ED446112 2000-09-00 Summarizing Change in Test Scores: Shortcomings of Three Common Methods. ERIC Digest.

ERIC Development Team

www.eric.ed.gov

Table of Contents

If you're viewing this document online, you can click any of the topics below to link directly to that section.

ERIC Identifier: ED446112
Publication Date: 2000-09-00
Author: Russell, Michael
Source: ERIC Clearinghouse on Assessment and Evaluation College Park MD.

Summarizing Change in Test Scores: Shortcomings of Three Common Methods. ERIC Digest.

ACCESS ERIC 1-800-LET-ERIC

The reliance on test scores to assess the impact of schools on student achievement has increased sharply during the past decade. This increase is reflected in the number of states that employ testing programs to hold schools, teachers and students accountable for improving student achievement. According to annual surveys by the Council of Chief State School Officers (1998), 48 states use statewide tests to assess student performance in various subject areas and 32 states currently use or plan to use test scores to determine whether to grant diplomas. In addition, many educational programs, including charter schools, depend on test scores to demonstrate the success of their programs. In many cases, however, educational leaders employ overly simplistic and, sometimes, misleading methods to summarize changes in test scores.

Educational leaders, institutions and the popular press have employed a variety of methods to summarize change in test scores. To aid public understanding of test score reporting, this Digest introduces the advantages and disadvantages of three commonly used methods:Change in Percentile Rank, Scale or Raw Score Change, and Percent Change. A separate Digest and article (Russell, 2000) describe two alternate approaches for summarizing change and demonstrate how a third method, namely Expected Growth Size, can be used to summarize change for vertically equated norm-referenced tests.

METHOD 1: CHANGE IN PERCENTILE RANK

As the name implies, the change in percentile rank method focuses on the increase or decrease of the mean percentile ranking for a group of students. Change in test performance is determined by subtracting one mean percentile rank from another. Since most people are familiar with percentile ranks and the mathematics required for this method are relatively simple, this method is often employed to express change in test scores to the general public. As an example, Edison Schools used this method to report that students, on average, "are gaining more than 5 percentiles per year " (1999, p. 2). With this approach, however, two problems can arise. First,calculating the mean percentile rank based on individual percentile ranks can provide an inaccurate estimate of a group's mean performance. Second, due to the unequal intervals separating percentile ranks, changes in mean percentile ranks represent different amounts of growth at each point on the scale.

Why Averaging Ranks to Determine the Mean Group Rank Is Misleading

As Glass and Hopkins (1984) point out, percentile ranks are ordinal measures in which the amount of the trait measured represented by each one-point increase in rank varies at each point on thescale. Due to the unequal intervals between each percentile rank, using each individual's percentile rank to determine the mean percentile rank can result in an inaccurate estimate of the group mean. As an example, Table 1 displays the ranks and corresponding times for five sprinters. In this example, the trait measured is running speed. Although the mean rank for the group of sprinters is three, the mean speed for the group is much slower than the time recorded by the third-place finisher. Due to the unequal time intervals separating each rank, simply using ranks to determine the mean rank results in an inaccurate estimate of the group's mean running speed.

See TABLE 1 at end of digest.

To overcome the problem associated with performing mathematical operations with ordinal ranks, one should use the score associated with each percentile rank to determine the group mean and then determine the percentile rank (PR) that corresponds to the group's mean score. As Table 2 demonstrates, using students' standard scores or Normal Curve Equivalent (NCE) scores to calculate the group mean and then finding the percentile rank that corresponds to either mean yields a mean percentile rank that provides a more representative estimate of the group's mean achievement.

See TABLE 2 at end of digest.

In Table 2, we see that the mean percentile rank of 27.2 corresponds to a standard score of approximately 163. However, the mean standard score for the group of students is actually 159.3, which corresponds to a percentile rank of approximately 21. In this example, the mean percentile rank over-estimates the group's mean language achievement and implies that on average students are performing six percentiles higher than their mean standard score indicates.

Why Differences Between Mean Percentiles Can Be Deceptive

Even when the mean percentile rank is calculated using students'standard scores or NCEs, summarizing change in score as the difference between the group's mean percentile ranks can be misleading because it implies that the same amount of change represents the same amount of growth at all points on the percentile scale. In reality, the further a percentile rank deviates from the mean, the more a student's score must increase for his or her percentile rank to increase. This relationship is a direct result of the distribution of scores within the normal curve. In a normal curve, a disproportionate number of people score in close proximity to the mean. As a result, a small change in a person's test score close to the mean will result in a much larger change in his or her rank relative to other test takers as compared to the same change at the extremes of the distribution.

As an example, on the Iowa Test of Basic Skills (ITBS) Language sub-test, the standard score for a third-grade student must increase seven points in order to move from the 10th to the 20th percentile. However, to move from the 50th to the 60th percentile, a student's standard score only needs to increase four points. Similarly, a ten-point increase at the 10th percentile represents a change of about .44 standard deviations. However, a ten-point increase at the 50th percentile represents a change of only .25 standard deviations. Depending upon a student's percentile rank, this method exaggerates or understates change in student performance.

METHOD 2: SCALE OR RAW SCORE CHANGE

A second method used to examine change in test performance focuses on change in scale scores or raw scores. For this method, the mean score for prior years is subtracted from the mean score for the current year. The result represents the change between the two time periods.
As with change in percentile ranks, this method is appealing because it involves basic arithmetic. However, there are several drawbacks. Foremost among them is that when raw scores are used to determine change, it is difficult to compare change across tests that have different score ranges.

As an example, a third-grade mathematics test may contain 30 items while the fourth-grade test contains 40 items. Each grade level may experience a five-point increase in its mean score. But, since the tests differ in length, these five point increases do not have the same meaning for both tests. For the test containing 30 items, a five-point increase suggests that students are answering about 17% more items correctly. For the test containing 40 items, a five-point increase indicates that students are answering only about 13% more items correctly.

One solution to this problem is to focus on change in scale scores. However, this too presents problems. Although most norm-referenced standardized tests report scores from different grade levels on the same scale, the standard deviations for the grade levels tend to differ. For this reason, a five-point change for two different grade levels represent different amounts of change within each grade level. The problem is similar to that experienced with the change in percentile rank method. As an example, for the ITBS Language test, the standard deviations for grades 3 and 4 are 19.05 and 24.25, respectively. Thus, five-point increases in the standard scores for grade 3 and 4 represents changes of .26 and .21 standard deviations, respectively. Clearly, a five-point change represents more growth relative to students within grade 3 than within grade 4.

METHOD 3: PERCENT CHANGE

Further distortion is caused by summarizing change in test performance as a percentage of prior performance. As an example, some charter schools have reported

20 to 30% improvements in their test scores. To obtain these figures, test scores for the current year are divided by past test scores to yield the percent change. In the best case scenario, this method focuses on percent change of standard scores or NCEs. In the worst case, this method focuses on percentile ranks. In all cases, however, this method assumes that the scores used to determine percent change are on a ratio measurement scale. Both standard scores and NCEs, however, are at best interval measures while percentile ranks are clearly ordinal measures. As Glass and Hopkins (1984) explain more fully, ratios based on interval and ordinal measures are meaningless.

The percent change method is particularly deceiving when initial performance is low. Take, for example, two schools that both experience five-point increases in mean scores. School A saw its mean scorei ncrease from 20 to 25, while the mean score for School B increased from 50 to 55. Although both schools experience the same amount of change in their scores, the percent change method suggests that scores for School A improved 25% while School B improved only 10%. Once again, although the arithmetic is simple, the percent change method produces a statistic that is both difficult to interpret and misleading.

SUMMARY

As Willet (1988) explores more fully, all methods of summarizing change may be threatened by low score reliability. However, even when scores are sufficiently reliable, the three methods described above can result in misleading estimates of score changes. In general, these methods are insensitive to the measurement scale on which scores are expressed and perform mathematical operations that are inappropriate for these measurement scales. These methods also assume that the same size difference represents the same amount of change at all points on the scale. As demonstrated above, this assumption is false. For these reasons, all three methods should be avoided when summarizing change in test scores. As is explained more fully in a separate Digest and article (see Russell, 2000), preference should be given to methods that yield standardized estimates of score changes which have the same meaning at all points on the measurement scale and which can be compared across tests and grade levels.

REFERENCES:

Council of Chief State School Officers (1998). Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance. Washington, DC.

Edison Schools. (1999). Second Annual Report on School Performance. New York, NY.

Glass, G. & Hopkins, K. (1984). Statistical Methods in Education and Psychology, 2nd Edition. Boston, MA: Allyn and Bacon.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B.(1996). ITBS: Norms and Score Conversions with Technical Information. Itasca, IL: Riverside Publishing.

Russell, M. (2000). Summarizing change in test scores part II: Advantages of expected growth size estimates. Practical Assessment, Research and Evaluation, 7(6). [Available online: http:// ericae.net/pare/].

Willett, J. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), Review of Research in Education 15 (pp.345-422). Washington, DC: American Educational Research Association.

This digest is is based on Russell, Michael (2000). Summarizing Change in Test Scores: Shortcomings of Three Common Methods. Practical Assessment, Research & Evaluation, 7(5).

●

Available online: http://ericae.net/pare/getvn.asp?v=7&n=5.

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract ED99CO0032. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or the U.S. Department of Education. Permission is granted to copy and distribute this ERIC/AE Digest.

TABLE 1

Rank and Finishing Time for Five Sprinters

Rank	Time
1	10.2
2	10.3
3	10.5
4	11.2
5	12.4
Mean	
3	10.9

TABLE 2

Standard Score, NCE and Percentile Rank for Students on the Third Grade Iowa Test of Basic Skills Language Test*

Student	Standard Score	NCE	PR

1	127	1	1
2	143	10	3
3	151	23	10
4	161	35	24
5	164	39	30

Title: Summarizing Change in Test Scores: Shortcomings of Three Common Methods. ERIC Digest.
Note: For a related Digest, see TM 031 856. Based on "Summarizing Change in Test Scores: Shortcomings of Three Common Methods" by Michael Russell, in "Practical Assessment, Research & Evaluation" 7(5).
Document Type: Information Analyses---ERIC Information Analysis Products (IAPs) (071); Information Analyses---ERIC Digests (Selected) in Full Text (073);
Available From: For full text: http://ericae.net/pare/getvn.asp?v=7&n=5.
Descriptors: Achievement Gains, Change, Evaluation Methods, Scores, Test Interpretation, Test Results
Identifiers: ERIC Digests, Percentile Ranks
###

▲

[Return to ERIC Digest Search Page]