MEASURES OF SCHOOL PERFORMANCE

James S. Coleman and Nancy L. Karweit
MEASURES OF SCHOOL PERFORMANCE

James S. Coleman and Nancy L. Karweit
-iii-

PREFACE

Under a Rand Corporation-sponsored program of research into
domestic problems, James S. Coleman, Professor of Social Relations at
Johns Hopkins University and a Rand consultant, was invited to spend
a period in Santa Monica working on questions that received national
attention with the publication of a report for the U.S. Office of
Education, *Equality of Educational Opportunity.* Professor Coleman
had been largely responsible for the latter report, and he is the
senior author of the present study.

What gave impetus to the present research was the main conclusion
of the government report that a child's performance in school was
primarily determined by the socioeconomic backgrounds of his family
and of his classmates' families. One could infer from this that,
within the normal range of variation, few of the factors that school
systems can control -- other than assignment of students to schools
and classrooms -- affect the performance of their students. Although
based on an extensive national survey, the government report pointed
up the dearth of thorough, consistent, and comprehensive data on
school inputs and outputs.

In response to these findings, The Rand Corporation undertook
to support a modest, initial inquiry into the relationships between
the inputs to a school system -- students, teachers, equipment, and
facilities -- and its output, as measured by the performance of stu-
dents on standardized achievement tests. An effort was made to examine
computerized systems for gathering educational data as a by-product
of regular school functions. This work was reported in *Multi-Level
Information Systems in Education*, The Rand Corporation, P-4377, June

A further question, how to improve existing methods of educa-
tional data collection and evaluation, is addressed in the present
Report, which was drafted while the authors were at Rand in the summer
of 1969. In it, the authors discuss the inaccuracies of using stan-
dardized tests (which were designed to measure individual student
performance) to evaluate the functioning of schools and school
districts, the impact of special programs, the comparative effects of home and school on student achievement, and other aspects of school performance.
SUMMARY

Standardized tests, designed to measure individual student performance, are being used increasingly in efforts to evaluate the functioning of schools and school districts, the impact of special programs, the comparative effects of home and school on achievement, and other aspects of school performance. Current usage and interpretation of test scores may lead to incorrect inferences, endangering the acceptance of this use of test scores. This Report examines the most common inaccuracies in present reporting practices and suggests ways in which current test results might be made more reliable and useful.

The scope of existing standardized tests is recognized as narrow, focusing mainly on the measurement of verbal and mathematical skills among the many qualities that schools seek to develop. Such areas as history and foreign language are not amenable to testing.

These available measures of school performance contain a major weakness, which arises from the fact that almost all standardized tests are expressed in relative terms — as grade equivalents or as percentiles. The performance of the standardizing population, usually the population of students at that given grade level, changes from year to year, making it difficult to compare performance at different time points and at different grades. The authors suggest the need for the development of absolute measures to facilitate these comparisons and for the development of a set of standards for test-norming procedures.

The results of standardized tests are expressed in a variety of ways — for example, as raw, standardized, percentile, or grade-equivalent scores. The Report examines problems associated with the use of each type of score for purposes of comparison.

Grade equivalents are seen as misleading. For example, results showing that Negro students fall farther behind in grade equivalents with each school year are contradicted by average percentile measures, which show both Negro and white averages as remaining at the same percentile in national distribution. Thus, a student who remains at
exactly the same position relative to others may, by the grade-equivalent measure, appear to be falling farther and farther behind. A student's grade-equivalent score is actually a relative score masquerading as an absolute one.

Percentile scores may also be misleading, but in a different way. Because of the bell-shaped distribution curve, many students are clustered closely in absolute score around the 50th percentile. At the 10th or 90th percentiles, students are more widely dispersed. Hence, an absolute score change from the 97th to 98th percentile approximates a change from the 50th to the 57th percentile.

Yet another area of confusion flows from the attempt to use ability or intelligence measures, scored in grade equivalents, to determine whether students are underachieving or overachieving. In criticism of this procedure, the authors stress that the faculties measured in ability tests do not necessarily correspond to those measured in achievement scores. Abilities tend to be measured in one or two dimensions, while achievement may be demonstrated in a variety of dimensions. Also, ability tests measure qualities that are subject to change by schools; students can be taught to score better on most ability tests. Raw scores on ability or intelligence tests show as much increment in performance with increase in grade level as do tests of specific achievement.

Current interest centers on assessing the effects of schools on achievement, but such inferences are hardly to be drawn with assurance from current measures. Much change in achievement derives from factors outside the school, and the amount attributable to school remains unknown. In the absence of a model showing how school factors combine with others in affecting performance, one cannot accurately predict the "expected" movement in achievement in a school whose children differ in nonschool factors; the authors suggest indirect approaches to that end. The difficulty of determining growth due to school factors by comparing average performance at different grade levels is posed by the possibility of population differences between the grades being compared. To allow for population changes and render the test scores of children at different grade levels comparable, the Report
proposes reweighting test scores at later grades before averaging, in such a way that they will closely reflect the population at the earliest grade. In practice, the first-grade test score has a special role. Subject to its degree of reliability, it embodies the child's background factors without containing school influences. The gain or loss in standard scores in particular educational settings from this first-grade score can be attributed to those educational settings. But even this assessment is subject to distortion as a result of changes in the school population under consideration.

Distribution of performance is an important but seldom-assessed measure of school effectiveness. A primary entrant population with diverse backgrounds will display a large standard deviation. If this deviation decreases over the years while the average remains constant, both high-performing and low-performing students are moving closer to the average, suggesting the effect of the educational process. Variations from these conditions may suggest specific effects of education, such as enhancement of the capabilities of low-performing entrants.

None of the measures discussed provides more direct evidence of the degree of equality of opportunity than does the change in the achievement of individual students, which shows the ability of a curriculum to reduce the dependence of the final on the initial level of achievement. Many schools will shortly be equipped to present this information; this report provides a formula, based on differences in each individual's test scores at two points in time, which promises to capture the movement in achievement, both up and down.

The more sophisticated data systems of the future will ultimately permit us to draw direct inferences about the effects of home and other influences on student performance as distinct from those of the school or program, and will greatly simplify the task of those who must assess the role of schools. Until then, the indirect inferences obtainable with present data systems must be rendered as reliable as possible by the use of correct measures and consistency in reporting procedures.
CONTENTS

PREFACE ........................................... iii
SUMMARY ........................................... v

Section
I. INTRODUCTION ..................................... 1
II. THE TESTS ......................................... 3
III. THE MISLEADING USE OF GRADE EQUIVALENTS ............ 7
IV. THE DIFFERENCE IN SIZE OF THE PERCENTILE AT 50% AND 90% . 17
V. ABSOLUTE CHANGES OVER A GIVEN PERIOD ................. 21
VI. THE INCORRECT USE OF ABILITY MEASURES TO ESTABLISH EXPECTED ACHIEVEMENT LEVELS ................. 23
VII. INFERENCES ABOUT SCHOOL EFFECTS FROM SCHOOL PERFORMANCE MEASURES .......................... 27
VIII. STANDARDIZATION FOR POPULATION CHANGES ............ 31
IX. THE DISTRIBUTION OF PROGRESS ........................ 33
X. TOTAL MOVEMENT IN ACHIEVEMENT ........................ 35
XI. FURTHER MEASURES IN ADVANCED DATA SYSTEMS .......... 37
I. INTRODUCTION

Increasingly, the results of standardized tests are being used to evaluate schools and curricula. For example, in response to public pressure, a number of school districts are now publishing school-average test scores, and some states have begun to give out average test scores for school districts. Federally-sponsored and state-sponsored experimental programs ordinarily are evaluated largely on the basis of such standardized test scores.

Designed to measure individual performance, these tests until recently were used mainly for student course placement. In their application to the evaluation of schools and educational programs, however, errors have been made that have led to incorrect inferences about the programs and their effects, and, partly for this reason, school administrators are offering strong resistance to this use of test scores. However, if the tests are valid for measuring the performance of individual students, they are valid also for measuring the school's program — provided the correct comparisons are made. It is important, therefore, to examine past practice for some of the errors and inaccuracies that lead to invalid conclusions, and to establish appropriate comparison measures that will render the test scores useful for judging the performance of a school or a program.

The various uses to which test results have been put in the measurement of school performance and program performance in recent years can be grouped under three headings:

a. Depicting the level of functioning of students in an already existing program, school, or school district. This is being done with the publication of school-average scores at different grade levels in various cities, and of school-district averages within states.

b. Describing the impact of a special program with a definite starting point. Examples of this use abound in the evaluation of More Effective Schools (New York City), Project Head Start, and many other experimental programs that are federally- or state-financed. Inquiries into the effectiveness of newly-instituted school
integration, as in White Plains, Buffalo, and Berkeley, are further instances.

c. Using test scores as "dependent variables" in research aimed at separating effects of student background from those of school environment. This is exemplified in Equality of Educational Opportunity, in the "Plowden Report" in England, in the international assessment of achievement in mathematics, in Racial Isolation in the Public Schools, and in many other recent analyses.

The test score for the individual student, the classroom average, or the school average at one point in time (or sometimes the difference in scores at two points in time) is taken as the dependent variable; then, various "independent variables" are introduced in a statistical analysis to account for variation in these scores between different students, classrooms, or schools.

In all these such questions arise as: What score should be used (raw score, standardized score, percentile, or grade equivalent)? How should achievement increments, or "growth," be measured? What kinds of comparisons among schools or students are necessary to permit valid inferences about school effects? Because even a seemingly straightforward publication of test scores for schools can be misleading, it is important to answer these questions and clarify some of the issues involved.

We are not concerned here with designing the ideal measures of school performance, for such a design could not be carried out in existing schools. Instead, taking as given the kinds of tests that schools use, the frequency with which they are administered, and their various uses, we will examine how current practices might be appropriately modified.

---


II. THE TESTS

At the outset, it is important to recognize that the tests used to measure the performance of districts, schools, or programs tend to be narrow in scope, most often covering only two areas of student performance: verbal and mathematical skills. Presumably, there are two reasons that the testing does not cover a broader range: Some areas (e.g., social studies, history, and foreign languages) do not have the consensus about goals that exists with regard to verbal and mathematical skills; and for some of the other school goals, such as motivation or attitude toward learning, the conventional paper-and-pencil tests are poor measuring devices, and no good alternatives have yet been found. In examining verbal and mathematical skill tests as measures of school performance, therefore, we must bear in mind that such tests were designed only to measure two central skills among those that schools seek to develop, and that they cannot measure all the qualities that schools are intended to develop in their students.

The second thing to understand about current testing is that the tests by which schools measure verbal or mathematical achievement are standardized, not on some absolute scale, but relative to other students, and their results are expressed in relative terms. Often, the expression is in grade-level equivalents (e.g., a student's reading may be equivalent to that of the average child in the United States beginning the 6th grade). Sometimes the scores are expressed in terms of a child's percentile position (or, for a school average, the percentile position of the average score in the school) relative to all children at the same grade level (or all schools) in the country. Or they may be expressed in standard scores based on the distribution of children at the same grade level. The standardizing population ordinarily is that of all students in the United States at the given grade level, creating "national norms," but there are exceptions (such as "large-city norms").

Although for our present purpose we will take these tests as given, it may be well to point out that the development of absolute
measures would greatly facilitate comparisons. The use of relative measures means that, as performance levels change in the norming population, either the norms must also be changed (which would preclude over-time comparisons), or the relative measures will no longer accurately express the child's or school's relative standing in the current population. And even if performance levels do not change, the fact that norms are always based on a given grade level makes it difficult to measure relative growth by comparing performance at different grade levels, as will be shown presently. Absolute measures could most easily be developed in mathematics, and with somewhat more difficulty in verbal skills.*

A final comment on test norming as it is currently carried out: Because the national norms of widely-used tests are the basis for comparisons and policy decisions, we need some standard for test norming to correct the present, grossly deficient procedures. The tests now used often are normed on small, unrepresentative samples; and even the fact that achievement increments are not equal over the months of the year is not taken into account in the establishing of grade equivalents, or percentiles, for different testing dates. Linear interpolation between yearly points is sometimes used to establish norms for different testing dates in the year. This interpolation gives results which make impossible percentile comparisons of test scores taken at different points in the school year. So long as test scores were being used only to measure individual performance, this source of error was obscured by the measurement error of the test; but when they are used to establish school averages, this measurement error is sharply reduced, leaving the interpolation error exposed. Schools throughout California, for example, administered standardized tests in the fall, and a second form of these tests the

---

*This greater difficulty would arise from the fact that language is a product of the particular population that uses it, and therefore differs from population to population. Nevertheless, it would be possible to design measures in which word difficulty was defined by frequency of usage in a well-defined type of textual material, and difficulty of textual passages was defined by linguistic structure and word difficulty.
following spring, according to which percentile scores were generally found to have fallen. This result, interpreted as a true result, very likely was due to interpolation errors in test norming.

Deficiencies such as these confirm the need for a set of standards for test-norming procedures, to be developed by a federal agency such as the U.S. Office of Education or the National Bureau of Standards.
III. THE MISLEADING USE OF GRADE EQUIVALENTS

Test scores sometimes are expressed by year in school. That is to say, a child's score may be reported as being "5.3 years," indicating that it is equivalent to the national norm three-tenths of the way through the fifth grade. Thus, if the child made such a score at the end of November of the fifth grade, he would be at about the national norm.

If one takes the tests and the norms as given, they can yield misleading results in comparisons of the growth of individuals or groups. Negro students, for example, are commonly shown to be below grade level, and to fall farther behind with each year in school, gaining only .6 to .8 of a grade equivalent per year, while whites gain about 1.0 grade equivalent, consonant with the national norm. The inference usually drawn from this smaller yearly increment is that the schools (or, less often, the home backgrounds) are causing Negro students to drop further and further below whites of the same age. This may then be taken as prima facie evidence either that the schools are performing more poorly for Negro students than for whites or that Negro students are unable to progress as fast as whites.

If, however, one uses the average percentile instead of the average grade equivalent as a measure for these same test scores, the inferences drawn may be the exact opposite. The scores will then show that the Negro average remains at the same percentile in the national distribution and that the white average also remains at the same percentile. Scores of Negroes could even show an increase in the percentile position, while their grade-equivalent scores were showing a smaller yearly increment than those of another group with higher percentile scores whose percentile position was not changing. Under the circumstances, which measures should be used, and what inferences could be drawn from them?

To answer these questions we must examine the source of the anomaly: the fact that at early grades in school the distribution of scores of children in a given grade (or of a given age) covers a smaller span of grade equivalents (that is, average scores at each
grade) than at higher grade levels. This may be illustrated by the use of three tests used in *Equality of Educational Opportunity* that were linked at grades 6, 9, and 12 to allow direct comparisons. Table 1 (p. 9) shows the difference in the conclusions to be drawn from comparisons of grade levels and those of percentiles. The population is whites in the urban Northeast, and we may take the average score of sixth-graders in that population as the norming population, defining the 6.0, 9.0, and 12.0 grade-level scores. The table shows, at each of these grades, the grade equivalent of students who were at the 16th and 84th percentile in this distribution. (These percentiles are one standard deviation below and above the mean, respectively, approximating a normal distribution.) By a grade-equivalent measure, stated in years and tenths-of-years, the students at the 16th percentile are shown to be falling farther and farther behind, although they remain at the same percentile. At grade 6, they are 1.5 years behind in verbal ability, 2.2 years behind in reading comprehension, and 1.8 years behind in mathematics achievement; by grade 9, they are 2.1, 2.5, and 3.0 years behind; by grade 12, they are 3.2, 3.5, and 4.8 years behind. The same is true in reverse for those at the 84th percentile, who are farther ahead in grade equivalents at grade 9 than they were at grade 6. (Their grade equivalent at grade 12 cannot be determined, because the average cannot be accurately extrapolated that far.)

In other words, a student who over these six years remains at exactly the same position relative to others at the 16th percentile appears, by the grade-equivalent measure, to be falling farther and farther behind. His average grade increment in mathematics between the 6th and the 9th grade is .6 grade per year, but between 9 and 12 it is only .4 grade per year. Yet 84 per cent of the students remain above him throughout these years. A student who remained exactly the same number of grade levels behind (i.e., whose grade increments were the same from one year to the next) would in fact be moving up in percentile position; for example, if he had been at the 16th percentile in verbal ability at grade 6, 1.5 years behind, and was still 1.5 years behind at the 12th grade, he would have had to rise to the
Table 1

Grade equivalents in years and tenths-of-years for grades 6, 9, and 12 for average and for students at the 84th and 16th percentiles, for verbal ability, reading comprehension, and mathematical achievement. Population: White students in urban Northeast (from Equality of Educational Opportunity Survey, September 1965), also used as the norming population.\(^a\)

<table>
<thead>
<tr>
<th></th>
<th>Equivalent grade level in years and tenths-of-years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.0</td>
</tr>
<tr>
<td><strong>Verbal Ability</strong></td>
<td></td>
</tr>
<tr>
<td>Average (norm)</td>
<td></td>
</tr>
<tr>
<td>Student at 16th percentile</td>
<td>4.5</td>
</tr>
<tr>
<td>Student at 84th percentile</td>
<td>7.8</td>
</tr>
<tr>
<td><strong>Reading Comprehension</strong></td>
<td></td>
</tr>
<tr>
<td>Average (norm)</td>
<td></td>
</tr>
<tr>
<td>Student at 16th percentile</td>
<td>3.8</td>
</tr>
<tr>
<td>Student at 84th percentile</td>
<td>8.7</td>
</tr>
<tr>
<td><strong>Mathematical Achievement</strong></td>
<td></td>
</tr>
<tr>
<td>Average (norm)</td>
<td></td>
</tr>
<tr>
<td>Student at 16th percentile</td>
<td>4.2</td>
</tr>
<tr>
<td>Student at 84th percentile</td>
<td>8.3</td>
</tr>
</tbody>
</table>

\(^a\)Data taken from Figures 3.121.1, 3.121.2, and 3.121.3 in Coleman et al., *Equality of Educational Opportunity*, pp. 274-275. To determine the 16th and 84th percentiles, scores are taken one standard deviation below and above the mean, an approximation which assumes a normal distribution of test scores. This may give minor differences from use of the true distribution, which do not, however, affect the inferences drawn from the table.
32nd percentile to do so, that is to say, he would have had to pass 19.05 per cent of the students who were above him at grade 6 in order to achieve this position.

Similar discrepancies arise in comparisons of schools. Suppose that school A's average in mathematical achievement is at the 16th percentile in grade 6, while school B's is at the 50th percentile.* The average grade level of school A is 4.2, and that of school B is 6.0, a difference of 1.8 years. If the percentile position of the average in both schools remains constant -- at the 16th and 50th percentiles -- the average in school A will show a grade level of 6.0 at grade 9 and 7.2 at grade 12, that is, a gain of only 1.8 years in the three years from grades 6 to 9, and of 1.2 years in the three years from grades 9 to 12. The average in school B will have gained exactly 3 years in each of these three-year periods.

One way toward understanding the source of this discrepancy is to look at a typical curve for such linked tests over a set of grades. It will have a form like that of Figure 1 (p. 11), which shows the mean curve (solid) and the 16th and 84th percentile curves (dashed) for grades 6, 9, and 12 for verbal ability, using the same data as Table 1. (The curve for grades between these is approximated by straight lines; and the mean curve is extrapolated back to grade 3.) Scale scores, derived from raw scores, are used for the vertical axis, and student grade level for the horizontal axis. To create the scale appropriate for linked tests of this sort, the raw score is so scaled that the variance among students at each grade level is constant. (In this test the scaling was not done perfectly; the variance increases slightly from grade 6 to 9 and from 9 to 12. For present purposes, however, we can assume that the scaling is correct.)

When tests are scaled to create constant variance at each grade level, they uniformly show a declining slope as years in school

*The 50th percentile is the median; for the purpose of using Table 1, we will assume that the mean equals the median and that the 50th percentile therefore is exactly at grade level 6, 9, or 12.
Fig. 1. Linked test scores with common mean and variance for grades 6, 9, and 12.
increase (see Figure 1). Starting at a given distance below the average thus means an ever-larger distance behind the average curve.

Referring again to Figure 1, we find the "grade-equivalent" measure for a student who is at the 16th percentile at actual grade level 6 by projecting from there, by way of line a, to the mean (or median) curve, and then reading off this point of intersection on the horizontal axis (in this case, 4.5). These "distances behind" (along lines b and c for grades 9 and 12, respectively) increase with the increase in actual grade level, though the "distances below" remain approximately constant. The student's grade-equivalent score is thus a measure, not of his performance relative to the average of others his age, but of the age level of those whose average performance equals his own.

A "year of growth" in reading at grade 12 is less, relative to the total distribution of 12th-graders' scores, than a year of growth at grade 6. A "grade-equivalent" score, therefore, means a different thing at every grade level. It does not compare the student to others of the same age or at the same actual grade level; it compares him to the average or median student at another grade level. It is a relative score masquerading as an absolute score.

The grade-equivalent score for the school, however, does show exactly what it purports to show: It uses as the yardstick the median student at different grade levels in the norming population, and picks out the grade level at which that median student receives the same score as the median student in the given grade in the given school. It thus shows the grade level at which the median student in the school is performing. But it is not appropriate for inferences about the effect of the school, or the performance of the school, or the rates of growth of children at different grade levels. The absurdity of making such inferences from grade equivalents is shown if one assumes that performance in verbal skills levels off shortly after the end of high school, say at age 20. This would produce a fixed difference in verbal skills, which would remain constant, so that, at age 40, the differences would be the same as at age 20. Yet by a measure of grade equivalents or age equivalents, those at the 16th percentile
would come to be 10, 15, and 20 "years behind" those at the 50th percentile as they reached age 30, 35, and 40, respectively, merely because the performance of the two groups leveled off at different points when schooling ended.

Another way of showing that one cannot accurately measure gain by comparing gains in grade equivalents is as follows: The true measure of an individual's or a school's gain between two points in time is the slope of the line joining the two points on a graph of scale score vs. time, like that of Figure 1. If the gains for two groups have the same slope, the lower group will have gained less in grade equivalents than the higher group so long as the median curve shows a decreasing slope with the increase of years in school, as in Figure 1. The lower group will be the same "distance below" at time 2 as at time 1, but a greater "distance behind." If the tests are correctly linked and scaled, however, with constant variance, and the percentile positions of the upper group remain the same, the percentile position of the lower group will remain the same also.

One way of looking at the changes that occur over the years of school is to modify Figure 1 by rescaling the scale score to stretch out the upper end, so that the mean is a straight line (see Figure 2). Instead of a constant variance, this new scale score would have an increasing variance with increasing mean. In such a graph, a particular "growth rate," or slope, is associated with each percentile position in the distribution; if a person remains at the same percentile throughout school, his growth is characterized by a given line. In Figure 2, those at the 84th percentile grow fastest; those at the 50th percentile have a smaller slope or growth rate; and those at the 16th show a still smaller growth rate. Thus, a school whose average student begins in the first grade at the 16th percentile and at the 12th grade is still the school's average and still at the 16th percentile will have kept him at exactly the same level compared to all students; but the growth rates for that school are lower than for a school whose average student is at the 50th percentile.

This is not to say that a percentile score is the ideal comparison for most purposes, for it is not. Suppose that two schools showed
Fig. 2. Linked test scores with increasing mean and variance for grades 6, 9, and 12.
the same gain (i.e., the same slope in scale score between two grades), and that one school began at the 50th and the other at the 10th percentile. Then, if the school at the 50th went to the 60th percentile, the same gain in scale score would raise the school at the 10th only to the 15th percentile. This point will be discussed at length in the next section. For the present, we only need to recognize that the percentile score is a measure, as it should be, of amount below or above, rather than, like the grade-equivalent measure, of the amount behind or ahead.

Ordinarily, comparisons between schools or school districts consist of measurements at several grade levels followed by (1) comparison of the scores at a particular grade, and (2) comparison of the gains between grade levels or the relative positions at different grade levels. Table 2, which is taken from figures published in California in 1968, illustrates a typical publication of scores. It shows percentile scores in reading in grades 1, 3, 6, and 10 for four of the largest school districts in California. (Percentile scores are based on the distribution of student scores, not the distribution of school scores. School scores and school-district scores are more closely grouped around the mean. Thus, if a school district is "at the 15th percentile," this means, not that 15 per cent of schools are below it, but that its average is above that of 15 per cent of all students in the population used to norm the test.)

Table 2

Reading achievement test scores by percentile for four of the largest districts in California (from The Los Angeles Times, February 23, 1968).

<table>
<thead>
<tr>
<th>District</th>
<th>Grade 1</th>
<th>Grade 3</th>
<th>Grade 6</th>
<th>Grade 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>29</td>
<td>37</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>San Diego</td>
<td>61</td>
<td>59</td>
<td>59</td>
<td>58</td>
</tr>
<tr>
<td>San Francisco</td>
<td>41</td>
<td>36</td>
<td>40</td>
<td>38</td>
</tr>
<tr>
<td>Oakland</td>
<td>49</td>
<td>40</td>
<td>33</td>
<td>31</td>
</tr>
</tbody>
</table>
In these comparisons, the first-grade reading level measures student performance approximately at school entrance. A gain in percentile position relative to that starting point indicates that students at higher grades in that school system are performing relatively better than those at lower grades; a loss in percentile position indicates that higher grades are performing relatively less well than lower grades.
IV. THE DIFFERENCE IN SIZE OF THE PERCENTILE AT 50% AND 90%

Percentile scores may be misleading in a way different from grade-equivalent scores. Because of the bell-shaped distribution curve, many students are clustered closely in absolute score near the 50th percentile, while, at the 10th or 90th percentile, students are more widely spaced. Thus, a change in a school's position from the 50th to the 51st percentile is a small change in scale score compared to a shift from the 10th to the 11th percentile. In contrast to the original scale score, constructed as indicated in Section 2, the percentile score stretches out the scale toward the middle, and compresses it at the ends. Figure 3 (p. 18), which represents the equivalence between standard scores and percentile scores, shows that a change from the 10th to the 11th percentile is about equivalent in standard score to a change from the 50th to the 52nd percentile; and that changing from the 2nd to the 3rd percentile (or from the 97th to the 98th) is comparable in standard score to going from the 50th to the 57th percentile.

Consequently, the comparison of percentile scores is misleading as a measure of the amount of change, although it is useful as a measure of the direction of change in relative standing. The amount of change in test scores in Table 2, for example, would have been more accurately represented by standard than by percentile scores.

Standard scores merely change the metric and zero point of the scale score. By creating separate standard scores for every grade, it is possible to arrive at the same mean for the average at each grade level. On the basis of standard scores in which the mean is given a score of 5 and the standard deviation is 1, the scores of the districts in Table 2 may be expressed as shown in Table 3 (p. 19), which gives an accurate picture of the amount of change in the average score of each district.
Fig. 3. Equivalence between standard scores and percentile scores.
Table 3

Reading achievement test scores by standard scores for four of the largest districts in California.

<table>
<thead>
<tr>
<th>District</th>
<th>Grade 1</th>
<th>Grade 3</th>
<th>Grade 6</th>
<th>Grade 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Angeles</td>
<td>4.45</td>
<td>4.67</td>
<td>4.82</td>
<td>4.90</td>
</tr>
<tr>
<td>San Diego</td>
<td>5.28</td>
<td>5.23</td>
<td>5.23</td>
<td>5.20</td>
</tr>
<tr>
<td>San Francisco</td>
<td>4.77</td>
<td>4.63</td>
<td>4.75</td>
<td>4.70</td>
</tr>
<tr>
<td>Oakland</td>
<td>4.98</td>
<td>4.75</td>
<td>4.56</td>
<td>4.50</td>
</tr>
</tbody>
</table>
V. ABSOLUTE CHANGES OVER A GIVEN PERIOD

Scores such as those presented in Tables 2 and 3 show the position of school districts relative to a norm at each grade level. They do not, however, show gain or growth in absolute position. In a school which remains at the 50th percentile on a standard score of 5.0, the average student has progressed at exactly the average rate. It may be desirable to show the rate of growth by the change in scores when different forms of the same test, or linked tests, are used at two or more points in time. For example, comparable tests are often given at the beginning and end of a special program to assess its effectiveness. It is for such purposes that the grade-equivalent scores, misleading as they are, are frequently used.

To make a valid comparison of this kind, one needs a score which shows amount of gain without the misleading properties of the grade-equivalent score. This need is met by a standard score based on a single standardization at the first of the two tests, rather than a separate standardization for each grade (as in Table 3). Take, for example, a program that begins with the 6th grade and continues through grades 7 and 8, with tests given at the beginning of grades 6 and 9. The effect of the program on each student, or on the average of all students, would be measured by the increase in number of standard deviations for each student on the basis of the 6th-grade normed distribution. Looking at the figures for reading comprehension in Table 1, for instance, consider three students, represented by the 84th percentile, the 50th percentile, and the 16th percentile, respectively, in the national distribution at grade 6; and suppose they remain at these percentiles relative to the students in the same grade at grade 9. In standard scores based on the distribution at grade 6, their scores are 6.0, 5.0, and 4.0, respectively. The standard deviation at grade 6 is 16 points in raw score. Between grades 6 and 9, the 84th percentile student who remains at the 84th percentile gains 19 points in raw score (or 1.2 in standard score); the 50th percentile student who remains at the 50th percentile gains 19 points in raw score (or 1.2 in standard score); and the 16th
percentile student who remains at the 16th percentile also gains 19 points in raw score (or 1.2 in standard score). The final standard scores are 7.2, 6.2, and 5.2, based on a 6th-grade norm of 5.0 standard score. In this example, the student at each percentile gained the same amount in standard score, although as shown in Table 1, the 16th percentile student "fell behind" an additional .3 years, and the 84th percentile student "moved ahead" an additional 1.0 year. Of course, a student who remains at the same relative or percentile position does not necessarily always gain the same amount; if, instead of reading comprehension, we had used verbal ability as the example, the standard scores would have been 7.9, 6.7, and 5.5 for the three students.

The foregoing is meant to show, not that grade equivalents are irrelevant information, but rather that they are not an appropriate test for the impact of a program on students who begin at different points. If a student who begins at a low level always gains only the same amount in standard scores as the student who begins at a higher level, he will never catch up; but to assess the effectiveness of the program correctly, one should be able to measure the "catching up" as extra gain, shown by gain in standard score, instead of allowing such extra gain to be obscured in grade equivalents.
VI. THE INCORRECT USE OF ABILITY MEASURES
TO ESTABLISH EXPECTED ACHIEVEMENT LEVELS

One method that is sometimes used to determine the "expected" level of performance of students in a given school or program is an ability, or intelligence, measure scored in grade equivalents. The grade level of the ability measure is then compared with the grade level of the achievement measure to determine whether, on the average, students in the school or program are achieving above, below, or at their ability level.

This procedure developed out of a similar one used at the individual student level to spot students who were "underachieving" relative to their ability. The method, whether applied to schools or to individual students, appears to allow a measurement of expected achievement through use of an ability and achievement test at a single point in time, rather than of achievement tests at two points in time, as do most other methods. However, this is an illusory possibility, both for individuals and for schools, because of two defects in the method.

The first of these lies in the fact that abilities ordinarily are measured in one or two dimensions (over-all ability, or verbal and nonverbal ability), while achievement has a number of dimensions. Further, even when achievement bears a name similar to that of the ability measure, it may be even more closely related to other abilities. For example, scores on mathematical achievement tests typically are more highly correlated with scores on verbal ability tests than with scores on nonverbal ability tests. Since ability test scores do not necessarily correspond to particular achievement scores, we cannot reliably use them to predict such achievements.

Even if there were such a correspondence, however, the second defect in the method would remain: the assumption that the ability tests measure qualities that are not subject to change by the schools. It is a faulty assumption, which can be challenged in any of several ways.
First, anyone examining group-administered tests of ability or intelligence and group-administered achievement tests may find it difficult to determine which is which, except for the section on spatial perception in nonverbal ability. Group-administered verbal ability or intelligence tests are largely tests of vocabulary in one or another form.

Second, when one looks at increases in raw scores on ability or intelligence tests, one finds just as much increment in performance with the increase in grade level as in tests designed to measure specific achievement, say in reading comprehension. In the three tests listed in Table 1, the verbal ability test showed a slightly greater increment in performance between grade 6 and grade 12 than either the reading or the mathematics tests. There is no a priori reason for assuming that this increment in ability is purely the result of maturation, unaffected by school environment, any more than is the increment in reading or mathematics achievement. Indeed, the verbal ability test is considered in this report, as it was in the original publication in which it was used, an achievement test like the others.

Third, empirical evidence suggests that increments in performance in ability tests tend to be just as much related to variations in school environment as are those in tests labeled achievement tests.

The use of so-called ability or intelligence tests to provide an expected level of achievement is thus not valid. If the school can affect either test score equally, then the "expected" measure is contaminated by school effects; and if some school programs affect one more than the other while some do just the opposite, inferences drawn about the functioning of school programs will be incorrect.

This does not negate the use of both standardized "achievement" and "ability" tests to determine the expected level of performance of each student relative to others in the same class. Everyday performance, which depends on sustained interest and motivation as well as on the skills manifested in such tests, may vary quite widely from standard test scores, so that comparison of the two can provide information about a student's interest and motivation. However,
precisely because classroom performance is not measured in standard fashion across schools or even classrooms, this diagnostic use of test scores cannot be made at the level of the school or program.
VII. INFERENCES ABOUT SCHOOL EFFECTS
FROM SCHOOL PERFORMANCE MEASURES

The principal reason behind interest in school achievement measures is the desire to assess the effects of schools on achievement. However, these cannot be inferred with any assurance from measures of school achievement of the type currently available. The essential difficulty derives from the fact that important effects on achievement lie outside the school, and that the amount of achievement attributable to school factors remains unknown. In the absence of a specific model showing how school factors combine with other factors in affecting achievement and how important they are, it is not possible to infer directly from the measures of achievement by school or school district just what the "expected" increment in achievement should be in a school for children who differ in nonschool factors. Suppose, for example, that all schools remained at the same percentile position from grade 1 to 12. There would then be some reason for saying that all schools were identically effective. Yet if a child's later achievement showed an increasingly large component due to the school, and if all schools were equally effective, there should be a movement of school mean toward the over-all mean in later school years -- a movement due to the increasing size of the school component of achievement -- instead of the school's maintaining the same percentile position. Thus, identically effective schools, or schools of equal quality, should show a convergence of percentile or standard scores over years in school, the extent of the convergence depending on the strength of the schools' compared to external effects. The apparent lack of any evidence of such convergence or regression toward the mean indicates that either (a) there is a near-perfect correlation between the effectiveness or quality of a school and the starting achievement level of its entering first-graders, or (b) school effects are quite weak relative to nonschool effects. The latter explanation would seem more credible than the existence of a near-perfect correlation; this suggests that, at the current levels of effectiveness of schools, the general expectation for two schools of
equal quality whose entering students differ in standard scores or percentiles is that the same differences will continue throughout the school years. In fact, to the question "What is the expected increment in achievement for a student with a given initial score, or a set of students with a given average initial score, when two forms of the same test are used as the initial and final tests?" -- a question that usually is implicit where inferences about effects of school programs are made -- the answer is unambiguous: If all the factors affecting achievement continue at the same level over the period between testings as in the period prior to the first test, the expected increment should be, not identical grade-level increments, but an identical increment in standard score for all students (standardized at the date of the earlier test), regardless of starting point.

The expected increment in standard scores will be independent of the starting point only when the same forces can be expected to continue between tests that were in effect before the testing. If these forces are changed during the period between tests -- for example, to increase educational resources for initially low-achieving children -- then the impact of this redistribution should show up in larger standard-score increments for the initially low-achieving students (though increments in grade equivalents may still be smaller).

The fact that expected increments in standard score should be identical for all students, and independent of starting point, if all achievement-affecting factors remained the same between testing points as they were before has an implication also for research on effects

*If the period between tests is so long as to require linked tests designed for different levels of performance (as in Table 1 or Figure 1), then these tests may have different standard deviations in raw score, making it necessary to create separate standard scores at the two periods. If so, the expected increment cannot be measured directly, but must be inferred as that increment in achievement which gives the same standard score at the two time points. However, linked tests should be so constructed as to have identical standard deviations of raw scores in the norming populations. The two tests can then be treated as two forms of the same test, using standard scores based only on the starting point, with the expected increments in standard score identified independently of starting point.
of school factors. Often, in such research, the aim is to isolate the effect of school factors on achievement from the effect of non-school factors such as family background. Given a single test score, background factors of the child are introduced into the analysis along with school factors, so that their effect can be controlled insofar as possible. When there are two test scores, however, at two points in time, the question arises how to use the two scores as measures of school effect, separating out the nonschool factors.

As indicated earlier, when factors that affect the first test score remain the same between tests, increments in standard score should be identical. Consequently, the results of the first test embody the effects of those background factors, and it is not necessary, then, to include the background factors in the analysis of the final test score. The only such factors which should show an effect beyond that incorporated in the first test score, and thus the only ones that need to be included in the analysis, are those which are different between testing periods from what they were before.*

The same principle holds for school factors: only those which are different between testing points from what they were before will have a bearing on whether the gain in standard score of the school average is more or less than the average gain. What this means in practice is that a first-grade test score, at the very beginning of school, has a very special role. It embodies, subject to its degree of reliability, the child's background factors, both environmental

---

*A caveat is necessary here. Because any test score has imperfect reliability, it embodies only imperfectly the background factors that are also embodied in the final score. In practice, therefore, it is of value to include family background measures as well as initial test score in the analysis of the final test score, in order to add those background effects that are missing in the initial score because of imperfect reliability. Thus, the effects shown by background factors will include two components: that due to imperfect reliability of the first test score, and that due to changes in the effects of background factors between first and second tests. Because, however, these background factors may be correlated with school factors that affect the second test score, they may reduce the regression coefficients of those school factors. There is some argument, then, for not including them along with the first test score, even though the reliability of that score is imperfect.
and hereditary, without containing school influences. The increase or decrease between the first-grade and later tests in standard scores of children in particular educational settings can be attributed to those educational settings. Technically, a regression analysis which has as dependent variable the second-test standard score, and as independent variables the first-grade test score, family background factors, and school factors, should lead to this basis for interpretation: regression coefficients on family background factors show, not total family background effects, but only changes in family background effects between the two test periods (since the basic background effects are embodied in the initial test scores); regression coefficients on school factors show total school effects since the beginning of school, as no school effects were embodied in the initial test scores.

* However, if variations in school factors are confounded with variations in initial test scores or family background scores, this will reduce the regression coefficients of both, and show up as a variance that can be accounted for by school factors, or family background factors, or initial test score. For this reason it may be desirable to exclude constant background factors, as mentioned in the preceding footnote.
VIII. STANDARDIZATION FOR POPULATION CHANGES

In comparing test scores at different grade levels in a school or school district, and noting the differences among schools or districts, it is tempting to infer changes from apparent trends. For example, looking at Table 2 or 3, one may be tempted to infer "gain" between grades 1 and 10 in San Francisco, and "decline" in Oakland. But it may well be that the population of families from which 10th-grade scores come in Oakland is different from the population from which first-grade scores come, and that the apparent decline is due only to population differences.

There is a method by which such population differences between grades can be standardized so as to eliminate this confounding factor. This is done, heuristically, by reweighting the test scores at later grades before averaging, to reflect as closely as possible the population that exists at the earliest grade. (The specific method of reweighting depends on the characteristics that are measured for children at later grades and at the first grade.) To illustrate the method, assume that for each of the students now in grades 3, 6, and 10 in Table 3, a standard test score had been obtained at grade 1 and entered in his record. Then, for each small interval of the grade 1 standard score, calculate the proportion of students at current grade 1 who have that score, say \( X_1 \), and the proportion of students currently at grade 10 who had that score at grade 1, say \( X_{10} \). Next, the grade 10 score of each current grade 10 student who had a grade 1 score in that interval is weighted by the ratio \( X_1 / X_{10} \). This is done for all intervals in the grade 1 test-score distribution, and an average is obtained of the weighted grade 10 test scores. This average is standardized to the current grade 1 population, and may be compared to the current grade 1 average. The average is not the actual grade 10 average in the district, but is a better estimate of what the average would be if the population composition were the same as that currently in grade 1.

If there is no record of a grade 1 test score for these higher grades, it is necessary to use characteristics for standardizing which
are closely related to the standardized test scores and on which information exists for all the students currently in grade 1, as well as those in grades 3, 6, and 10. Where data on race or father's occupation, or both, are available for each student, either of these characteristics, or the two jointly, can be used for standardization if the grade 10 (or 3 or 6) score of each student in a given population group is weighted by the ratio of the proportion of that group in grade 1 to that in grade 10 (or 3 or 6). This will standardize the test averages at each of the later grades to a population base which is the same, in those characteristics used in standardizing, as the current grade 1 population.
IX. THE DISTRIBUTION OF PROGRESS

In the measurement of school or district performance levels, the distribution of performance is seldom assessed. Yet such information can be just as important a measure of a school's performance as the average level of student performance, and may be an even more important measure of the effectiveness of a school's programs.

The measures by which schools or districts are ordinarily compared are school averages; but these averages conceal the size of the gap between the highest-performing students and the lowest-performing ones. Measures of the distribution of performance, and changes in this distribution, may be presented in several ways. Of the two that are most reasonable, the first would give both the mean standard score and the standard deviation within the school or the district at each level. If a school has a student population of diverse backgrounds, its standard deviation will be large at grade 1. (Ordinarily, the standard deviation of a school's score will be less than that of the norming population. If standardized scores with a mean of 5.0 and a standard deviation of 1.0 are used, the schools' or districts' standard deviations will be less than 1; it should average around .9 for a school, since about .8 of the variance of standardized achievement-test scores lies within schools.)*

If the standard deviation of a school's standard score decreases from grade 1 to 12, while the average remains constant, this means that, relative to the norming population, the school's students are becoming more homogeneous in performance, with both the high- and the low-performing students moving closer to the average than is true in the population as a whole. If the decline in the standard deviation is accompanied by an increase in average standard score, this represents principally extra gain among the initially low-performing students. If it is accompanied by a decrease in average standard score, it represents principally deficient gain among the initially high-performing students.

---

*Equality of Educational Opportunity, Table 3.22.1, p. 296.
The second concise way of providing information about the distribution of scores within the school is to present more than one average. Giving the averages for the top and bottom quartiles of the school or district, for example, provides distributional information with two numbers, which is more easily grasped than a number showing the standard deviation. (It does not, however, give the mean for the school as a whole; to do that would require a third number at each grade level.) There are numerous variations on this mode of presenting distributional information, such as giving averages for the first, third, and fifth quintile, or, for more extreme scores, the middle, top, and bottom ninths of the students in a school or district.

The interest in distribution of performance derives from the same source as that in grouping or tracking. If grouping or tracking in a school has the effect of widening the divergence between students in the high and the low tracks, this fact should show up in the comparison of trends in the distribution of test scores for districts with different degrees of grouping or tracking. If tracking is increasing the diversity of achievement in a school, distributional information of the sort described above will show it; if it is not, this fact will be reflected in a nondivergent distribution.
X. TOTAL MOVEMENT IN ACHIEVEMENT

Few schools and school districts are as yet equipped to present information about changes in the achievement of individual students without special processing of the students’ records. Within a few years, however, many schools will be so equipped. Once such changes can be examined as part of regular administrative data-processing, it will be possible to obtain measures of a student’s average movement in achievement. Even at present, this information can be tabulated easily in some schools. In the assessment of the impact of special programs, such measures are readily obtainable because of the before-and-after tests which ordinarily are given to students in these programs.

Whereas the change in average achievement between two tests is the average of the differences in test scores, or, equivalently, the difference of the averages -- i.e., \( \frac{\sum_j (y_{2j} - y_{1j})}{n} = \frac{\sum_j y_{2j}}{n} - \frac{\sum_j y_{1j}}{n} \) (where \( y_{1j} \) is the standard test score of person \( j \) at time 1) -- the measure of total movement is the average of the absolute differences, \( \frac{\sum_j |y_{2j} - y_{1j}|}{n} \). It captures all movement, up and down, and must be based on differences in the test scores of each individual at the two time points.

This is another important indicator of a school’s functioning. It shows the ability of a curriculum to reduce the dependence of final achievement on the initial level of achievement. For example, given two schools whose curricula are highly stratified in tracks, it may be relatively easy in one school to move between tracks but relatively difficult in the other -- a difference that will be reflected in a higher rate of total movement in achievement in the first school. This particular information is, in fact, more often needed to assess the effects of tracking than is information merely on the distribution of achievement. It is more directly indicative of the degree of equality of opportunity in schools than are any of the measures discussed in preceding sections. For the degree to which a school is able to make achievement levels independent of initial achievement levels constitutes the degree of equality of opportunity
that it provides. (This is not to say, of course, that equality of opportunity within a school is the sole important question about a school's functioning. The over-all magnitude of opportunity reflected in the average increment in achievement, which is the focus of current measures of school performance, is of obvious importance as well.)
XI. FURTHER MEASURES IN ADVANCED DATA SYSTEMS

As data systems in schools and school districts grow more sophisticated, increasingly detailed information will become available. Appropriately used, this information can simplify, rather than confuse, the task of the parent or legislator or school-board member attempting to assess the functioning of the school. In particular, it can be expected to separate out the effects of school programs from those effects on student performance that arise from home and other influences; the measures described above and others currently available only allow indirect inferences about school effects.

Here is an example of the kind of measure that would become available if existing data in student records were analyzed with more sophisticated data systems: For a child whose standard score (or percentile score, or grade equivalent) is at a given level at grade 3, the expected distribution of scores at any later grade could be ascertained for a given school. In that school, if he is reading at standard score 4.5 at the beginning of grade 3, he may have a 30 per cent chance of reading at 5.0 or above at the start of grade 6, a 40 per cent chance of reading between standard scores 4.5 and 6.0, and a 30 per cent chance of reading below 4.5. (In another school, a child with the same reading score at the start of grade 3 would have different probabilities of achieving at these levels.) More precise estimates could be obtained if one included additional information on the child, such as his age and his reading score at grade 1.

Such measures lie in the future; though the data for them exist, the data systems in schools and school districts generally lack both the computer hardware and the programming software for making the information available. In contrast, the measures discussed in preceding sections (except for section X) are easily obtainable with present data systems. It is important, however, to develop consistency in reporting procedures, and to use correct measures in place of the inaccurate ones frequently presented in current reporting, so that measures of school performance may not be discredited, nor used incorrectly as the basis for policy decisions.