the tests are acceptable measurements that need to be brought to some common base line to make results comparable.

Transformations and Scaling of Marks

Two kinds of measurements are useful when talking about a group of scores, measures of central tendency and measures of dispersion. Measures of central tendency provide information about the "centre" of the group. The mean (that is the arithmetic mean, sometimes called the average) and the median (that score above which 50% of the students fall and below which 50% of the students fall) are the most common measures of central tendency used with marks.

These measures provide a number which can be used to represent the whole group. The scores may be widely scattered about this central point or tightly bunched about it: a measure of dispersion provides information about the grouping of the scores around the measure of central tendency. The standard deviation is one measure of dispersion, and is used when the arithmetic mean has been adapted as the measure of central tendency. The interquartile range (the difference between the scores at the 25th percentile and the 75th percentile) is used in conjunction with

the median. The mode and the range of scores, other measures of central tendency and dispersion, provide less information about the scores made by a group of students and are of limited statistical value.

Given two groups of scores their total number cannot be changed. Transformations can only accomplish two things: "centre" both scores at the same point and/or make their dispersions equivalent.

In the distant past the mark of fifty was granted some mystical power of dividing success from failure. Certainly some criterion has been, and will continue to be, used for dividing students into two groups, passes and failures. The arbitrariness of this division cannot be emphasized too much however. Following an examination the teachers may rate it as "a difficult examination" or "an easy examination". Such statements are made in terms of the actual scores obtained by the students with reference to some undefined performance that was expected. Any transformation will bring home dramatically the arbitrariness of passing and failing standards.

Transformations are required because different examinations have essentially set different standards. Before a transformation should be attempted, one must consider what purpose the transformation will serve. Some subjects, such as mathematics tend to have widely dispersed scores while other subjects such as English Composition tend to have narrowly dispersed scores. Thus a student who is excellent in both subjects will find his mathematics grade of more value to his average than his English grade; the converse is of course true for the student who is poor in both subjects. If you wish to have all subjects equally weighted you must perform a transformation that will equate their dispersions.

In any given year, the means and/or the medians for different subjects will vary. It can be logically argued that all students should

have an equal chance of obtaining a mark of say 63, in all subjects. Where the mean or median is high, this mark will be much easier to obtain than when the mean or median is low. To provide an equal opportunity in all subjects, some kind of transformation affecting mean or median is required.

It will be immediately obvious that as you perform these transformations or scaling operations, the proportion of students falling below a mark of 50 will change. Some transformations are indeed worked out using a number other than 100 as the total possible score. In such a transformation a score of 50 would of course have an entirely different meaning than it had when a student was being marked out of a hundred.

The transformation that converts marks to "T scores" provides a very good example of some of the advantages and problems in a transformation. Even if the initial marks are not normally distributed this transformation redistributes them so that they are normally distributed. This transformation provides a mean of 50 and a standard deviation of 10. When marks are converted to T scores it is possible to operate with them using powerful statistical techniques. On the other hand such a transformation faces one squarely with the problem of how many students ought to fail for the cutting point must be arbitrarily set. A second problem can be easily demonstrated with such a transformation. Let us assume that all Grade 9 students in school "X" received the same mathematics examination and it was marked according to a rigidly specified marking system. Let us then suppose that all the Grade nine students in school "Y" had a different mathematics examination that also was marked according to a rigid marking system. By converting the results of both examinations to T scores you have essentially said that the students in the two schools were equal, that the only differences were due to differences between the examinations and marking systems. Converting their marks to T scores gives both groups the same mean and the same distribution of marks.

These points direct us to a consideration of the common elements that are required before a comparison can be made. Thus in school "X" we want to compare the performances of students on Mathematics and English tests. There is some immediate basis for making comparisons among students with respect to either English or Mathematics. Both examinations have been written by the same group of students, so there is some justification for transforming the scores on the assumption that students will show the same range of ability in both subjects, regardless of whether the examinations were easy or difficult. Even if this assumption cannot be met, equal treatment for all students on all their examinations, within a given setting, can be justified. Note carefully that if you were trying to compare these students on these examinations to another group of students in school "Y" who had written a different examination there would likely be no basis for assuming either that the two groups had the same range of ability or that their mean performance was equivalent. Only an external test marked to a common standard could provide the basis for such an assumption. Then we could compare different groups of students who had written the same examination.

It is obvious that scaling and transforming marks are valuable but there are certain limitations in such schemes. Transformations can be used as the basis for providing, to all students within a given school, the opportunity of being evaluated according to a common, known, standard on all examinations. For the final, external examinations at the end of secondary school one can safely assume that with so many students representing such a wide range of ability and background their performance will be normally distributed, providing a sound basis for transforming scores following the marking. Such a transformation would provide a

common standard for all students in all subjects; scholarship students would not be penalized for having taken courses that either had a low mean or a narrow range of scores.

The Mean or the Median?

Which is the most appropriate measure of central tendency, the mean or the median? Each measure provides one kind of picture of the "centre" of the group. A hypothetical example will demonstrate the differences. Figures 1 and 2 display two frequency polygons. These records of tire life were obtained by testing large numbers of tires from both companies. As can be seen, the mean tire life for an "Everwear" tire is 16,000 miles while the mean tire life for a "Neverfail" tire is 18,000 miles; conversely the median tire life for an "Everwear" is 18,000 miles and the median life for a "Neverfail" is only 16,000 miles. Two problems can be posed. If you wanted to buy only one tire for your car which brand would you purchase? Which brand would you purchase if you wanted to buy several hundred tires for a fleet of cars? The decision to purchase a single tire would be best based on the median. The median for "Everwear" is 18,000 miles, that is to say, half the "Everwear" tires last for more than 18,000 miles; in contrast, half the "Neverfail" tires will last for less than 16,000 miles. The decision to buy a single "Everwear" tire is based on the fact that you would have a fifty-fifty chance of getting a tire that would last for at least 18,000 miles. In buying for a fleet of cars, the mean would provide the best basis for your decision. Since a large number are being purchased, you would be more interested in the total wear of ALL the tires. The best estimate of this figure is provided by the mean, that is, the average mileage.

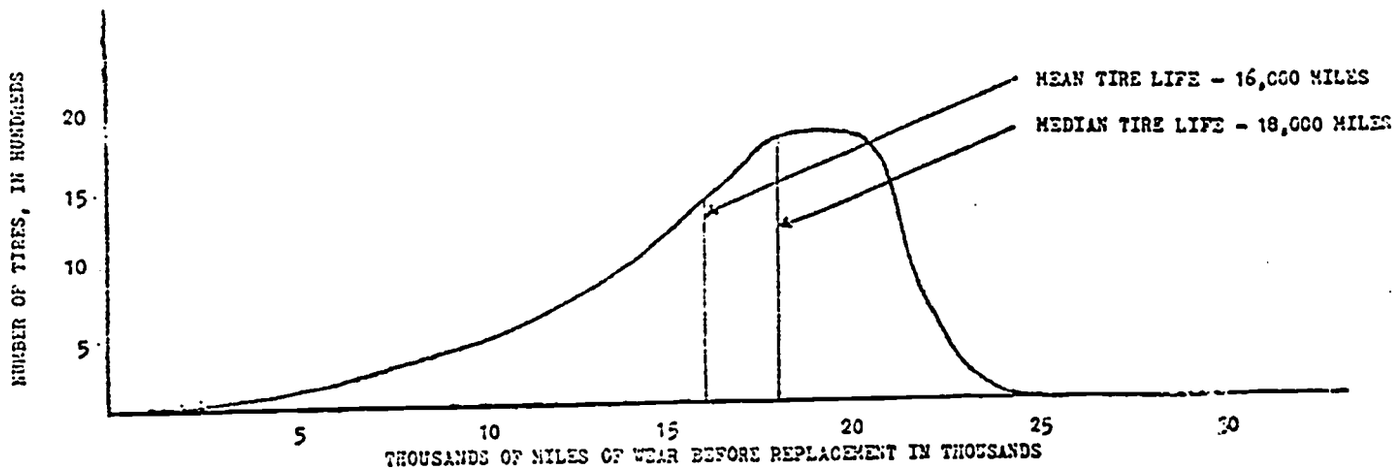


Fig. 1. Tire life of "Everwear" tires.

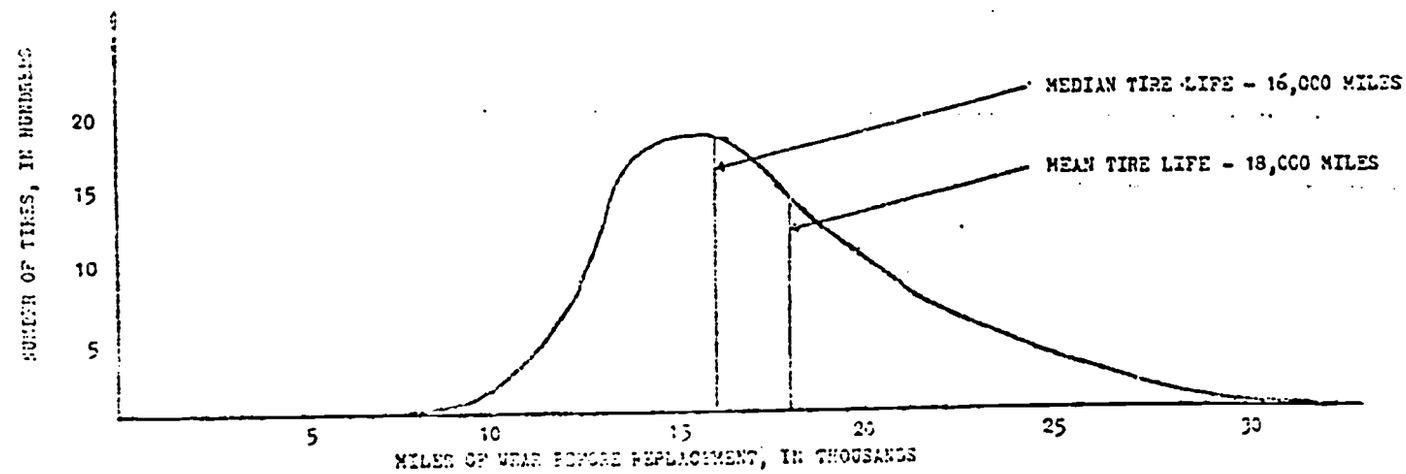


Fig. 2. Tire life of "Neverfail" tires.

In a normal distribution, the mean and the median are identical. In most classroom situations, however, the distribution, as in our example, is "skewed"; the graph gives the impression of a normal curve slightly squashed to one side and the mean and the median are different. The example suggests that if one single measure is desired to represent the whole group, the mean would be most appropriate. In comparing two groups you would likely base your comparison on means. If you were attempting to look at a single individual in comparison with his group, the median would seem to be more appropriate.

Percentiles, Ranks or Standard Scores?

Ranks express order, using the highest, best, or most noteworthy performance as the top or first place. An individual who comes tenth has been surpassed by nine other people. This statement of rank is meaningless unless we know the number of people involved. To come tenth among a group of eleven students is quite different from coming tenth among a group of one thousand.

Percentiles represent one attempt to standardize the procedure of ranking for percentiles are ranks converted to a base of 100. An individual at the 85th percentile surpassed 85% of the students he was compared to but in turn did more poorly than the top 15%. Percentiles indicate relative position in a group without reference to the number of people in the group. Percentiles do not give any information about the individual's performance. They do not report whether the student had a high or low score. They only report an individual's rank with reference to a comparison group.

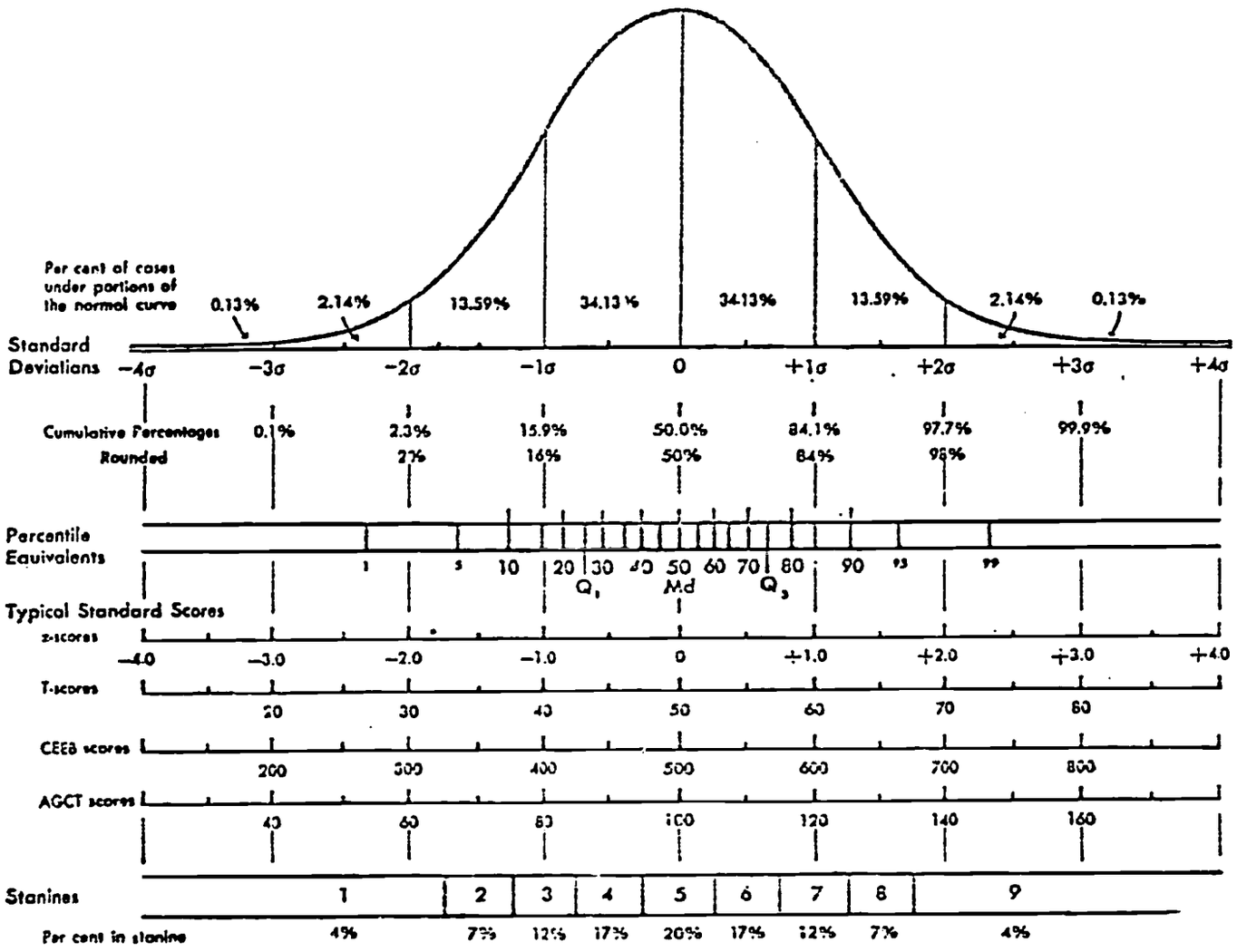
Percentiles have enjoyed considerable popularity because they seem easy to explain and understand. A student whose results place him

at the 48th percentile can be told that 52% of the students made a "better" score and 48% of the students made a "poorer" score.

Figure 3 presents a normal curve. This is the kind of distribution of scores that usually occurs when a large number of students take the same examination. As can be seen, the majority of scores cluster around the mean. A relative small number of scores are scattered in the extremes. The dramatic effect this has on percentiles is illustrated by the line "Percentile Equivalents". As can be seen, a small difference in score separates people at the 50th and 60th percentile, while a large difference in score separates the people at the 85th and 95th percentile. The use of percentiles tends to be misleading because scores are compressed around the centre point. Examination of the chart shows that the 75th percentile is much closer to the median (50th percentile) than might at first be expected.

To avoid this misleading characteristic of percentiles, a variety of standard scores have been developed. A standard score is obtained through a conversion. The mean and standard deviation are arbitrarily determined in advance (i.e., a number to represent the "centre" score and a number to represent the "clustering" of scores are selected). The chart shows, as examples, T scores with a mean of 50 and a standard deviation of 10, and College Entrance Examination Board scores with a mean of 500 and a standard deviation of 100. Standard scores have the advantage of providing a common base to which marks can be converted. Because the common base has a statistical foundation, a wide variety of comparisons can legitimately be made. A major drawback is the amount of calculation involved in the conversion.

Ranks and percentiles readily convey an individual's standing with respect to his classmates. They do not provide an adequate basis for



NOTE: This chart cannot be used to equate scores on one test to scores on another test. For example, both 600 on the CEEB and 120 on the AGCT are one standard deviation above their respective means, but they do not represent "equal" standings because the scores were obtained from different groups.

Fig. 3. Chart illustrating the equivalence of various scores related to the normal curve, adapted from Test Serv. Bull., No. 48, of The Psychological Corporation, p. 8.

comparisons, even among an individual's different school subjects because this is not an equal interval scale. Standard scores are essential for making detailed analyses and comparisons of marks but are time consuming to apply. If the original marks of the students tend to pile up at one extreme, the transformation to standard scores is appropriate only if one can accept the assumption that this distortion is a function of the test and not of the students' ability. The conversion to standard scores essentially places the results in a normal distribution.

As with the mean and median, one's choice depends on the use that is going to be made of the results.

Conclusion

A final note of caution should be added regarding normal distributions. Statisticians look for groups numbering in the thousands when seeking to determine whether observations are normally distributed. The application of the normal curve to small groups carries with it a variety of hazards because just as individuals vary so do small groups. The larger the group the more closely it resembles the whole population. Great variation can be expected among students in a class, smaller variations can be expected among schools. All this suggests that equivalence cannot be assumed among groups. While it is possible to argue for using the same scale of transformation for marks in every school within a system, it would then be necessary to provide a different score as the failure mark for each school to keep the standards comparable.

Marks are numbers, numbers that are added, averaged, and compared. Though most educational decisions essentially reflect value judgements, these judgements are typically interpreted in the form of these numbers, or marks. Transformations and the scaling of marks can provide a just basis for comparisons that would otherwise be inappropriate. There are

many problems related to transformations and the scaling of marks, not the least of which is the amount of work required for some of these transformations. Many of the problems are not directly related to scaling or transformations, instead certain basic educational problems which are often ignored find themselves dramatized and brought to the forefront when such procedures are adapted.

References

For a straightforward examination of the statistical concepts found in this paper and formulas relating to transformations, the reader is referred to Walker and Lev (1958). Test Service Bulletin No. 48 provides a concise description of different methods of expressing test scores.

Walker, Helen M., & Lev, J. Elementary Statistical Methods. New York: Henry Holt, 1958.

The Psychological Corporation. Methods of Expressing Test Scores. Test Serv. Bull., 1955, No. 48.