EDUCATIONAL ASSESSMENT, *10*(4), 377–399

Regression to the Mean in Average Test Scores

Gary Smith
Department of Economics
Pomona College

Joanna Smith
Rossier School of Education
Los Angeles, CA

A group's average test score is often used to evaluate different educational approaches, curricula, teachers, and schools. Studies of group test scores over time often try to measure "value-added" by holding constant certain student characteristics such as race, parents' education, or socioeconomic status; however, the important statistical phenomenon of regression to the mean is often ignored. There is a substantial literature on the importance of regression to the mean in a variety of contexts, including individual test scores. Here, we look at regression to the mean in group averages. If this regression is not taken into account, changes in a group's average test score over time may be misinterpreted as changes in the group's average ability rather than natural and expected fluctuations in scores about ability. California Academic Performance Index scores are used to illustrate this argument.

Average test scores are often used to compare groups of students to make summative or formative evaluations of teaching methods, teachers, and schools. In today's climate of increased accountability and the high stakes of The No Child Left Behind Act of 2001 (2002), it is particularly important that educational leaders and policymakers understand what they can and cannot infer from test scores. The nonrandom nature of many groups (for example, school populations) means that the average group score can be a misleading measure of a school's effectiveness. Some schools have higher scores than other schools not due to the school, but

Correspondence should be addressed to Gary Smith, Department of Economics, Pomona College, 425 N. College Avenue, Claremont, CA 91711. E-mail: gsmith@pomona.edu

because of the students. Educational value added might be better measured by changes in test scores over time but the interpretation of such changes is muddled by the statistical phenomenon known as regression to the mean. While much attention has been paid to controlling for students' background characteristics (e.g., race, parents' education, socioeconomic status) in comparisons of nonrandom school populations, regression to the mean has been virtually ignored. In this article, we show how the average test scores of groups regress to the mean and use scores from California's accountability system to illustrate this argument.

ACCOUNTABILITY AND INTERNAL VALIDITY

All accountability systems are affected by the well-known issues of sampling error, test reliability, generalizability, and validity that should be addressed when making judgments of teaching effectiveness. The fundamental concept of internal validity is whether the test measures what it is intended to measure. Does the test provide an authentic assessment? Common threats to an authentic assessment include history, maturation, testing, instrumentation, and statistical regression (Isaac & Michaels, 1995).

A *history* effect occurs when an event separate from the treatment affects the outcome. For example, a test might require students to write vividly on the topic of "hardship." A hurricane that leaves the town without electrical power for 3 days right before the test may have more to do with the students' responses to the test question than the poems or stories read in class.

A *maturation* effect occurs when natural biological and psychological changes that happen during the time between the pretest and posttest affect student outcomes apart from the actual treatment. For example, young students may develop better motor skills that improve their test scores in drawing or writing regardless of what happens in school.

A *testing* effect occurs when the administration of the pretest alters the outcome of the posttest. For example, students who become motivated by pretest questions to watch a television documentary about World War I will answer more posttest questions about the war correctly, even if their school curriculum ignores the war.

An *instrumentation* effect occurs when there are changes in the test or how the test is administered. For example, California has used completely different tests, some of which are scored on the basis of nationwide quintiles and some of which are scored on the basis of performance to a standard.

A statistical *regression* effect causes extremely high or low pretest scores to tend to move toward the mean on the posttest regardless of treatment. Although it is obviously not the only problem inherent in interpreting test scores, we focus here on regression to the mean because it is so widely overlooked and misunderstood.

REGRESSION TO THE MEAN

Regression occurs in a variety of contexts (Schmittlein, 1989). For instance, any two variables with equal variances and a joint normal distribution with correlation between 0 and 1 exhibit regression to the mean (Maddala, 1992, pp. 104–106). Suppose, for example, that height and weight are bivariate normal and have been scaled to have equal variances. Because height and weight are imperfectly correlated, the tallest people are not the heaviest (weights regress to the mean) and the heaviest are not the tallest (heights regress to the mean).

In educational testing, the "true score" is the statistical expected value of a person's test score (Lord & Novick, 1968). A person's observed score on any single test depends on the questions asked, the person's health during the test, and even the person's luck when unsure of the answer. The difference between an observed test score and the true score is the error score. Observed scores regress toward the mean because those who score highest on a test are likely to have had positive error scores that put their observed scores farther from the mean than are their abilities.

For example, a second grader who scores in the 90th percentile is more likely to be someone of somewhat more modest ability who did unusually well than to be someone of higher ability who had an off day, because the former outnumber the latter. When the modest-ability student takes another test—the next day, the next week, or 3 years later—he will probably not do as well. Students, parents, and teachers should anticipate this drop-off and not blame themselves if it occurs. Similarly, students who score below average are likely to have had an off day and should anticipate scoring somewhat higher on later tests. Their subsequently higher scores may be a more accurate reflection of their ability rather than an improvement in their ability.

An Overlooked Phenomenon

Regression toward the mean is a pervasive but subtle statistical principle that is often misunderstood or insufficiently appreciated. An anonymous reviewer for an article one of the authors wrote on regression to the mean in sports performances noted that

> There are few statistical facts more interesting than regression to the mean for two reasons. First, people encounter it almost every day of their lives. Second, almost nobody understands it. The coupling of these two reasons makes regression to the mean one of the most fundamental sources of error in human judgment, producing fallacious reasoning in medicine, education, government, and, yes, even sports. (Lee & Smith, 2002)

There is well-established evidence that most people are blind to regression to the mean in a variety of contexts (Campbell, 1969; Kahneman & Tversky, 1973). Regression to the mean in educational testing is discussed by Kelley (1947), Lord and Novick (1968), and Thorndike (1963). Kahneman and Tversky (1973) noted that regression to the mean is all around us—including scores on consecutive tests—yet most seem blind to it:

> First, they do not expect regression in many situations where it is bound to occur. Second, as any teacher of statistics will attest, a proper notion of regression is extremely difficult to acquire. Third, when people observe regression, they typically invent spurious dynamic explanations for it. (p. 250)

One of their examples is a problem they presented to graduate students in psychology, which described an actual experience:

> The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with this training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try. What should the psychologist say in response? (pp. 250–251)

None of the graduate students suggested regression as a possible explanation for the observed data. “The respondents had undoubtedly been exposed to a thorough treatment of statistical regression. Nevertheless, they failed to recognize an instance of regression when it was not couched in the familiar terms of the heights of fathers and sons” (Kahneman & Tversky, 1973, p. 251).

Thorndike (1963) gave several examples of education research that was seemingly unaware of regression to the mean. In one study, college students whose grade point averages (GPAs) were relatively low compared to their scores on a standardized test were labeled underachievers, and students whose GPAs were relatively high compared to their scores on a standardized test were labeled overachievers. Regression to the mean implies (just as with height and weight) that those with the highest test scores are unlikely to have the highest GPAs and that those with the highest GPAs are unlikely to have the highest test scores. Thorndike noted that “the ‘underachievers’ could just as truly be called ‘overintelligent,’ and the ‘overachievers’ called ‘underintelligent’ ” (p. 13).

Another study looked at the reading gains of college students whose initial scores placed them in a remedial group. Thorndike (1963) observed that their scores could be expected to improve on a second test even if the instructor “had no more than passed his hand over the students’ heads before he retested them” (p. 14).

Wainer (1999) gave this example, which occurred when he was a statistician with the Educational Testing Service:

> My phone rang just before Thanksgiving. On the other end was Leona Thurstone; she is involved in program evaluation and planning for the Akebono School (a private

> school) in Honolulu. Ms. Thurstone explained that the school was being criticized by one of the trustees because the school's first graders who finish at or beyond the 90th percentile nationally in reading slip to the 70th percentile by 4th grade. This was viewed as a failure in their education. Ms. Thurstone asked if I knew of any longitudinal research in reading with this age group that might shed light on this problem and so aid them in solving it. I suggested that it might be informative to examine the heights of the tallest first graders when they reached fourth grade. She politely responded that I was not being helpful. (p.26)

Similarly, in December 2002, one of the authors of this article was on a committee interviewing candidates for a tenure-track academic position. One candidate, whose avowed specialty was educational testing, was asked how she would interpret data showing that a student who had scored 1.0 standard deviation above the mean on a test administered at the start of the school year scored 0.8 standard deviations above the mean on a similar test administered at the end of the year. Her answer was that the school had failed this student.

The Basic Regression Framework

Each student's ability μ is the statistical expected value of his or her test score. We assume that a student's score X on any particular test differs from ability by an independent and identically distributed error term ε:

$$X = \mu + \varepsilon \tag{1}$$

Looking at a test involving a group of students, there is a distribution of abilities and scores across students. If the error scores are independent of abilities, then the variance of the observed scores is equal to the variance of abilities plus the variance of the error scores:

$$\sigma_X^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$$

Thus the variance of scores is larger than the variance of abilities: Observed differences in scores typically overstate the differences in abilities.

A test's reliability is gauged by the squared correlation ρ^2 between scores and abilities, which equals the ratio of the variance of abilities to the variance of scores[1]:

$$\rho^2 = \frac{\sigma_\mu^2}{\sigma_X^2} \tag{2}$$

[1] Mathematical derivations of all formulas are available from the authors.

The correlation between scores and abilities approaches 1 as the standard deviation of the error term approaches 0, and approaches 0 as the standard deviation of the error term becomes infinitely large.

If we knew the students' abilities, Equation 1 could be used to make unbiased predictions of each student's score. However, we must use the observed scores to predict the unobserved abilities. Suppose hypothetically that we have data on abilities and use a large sample to estimate

$$\mu = \alpha + \beta X + \upsilon \tag{3}$$

by ordinary least squares. The slope will be very close to ρ^2 and, thus, the predicted deviation of a student's ability $\hat{\mu}$ from the mean ability $\bar{\mu}$ is a fraction of the deviation of this person's score X from the mean score $\bar{X}$:

$$\hat{\mu} - \bar{\mu} = \rho^2 \left(X - \bar{X}\right)$$

The squared correlation between scores and ability is used to shrink each student's predicted ability toward the mean. For example, if a test's reliability is 0.8, students who score 10 points above the mean are predicted to have an ability 8 points above the mean. On a comparable test, their scores can consequently be expected to average 8 points above the mean; that is, to regress to the mean.

In a large sample, the mean of ε will almost certainly be close to 0, so that the mean score $\bar{X}$ will almost certainly be very close to the mean ability $\bar{\mu}$. Therefore, we can also write

$$\hat{\mu} = \left(1 - \rho^2\right)\bar{X} + \rho^2 X \tag{4}$$

Predicted ability is a weighted average of the score and the mean score, using the squared correlation coefficient as the weight. In classical test theory, this is Kelley's equation (Kelley, 1947).

In practice, we do not observe abilities and consequently cannot use data for scores and abilities to estimate ρ^2. Instead, we can use the observed scores X and Y on two comparable tests because the population correlation between the two scores ρ_{XY} is proportional to the squared correlation between scores and abilities:

$$\rho_{XY} = \frac{\sigma_X}{\sigma_Y}\rho^2$$

Thus our estimate of the squared correlation between scores and ability is

$$\rho^2 = \frac{\sigma_Y}{\sigma_X}\rho_{XY} \tag{5}$$

If the two tests have equal standard deviations, then $\rho^2 = \rho_{XY}$.

We can also use least squares estimates of this equation to predict the scores on a comparable test:

$$Y = \alpha + \beta X + \upsilon$$

With a large sample, the estimated slope is

$$b = \frac{\sigma_Y}{\sigma_X}\rho_{XY} \tag{6}$$

The predicted deviation of a student's score $\hat{Y}$ from the mean score on one test is equal to this slope times the deviation of this person's score from the mean score on the other test:

$$\hat{Y} - \bar{Y} + b\left(X - \bar{X}\right) \tag{7}$$

If the two tests have the same mean, Equations 4 and 7 are equivalent.

A Hypothetical Example

To illustrate these arguments, consider the hypothetical abilities and scores in Table 1. These 20 students have abilities ranging from 550 to 750, which we initially assume to be constant between third and fourth grade. For the 4 students at each ability level, 2 have –25 error scores and 3 have +25 error scores on the third-grade test. On the fourth-grade test, 1 of the students with a –25 error score on the third-grade test again has a –25 error score, while the other now has a +25 error score. Similarly, 1 of the students with a +25 error score on the third-grade test has a –25 error score on the fourth-grade test and the other has a +25 error score. For example, of the 4 students with 550 ability, 2 score 525 on the third-grade test and 2 score 575. On the fourth-grade test, 1 of the students with a 525 third-grade score scores 525 again and the other scores 575. The uniform distribution of abilities and error scores is unrealistic, as are the severely limited values. This stark simplicity is intended to clarify the argument.

TABLE 1
Illustrative Abilities and Scores

Ability μ	*3rd-Grade Scores* Y_3	*4th-Grade Scores* Y_4
550	525	525
550	525	575
550	575	525
550	575	575
600	575	575
600	575	625
600	625	575
600	625	625
650	625	625
650	625	675
650	675	625
650	675	675
700	675	675
700	675	725
700	725	675
700	725	725
750	725	725
750	725	775
750	775	725
750	775	775

Figure 1 shows a scatter plot of ability and third-grade scores. (Duplicate observations are offset slightly.) The least squares line, $Y_3 = 0.0 + 1.0\mu$, means that the best predictor of a student's score is his or her ability. Figure 2 reverses the data, now using scores to predict ability. This least square line is $\mu = 72.2 + 0.89Y_3$, as predicted by Equation 4:

$$\begin{aligned}\hat{\mu} &= \left(1-\rho^2\right)\bar{X} + \rho^2 X \\ &= \left(1-0.89\right)650 + 0.89X \\ &= 72.2 + 0.89X\end{aligned}$$

The slope is less than 1 due to regression to the mean. Specifically, the predicted deviation of a student's ability from average ability is equal to 0.89 times the observed deviation of this student's score from the average score.

The larger the variation in error scores relative to the variation in ability, the greater the regression to the mean, For example, Figure 3 shows the relationship between scores and abilities when the error scores are increased from plus-or-minus 25 to plus-or-minus 50. As shown, the fitted line flattens farther

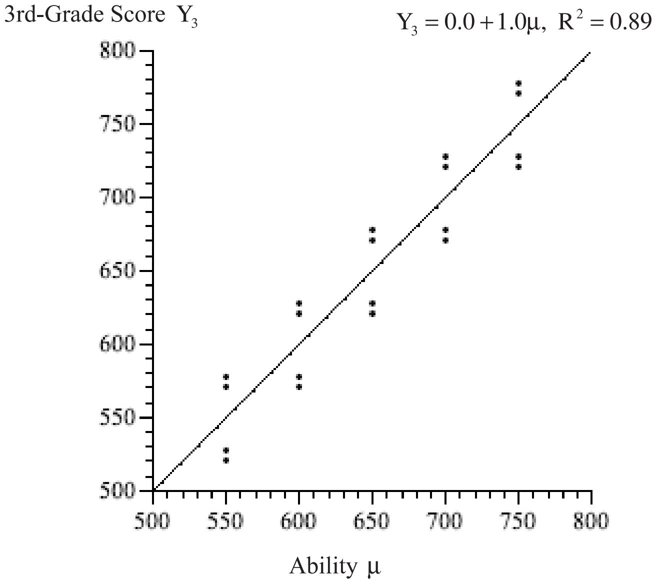

FIGURE 1 Abilities and third-grade scores.

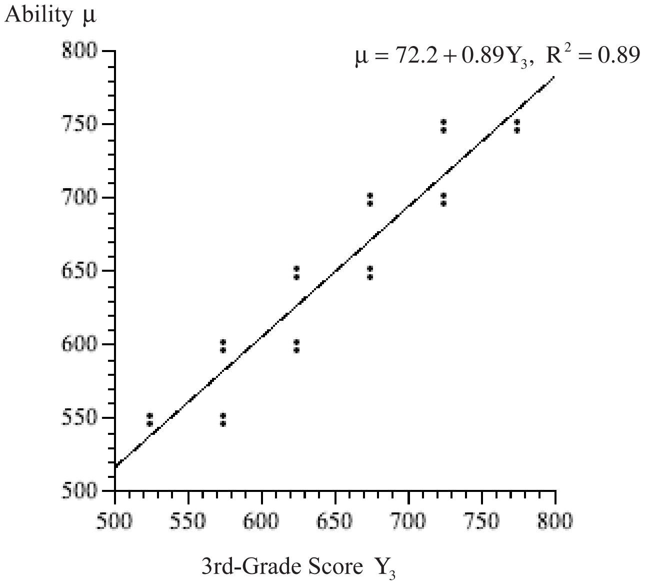

FIGURE 2 Third-grade scores and abilities.

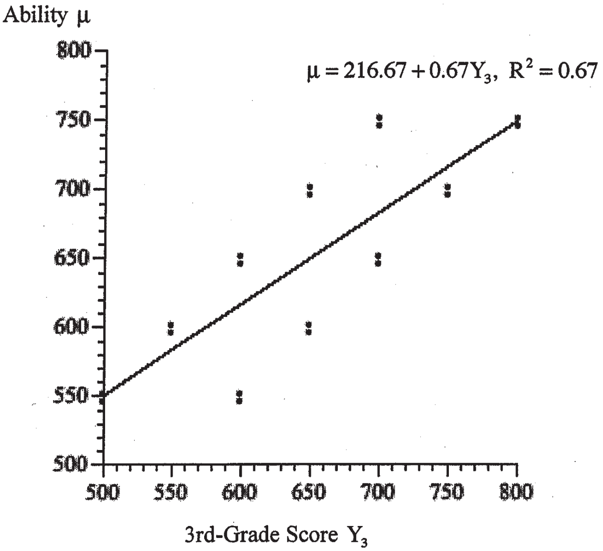

FIGURE 3 Third-grade scores and abilities, larger ε.

away from a 90-degree line. With error scores of plus-or-minus 25, a student who scores 10 points above the mean on the test is predicted to have an ability 8.9 points above the mean; with error scores of plus-or-minus 50, a student who scores 10 points above the mean is predicted to have an ability 6.7 points above the mean. As the error scores become infinitely large, the regression line becomes horizontal because test scores are no help in predicting ability.

Returning to the plus-or-minus 25 error scores in Table 1, suppose that we use the regression line $\hat{\mu} = 72.7 + 0.89X$ to predict fourth-grade scores. Figure 4 shows that, because there are no changes in abilities, the predicted fourth-grade scores that take regression to the mean into account are, in fact, unbiased predictors of the actual fourth-grade scores.

Figure 5 shows the relationship between third-grade and fourth-grade scores. The line shown is the regression-to-the-mean line (Equation 7), $\widehat{Y_4} = 72.2 + 0.89Y_3$, so that a student who scores 10 points above or below the mean as a third grader is predicted to score 8.9 points above or below the mean as a fourth grader. This makes sense because predicted ability is an unbiased predictor of fourth-grade scores and predicted ability regresses the third-grade scores toward the mean. Because our data assume no change in abilities, the regression-to-the-mean line is also the least squares line fit to the data.

The regression-to-the-mean argument does not depend on the timing of the tests. The highest scorers on the third-grade test are predicted to not do as well on

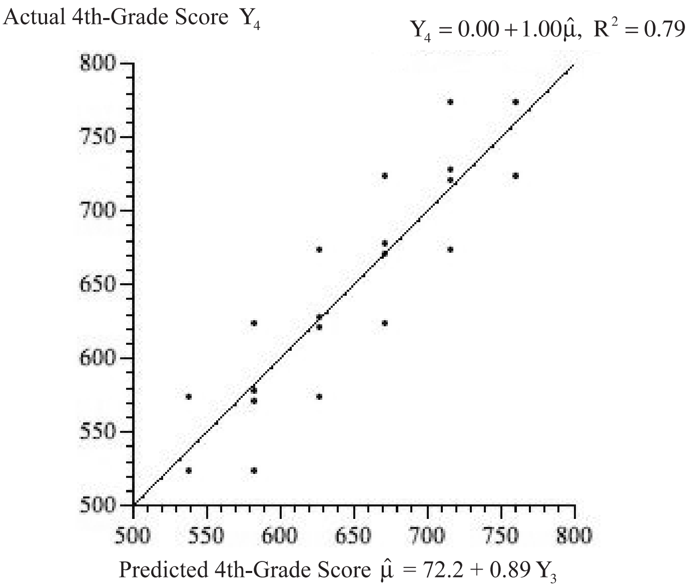

FIGURE 4 Predicted and actual fourth-grade scores.

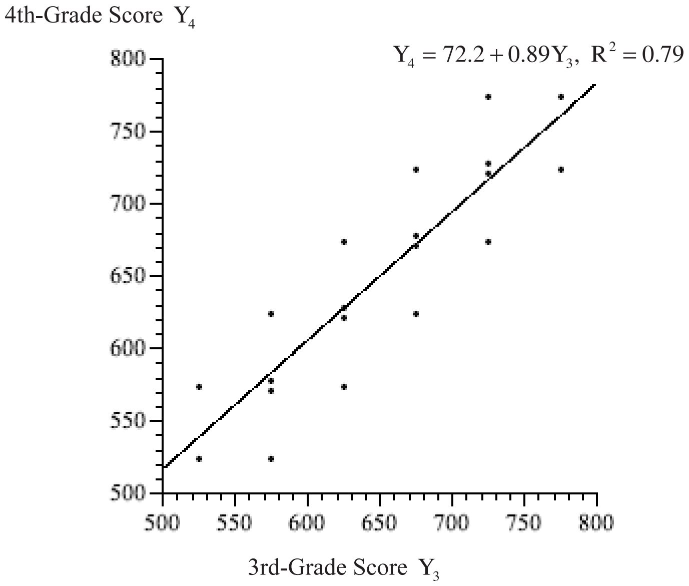

FIGURE 5 Third-grade and fourth-grade scores.

the fourth-grade test and the highest scorers on the fourth-grade test are predicted to not do as well on the third-grade test. Specifically, with our symmetrical data, the regression line relating third-grade scores to fourth-grade scores is $Y_3 = 72.2 + 0.89Y_4$. Students with third-grade scores 10 points above the mean are predicted to have fourth-grade scores 8.9 points above the mean and students with fourth-grade scores 10 points above the mean are predicted to have third-grade scores 8.9 points above the mean.

Nonrandomly Grouped Students

The previous example applies to an individual's scores on two tests. A student's average score on a great many tests will almost surely be very close to this student's ability which is, after all, the expected value of the student's score on these tests. It might be thought that the average score for a group of students will similarly almost surely be close to the average ability of these students. If so, regression to the mean would apply mainly to individual students rather than to a group's average test score.

However, just as a single student who does unusually well on a test probably had a positive error score, so a group of students that does relatively well probably had a positive average error score. Groups that are far from the mean on one test are likely to regress toward the mean on another test. If error scores are positively correlated within groups, this regression can be substantial. We will use a model with a hierarchical structure to extend the individual-student framework to the average score of students who are nonrandomly grouped; for example, in neighborhood schools.

Suppose that there are m groups with group i having n_i students whose average ability is μ_i. The distribution of average abilities across groups has mean μ and standard deviation θ; the ability μ_{ij} of student j in group i is described by a distribution with mean μ_i and standard deviation ϕ. This student's test score X_{ij} depends on ability and the error score:

$$X_{ij} = \mu_{ij} + \varepsilon_{ij}\,, \quad E\left[\varepsilon_{ij}\right] = 0 \qquad SD\left[\varepsilon_{ij}\right] = \sigma$$

For simplicity, we assume that the standard deviation of abilities within groups is the same for each group, and that the standard deviation of test scores about ability is the same for each student.

The error score ε_{ij} is likely to have a group component. The questions that are chosen for a particular test may be unusually well suited for a school's curriculum. Or perhaps the questions are ill suited for the neighborhood's culture. Or perhaps the test is given on a hot day and a school does not have air-conditioning. Or perhaps the test is given on a cold day and the heating system is not working properly.

Or perhaps a school has been hit by an infectious disease and many students are feeling poorly. Thus we break the error score into a group component ν_i and a student component ω_{ij},

$$\varepsilon_{ij} = \nu_i + \omega_{ij}$$

where the variance of ν_i is a fraction γ of the variance of ε_{ij}:

$$\begin{aligned} E[\nu_{ij}] &= 0 \qquad SD[\nu_{ij}] = \sqrt{\gamma}\ \sigma \\ E[\omega_{ij}] &= 0 \qquad SD[\omega_{ij}] = \sqrt{1-\gamma}\ \sigma \end{aligned}$$

The parameter γ is the correlation between the error scores for two students in the same group.

The reliability of the test with respect to individual students is equal to the square of the correlation between student scores and student abilities:

$$\rho_{ij}^2 = \frac{\phi^2 + \theta^2}{\phi^2 + \theta^2 + \sigma^2}$$

The reliability of the test with respect to average group scores is equal to the square of the correlation between average group scores and group abilities:

$$\rho_i^2 = \frac{\theta^2}{\gamma\sigma^2 + \dfrac{\phi^2 + (1-\gamma)\sigma^2}{n_i} + \theta^2}$$

This reliability measure correlates scores and abilities across groups. It can also be derived by correlating scores on parallel tests across groups.

A somewhat different measure is the correlation across tests of the average scores for a given group:

$$\tau_i = \frac{\phi^2}{n_i\gamma\sigma^2 + \phi^2 + (1-\gamma)\sigma^2}$$

Table 2 shows the reliability of a group's average score across groups and across tests for various values of the group effect and the reliability of the test for individual students. These illustrative calculations assume that the standard deviation of

TABLE 2
Reliability of Group Scores; Standard Deviation of Ability Across Groups = 100, Standard Deviation of Ability Within Groups = 50; Group Size = 500

	Reliability Across Groups ρ_i^2		*Reliability Across Tests* τ_i	
γ	$\rho_{ij}^2 = 0.8$	$\rho_{ij}^2 = 0.9$	$\rho_{ij}^2 = 0.8$	$\rho_{ij}^2 = 0.9$
0.00	0.999	0.999	0.444	0.643
0.25	0.927	0.966	0.006	0.014
0.50	0.864	0.935	0.003	0.007
0.75	0.810	0.905	0.002	0.005
1.00	0.762	0.878	0.002	0.004

Note. γ: correlation between student error scores within a group. ρ_{ij}^2: test reliability for individual students.

ability across groups is $\theta = 100$, the standard deviation of ability within groups is $\phi = 50$, and group size is $n_i = 500$.

If the within-group correlation of student error scores is 0, the reliability across groups is very close to 1; that is, a comparison of the average test scores for two groups is a reliable measure of the difference in average ability of these groups. However, as the within-group correlation of error scores increases, the reliability across groups falls and may be not much higher than the reliability across students. The reliability across tests is even lower.

AN EXAMPLE: CALIFORNIA'S STANDARDIZED TESTING AND REPORTING (STAR) PROGRAM

Many states have instituted high-stakes tests that directly affect students (grouping, promotion, and graduation), teachers (tenure, compensation, and bonuses), and schools (allocation of resources and imposition of sanctions). For example, California's STAR program requires all public school students in Grades 2 to 11 to be tested each year using statewide standardized tests. All schools are given an Academic Performance Index (API) score that is used for ranking the school statewide and in comparison to 100 schools with similar demographic characteristics. The API rankings are released to the media, placed on the Internet, and reported to parents in a School Accountability Report Card.

There are three audiences for these test results: the state government's assessment of schools, school administrators' assessment of students and teachers, and parents' assessment of their children, teachers, and schools. The scores can obviously have curricular effects on the schools and psychological effects on the families.

API Scores

API scores fall within a possible range of 200 to 1000. The state has determined a target API for every school of 800. Any school with an API below 800 has a 1-year API growth target equal to 5% of the difference between its API and 800 or one point, whichever is larger. Thus a school with an API of 600 has an API growth target of 0.05(800 – 600) = 10, to 610. A school with an API between 780 and 800 has a growth target of 1 point, because 5% of the difference is less than 1 point. The target for a school with an API above 800 is to maintain its API above 800.

In addition to the API score, each school is ranked by deciles in comparison to 100 schools with similar characteristics, including pupil mobility, ethnicity, and socioeconomic status; the number of emergency credentialed teachers; and average class size.

API scores in the 1999–2000 and 2000–2001 school years were based solely on Stanford 9 scores in reading, language, spelling, and mathematics for students in Grades 2 through 8, and Stanford 9 scores in reading, language, mathematics, science, and social science for students in Grades 9 through 11. A school's API score was determined by the percentage of students in each of five quintiles determined by nationwide scores in 1995. These calculations are done for each of the Stanford 9 scores and the school's API is a weighted average of the total weighted scores. A truly average school that has 20% of its students in each quintile on each test will have an API of 655, well below the state's 800 target. A Lake Woebegone school,[2] with scores all above average and evenly distributed between the 50th and 99th percentile would have an API of 890.

In practice, the average API score in 1999 was 631. The growth targets give schools considerable time to reach the 800 target. A school with a 631 API that exactly met its growth target every year would take 62 years to reach an 800 API. It is problematic though whether 800 is a realistic target. Can every school be a Lake Woebegone school?

API scores in 2001–2002 were based on the Stanford 9 scores and the California Standards Test (CST) in English-language arts. API scores in 2002–2003 were based on the Stanford 9 scores, the CST in English-language arts and mathematics, and (for high school students) the CST in social sciences and the California High School Exit Examination in English and mathematics. As in 1999–2000 and 2000–2001, Stanford 9 scores were weighted using nationally normed quintiles. The CST compares performance to a standard rather than to the performance of a reference group.

[2] Garrison Keillor, host of the radio program *A Prairie Home Companion*, describes the fictitious town of Lake Woebegone as a place where "all the children are above average." This impossibility has been termed the "Lake Woebegone Effect" by educational researchers to identify the flaw in claims made by states that all of their schools perform above average on state tests (see, for example, Cannell, 1988).

Evidence of Regression

Scores that have important consequences should be interpreted properly. In particular, the interpretation of these scores, indexes, and rankings should be informed by an awareness of possible statistical, rather than educational, reasons for these results. A comparison of school APIs across years shows evidence of regression toward the mean. We will use the 2001–2002 and 2002–2003 API scores to illustrate these effects. A least squares regression of 2002–2003 API scores Y on the 2001–2002 API scores X (Figure 6) gives this estimated equation:

$$Y = 100.970 + 0.8655X, R^2 = 0.94 \tag{8}$$

The *t* value for the slope is a staggering 334.9, indicating a highly significant relationship. Equations 5 and 6 show that the 0.8655 slope is our estimate of the squared correlation between scores and ability; that is, the reliability of the test across schools. This 0.8655 slope means that a school that is 100 points from the mean API in 2001–2002 is predicted to be 86.6 points from the mean API in 2002–2003. Specifically, a school with an 2001–2002 API of 550 is predicted to have a 2002–2003 API of

$$Y = 100.970 + 0.8655(550) = 577$$

To put this into perspective, the growth target for a school with a 550 API is 550 + 0.05(800 – 550) = 562.5.

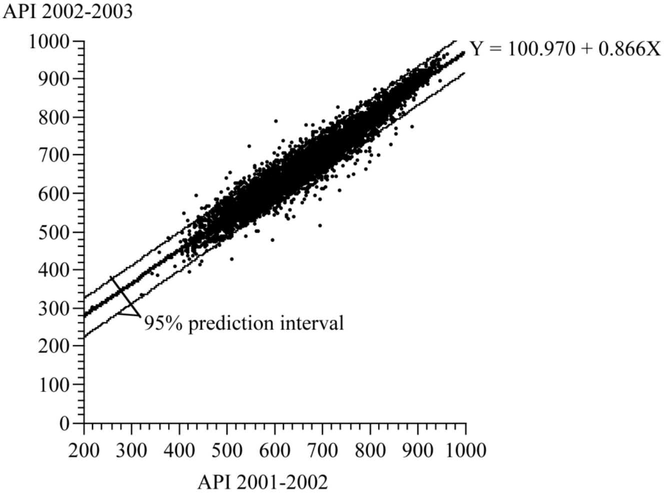

FIGURE 6 API scores in 2002–2003 and 2001–2002.

TABLE 3
Movements Out of Deciles for Similar Schools, 2001–2002 to 2002–2003

Decile	*Up*	*Same*	*Down*
10	0.0	54.9	45.1
9	21.5	24.8	53.7
8	28.7	22.4	49.0
7	31.0	16.5	52.5
6	39.0	16.3	44.6
5	44.9	16.2	38.9
4	50.0	14.6	35.4
3	50.2	21.3	28.5
2	51.3	24.9	23.8
1	49.2	50.8	0.0

Among those schools with below-average APIs in 2001–2002, 745 had their API scores fall and 2806 had their APIs increase in 2002–2003. Among those schools with above-average API scores in 2001–2002, 2060 had lower APIs and 1457 had higher APIs in 2002–2003. Overall, more scores increased than fell during this period, because the mean API increased from 681 to 691. Using standardized scores, among the 3,599 schools with negative Z scores in 2001–2002, 1,919 had higher Zs the next year and 1,680 had lower Zs; among the 3,592 schools with positive Z scores in 2001–2002, 1,946 had lower Zs the next year and 1,646 had higher Zs.

The state forms groups of 100 schools with similar characteristics and divides the schools in each group into deciles based on API scores. If regression to the mean is an important phenomenon, we expect significant year-to-year movements between deciles. Table 3 shows the movements out of deciles for similar schools between the 2001–2002 and 2002–2003 academic years. Decile 10 contains the most highly ranked schools, decile 1 the lowest ranked. At the extremes (deciles 10 and 1), there is nowhere to move but toward the middle and approximately half the schools do so. The movements are less lopsided for deciles nearer to the middle but it remains true that the schools moving toward the middle consistently outnumber those moving away from the middle. Regression to the mean!

Evidence of the Group Effect

Because changes in a school's API score are the cumulative result of a great many influences, the central limit theorem suggests an approximate normal distribution. The actual distribution of changes in API scores is roughly symmetrical but too fat in the tails, which strongly suggests that a group component causes individual student error scores within schools to be positively correlated. As noted earlier, such a group component can explain why there is significant regression to the mean even in aggregated data.

The change in API scores between 2001–2002 and 2002–2003 for the 7,191 schools with valid scores in both years has a mean of 9.36 and a standard deviation of 30.02. If these changes were normally distributed, a fraction 0.001350 of the observations would be more than 3 standard deviations above the mean change and a similar fraction would be more than 3 standard deviations below the mean change. Thus approximately 0.001350(7191) = 9.71 of the schools should be more than 3 standard deviations above the mean change and another 9.71 schools should be more than 3 standard deviations below the mean change. In practice, there were 30 schools with API changes larger than 9.36 + 3(30.02) = 99.42 points, with the largest increase being 185 points (5.84 standard deviations above the mean), and 28 schools with APIs that fell by more than 9.36 – 3(30.02) = –80.7 points, with the largest decline being 178 points (6.24 standard deviations below the mean). If the changes in school scores were independent and normally distributed, the probability that 58 or more schools would be more than 3 standard deviations from the mean is 1.48 10^{-12}.

Table 4 summarizes similar calculations for changes that are more than 3, 4, and 5 standard deviations from the mean. Of the 15 schools more than 4 standard deviations from the mean, 6 were above and 9 below the mean; of the 4 schools more than 5 standard deviations from the mean, 3 were above and 1 below the mean. These fat tails strongly suggest that group effects are important.

It is implausible that school abilities would change dramatically in a single year. Presumably, this is why the API growth targets are so modest. (As noted earlier, a school with an API of 600 has an API growth target of 10, to 610.) Could the average ability at 30 schools increase sufficiently between the 2001–2002 and 2002–2003 school years to cause scores to increase by 100 to 185 points in a single year? It is even harder to imagine that average ability at 28 schools dropped sufficiently to cause scores to fall by 81 to 178 points in 1 year. A more plausible explanation for the fat tails is that substantial correlations in error scores within schools caused scores to fluctuate greatly around ability.

If we assume only minor changes in school ability from 1 year to the next, further evidence of correlated error scores is provided by the illustrative calculations

TABLE 4
Large Changes in API Scores Between 2001–2002 and 2002–2003

		Number of Observations		
SD	*Probability*	*Expected*	*Observed*	*p-value*
3	0.0027	19.42	58	1.48 10^{-12}
4	0.0000634	0.456	15	3.11 10^{-13}
5	0.000000574	0.0041	4	1.19 10^{-11}

in Table 2. If student error scores within schools were uncorrelated ($\gamma = 0$), this reliability would be very close to 1 since the average score of hundreds of students would be an extremely accurate estimate of a school's ability. Here, the estimated reliability of these API scores across schools is in fact 0.8655. For the assumed parameter values in Table 2 and a 0.8 test reliability for individual students, the estimated 0.8655 reliability across schools is consistent with a 0.50 correlation for student error scores within schools.

INTERPRETING A SCHOOL'S AVERAGE TEST SCORE

A school's average test score is a measure of the level of achievement, not improvement. While it is certainly a laudable goal to have every student achieve some level of proficiency, should some schools be rewarded and others penalized for factors beyond their control using a statistic that does not measure how much students learn during school year?

This argument suggests that an authentic measure of educational value added should look not at the level of a school's average test score, but at changes in test scores over time. However, such scrutiny can be clouded by regression to the mean. Because observed test scores are an imperfect measure of ability, high scores are typically an overestimate of ability and low scores are typically an underestimate—causing high and low scores to regress to the mean in subsequent tests. If this statistical phenomenon is not taken into account, changes in tests scores over time may be misinterpreted as changes in ability rather than fluctuations in scores about ability. These two issues—the desire to assess changes in abilities and the need to account for regression to the mean—suggest a different way of using test scores to evaluate schools.

Reasonable Growth Targets

The regression argument can be used to predict the magnitude of the drop-off in above-average scores and the improvement in below-average scores. The appropriate question is whether the observed regression is larger or smaller than that predicted by purely statistical arguments. Instead of focusing on a school's overall score, we can examine each school's scores by grade level; for example, the test scores for a school's fourth graders. Then, instead of comparing their average score to the average scores of other fourth graders statewide or to the scores of fourth graders at schools with similar demographic characteristics, we can compare their fourth-grade scores to their earlier scores. If those students who were above average slip less and those who were below average improve more than predicted by regression to the mean, the school is succeeding.

Specifically, suppose that the test's reliability for third-grade API scores is ρ^2 and that a group of California third-grade students (who have third-grade and fourth-grade scores) have an API of X in a year when the average third-grade API is $\overline{X}$ statewide. (Alternatively, we could use the average third-grade API for schools with similar characteristics.) We then estimate the ability of this school's third graders by Equation 4:

$$\hat{\mu} = \left(1 - \rho^2\right)\overline{X} + \rho^2 X$$

This school would demonstrate value added if, as fourth graders, these students' API is above $\hat{\mu}$.

To illustrate, we can now interpret the abilities and scores in Table 1 as school API scores. Consider a group of third-grade students with an API of 775 in a year when the average third-grade API is 650 and the test's reliability for third-grade API scores is $\rho^2 = 0.89$. The estimated ability of this school's third graders is 761:

$$\begin{aligned}\hat{\mu} &= \left(1 - \rho^2\right)\overline{X} + \rho^2 X \\ &= 0.11\left(650\right) + 0.89\left(775\right) \\ &= 761\end{aligned}$$

This school would demonstrate value added if, as fourth graders, these students' API is above 761. Notice particularly that if the API score falls from 775 to 770, this is actually good news in that the score did not fall as much as predicted by regression to the mean. There is no way to know for certain whether the 770 score is due to effective teaching or a fortuitous match between the test and the school's curriculum. All we can say is that, even though the score is lower than the score a year earlier, it is higher than the regression-adjusted estimate of the school's ability a year earlier.

If we were to make a new estimate of ability based on the latest score, we would regress this 770 score toward the previous 761 estimate of ability:

$$\begin{aligned}\hat{\mu} &= 0.11\left(761\right) + 0.89\left(770\right) \\ &= 769\end{aligned}$$

Thus the 770 score leads us to revise our estimate of the school's ability from 761 to 769. In general, the estimate of ability is revised upward or downward depending on whether the most recent score is above or below the previous estimate of ability.

Figure 4 summarizes the relationship between fourth-grade API scores and estimated ability for our hypothetical data. Because these data assume no change in abilities between third and fourth grade, 10 schools are above the line and 10 below. If there were changes in abilities (up or down), the results could obviously be quite different. For example, if all of the scores were 50 points higher, 18 of the 20 schools would have scores above their previously estimated abilities—evidence that abilities increased.

For the actual California 2001–2002 API scores, Equation 4 with a 0.8655 reliability and 681.33 mean gives this regression-to-the-mean equation for estimating a school's ability from its 2001–2002 score:

$$\begin{aligned}\hat{\mu} &= \left(1-\rho^2\right)\bar{X}+\rho^2 X\\ &= \left(1-0.8655\right)681.33+0.8655X\\ &= 91.61+0.8655X \end{aligned} \tag{9}$$

It is a sign of success if a school's 2002–2003 score is above its estimated ability, as given by Equation 9. In practice, 4,749 schools scored above their estimated ability and 2,442 scored below, evidence that abilities did increase, on average, between these 2 school years.

A perhaps more easily explained procedure is to look, as in Figure 6, at a least squares regression of 2002–2003 API scores on 2001–2002 API scores. The least-squares Equation 8

$$Y = 100.970 + 0.8655X$$

has the same 0.8655 slope as the regression-to-the-mean Equation 9 but the intercept is 9.36 points higher because the mean score increased by 9.36 points, from 681.33 to 690.69.

Schools with APIs above the least-squares line did better than predicted in 2002–2003, taking into account regression to the mean (the 0.8655 slope) and their student population (as gauged by the base-year API score). The two (slightly) curved lines in Figure 6 are 95% prediction intervals for the 2002–2003 scores. Schools outside these bands displayed statistically significant changes (up or down) in their API scores.

A comparison of school scores to the least-squares line is equivalent to a comparison of school scores to the regression-to-the-mean line if the sample means are equal, because Equations 8 and 9 then coincide. The advantage of using the regression line is that its logic is compelling. The advantage of using the least squares line is that reports of ability estimates that have been adjusted for regression to the mean add layers of complexity to be explained to parents, educators, and politicians. One way to tiptoe around this issue is for the state to report "adjusted" scores that are regression-adjusted estimates of ability. The regression process could be explained in a statistical appendix, much as the details of score calculations are relegated to lengthy appendixes that are read only by those with an appetite for statistics.

There are several possible ways of combining scores across grades to evaluate a school as a whole. For example, we might sum across grades each school's deviations from the regression or least squares lines. Another possibility is to calculate a current API for the entire school and also calculate a base API using the students' scores in the previous year. We can then estimate a regression (or least squares) line using data for all schools (or for similar schools) relating current API to the base API. Educational value added is gauged by the school's deviation from this regression line.

CONCLUSION

There is a substantial literature on regression to the mean in individual test scores, though its lessons are often ignored. In this article, we show that if error scores are positively correlated, there can be substantial regression to the mean in group scores, which are commonly used to judge the effectiveness of different teaching methods, teachers, and schools. For example, many states now administer high-stakes tests that are used by parents, school administrators, and state education officials to assess students, teachers, and schools.

In some cases, schools are judged by the level of their test scores, based on the presumption that what matters is the product (Can they read?), not the process (How did they learn to read? Could they already read?). In other cases, schools are judged by changes in test scores over time. In either case, state-level administrators and policymakers should work to create accountability systems that are informed by an appreciation of regression to the mean. As noted earlier, this does not mean that regression has to become part of the public discourse in discussing test scores; rather, the formulas used in calculating school scores that form the basis of their rankings should take regression effects into account and be made available to those interested but not at the expense of providing a meaningful estimate of abilities for teachers, parents, and other stakeholders.

When schools are judged by the level of their scores, these scores should be shrunk toward the mean because observed differences in scores overstate differences in abilities. Schools with scores that are high relative to other schools are likely to have scores that are also high relative to their students' ability; schools with relatively low scores are likely to have scores below their students' ability. This observation is also relevant for comparing scores over time, because the base-year benchmarks should be shrunk toward the mean. Changes in ability are best assessed not by whether this year's score is higher than last year's score but by whether this year's score is higher than last year's estimate of ability—which has been adjusted for regression to the mean. If students who were above average slip less and students who were below average improve more than predicted by regression to the mean, the school is succeeding.

Of course, as long as observed scores are an imperfect measure of ability, some schools will appear to be more successful and some less successful than they really are. This observation cautions against overreacting to a single year's scores. Evidence accumulated over several years is likely to be more reliable than evidence from a single year. Our argument is that we should be looking at the right evidence—test scores that take regression to the mean into account.

ACKNOWLEDGMENTS

Support for this research was provided by the Haynes Foundation.

REFERENCES

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist, 24,* 409–429.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice, 7*, 5–9.

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation* (3rd ed.). San Diego CA: Educational and Industrial Testing Services (EdITS).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University.

Lee, M., & Smith, G. (2002). Regression to the mean and football wagers. *Journal of Behavioral Decision Making, 15,* 329–342.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores.* Reading, MA: Addison-Wesley.

Maddala, G. S. (1992). *Introduction to econometrics* (2nd ed.). New York: Macmillan.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Schmittlein, D. C. (1989). Surprising inferences from unsurprising observations: Do conditional expectations really regress to the mean? *The American Statistician, 43,* 176–183.

Thorndike, R. L. (1963). *The concepts of over- and under-achievement.* New York: Teachers College, Columbia University.

Wainer, H. (1999). Is the Akebono School failing its best students? A Hawaii adventure in regression. *Educational Measurement: Issues and Practice, 18*, 26–31, 35.