Regression to the Mean: How Big Is It?

John R. Hills' article, "Regression Effects in Educational Measurement" [Educational Measurement: Issues and Practice 12(3), 31–34], was informative and imaginatively presented. Professionals in the field are well acquainted with the phenomenon of regression toward the mean either through standard texts or bitter experience. What is often not known is the magnitude of such effects.

Some years ago, I had a highly instructive experience about the magnitude of a regression effect. A colleague of mine who is a specialist in the education of the gifted received a letter from the Chair of the California State Assembly Legislative Subcommittee that had recently sponsored legislation to provide funds for the education of all gifted children in the state. The amount of money involved was considerable. A year after the program had started, the Chair received a letter from an official in one school district in the state reporting on how the program had worked in his district. Students had been tested at the end of the school year with a group scholastic ability test, and the top 3% of the students were selected for the program. The program ran for the next entire school year. At the end of the school year, the test was readministered to all students in the school to select the students who would be included in the program for the following year. To the horror of the school official, the students who had been in the program that year had shown an average decline of eight IQ points. It seemed that the program was making the students dumber! The Chair of the Assembly Subcommittee, obviously distressed, wrote to my colleague for guidance.

My colleague showed me the letter and asked how he should respond. I replied that this was an example of regression to the mean and suggested that he write the Chair of the Subcommittee and tell him so. My colleague was not willing to accept my explanation without some supporting evidence. I recalled Terman’s longitudinal study of gifted children and suggested that we check the results of that study. Fortunately, my colleague had the first five volumes of that study on his shelf. Within minutes, we located the results of the 1927–28 retesting of the sample that was originally identified in 1920–21 (Burks, Jensen, & Terman, 1930). The mean decrease in IQ was eight points! It was identical to the mean decrease of the gifted students in the school district. What is notable about this example is that two test–retest studies of changes in the intelligence of gifted school children in the state of California, done 50 years apart, yielded identical results. Because the two intelligence tests had a standard deviation of 16 points, the regression to the mean, from about two standard deviations above the mean, was about one half of a standard deviation. I have used this as a rough guide to estimate the magnitude of the regression to the mean in subsequent data sets that I have examined.

Of course, an analytical solution to the problem would be ideal. Such a solution would need to take into account the magnitude of the difference of a score from the mean, the reliability of the tests used, the correlation between scores on the first and second testings, the age of the students at the initial testing, and perhaps the length of the interval between testings. Until then, we need to be aware of the phenomenon, as Hills has so ably shown in his article.

Richard M. Wolf
Teachers College, Columbia University

Gain Score Grading Revisited

In one of John Hills’ scenarios (1993) intended to illustrate the effect of regression toward mediocrity, a teacher gave a pretest and a posttest, calculated gain scores, and based grades on those gain scores. According to Hills, “Without a doubt...those scoring low on the pretest will tend to get good grades, and those scoring high on the pretest will tend to get low grades (p. 31).”

This seems to assert that the correlation between pretest scores and gain scores must be negative. This is not, in fact, necessarily the case, although it seems to be a pervasive misconception, for the same assertion can be found in a number of statistical textbooks (e.g., Cohen & Cohen, 1983, p. 73; Glass & Hopkins, 1984, p. 129).

Consider a pretest score X, a posttest score Y, and a gain score G = Y - X. Then Cov(X,G) = Cov(X,Y) - Var(X). Dividing both sides by Var(X), b_G,Y = b_Y,X - 1. The sign of the simple regression coefficient predicting gain score from pretest score (and therefore of the correlation between gain and pretest) thus depends on whether b_Y,X, the regression coefficient predicting posttest from pretest, is greater than or less than unity.

Now let h = s_G/s_X and k = s_Y/s_X. Var(G) = Var(Y) - Var(X) = Var(X) + Var(Y) - Cov(X,Y). Dividing both sides by Var(X), h^2 + 1 + k^2 - 2rhk = 1. Since b_G,Y = rhk and b_Y,X = rhk, r_g = (rhk - 1)/h, or

r_g = 1 - (h - 1)/\sqrt{h^2 + 1 - 2rhk}.

This is an explicit function of two arguments: the correlation between pretest and posttest, and the ratio of the standard deviations of the two tests. (Equation 2.11.9 in Cohen and Cohen [1983, p. 73] is a simplified version of this formula for the special case of k = 1.) Figure 1 displays r_g as a function of r_Y,X for selected values of h.

When posttest variance is smaller than pretest variance (h < 1), the

References


pretest : gain correlation is always negative.

When posttest variance is larger than pretest variance \((k > 1)\), the pretest : gain correlation is positive whenever \(r_{xy} > 1/k\), or \(b_{yx} > 1\).

For the special case where \(\text{Var}(X) = \text{Var}(Y) (k = 1)\), the pretest : gain correlation \(r_{xy}\) is indeed negative (unless the pretest correlates perfectly with the posttest, in which case \(r_{xy}\) is zero). We conjecture that this special case is the source of the misperception that gain scores must be negatively correlated with pretest scores. For the classical measurement situation of parallel forms administered under parallel conditions, it is expected that the means and variances of parallel scores \(X\) and \(Y\) would be equal; then difference scores between the parallel forms are negatively correlated with whichever parallel form is taken as pretest.

But is it to be expected that pretest and posttest variances would be equal in the situation described in the scenario? We think not. Consider two stereotypical cases.

A. Mastery of Minimum Competence
If the purpose of the instruction is mastery of a minimum level of competence, it is reasonable for most of the items on the tests to be at, or close to, the level of difficulty corresponding to minimum competence. In this case, the posttest variance may be small or even zero (due, from a point of view, to ceiling effects), and those who had some prior knowledge of the content (as reflected in higher pretest scores) would indeed show smaller gains than those who had none. But then, gain scores are not a proper basis for grades, because the purpose is for all students to attain mastery of the minimum competence. This purpose is not consistent with the expressed intent of the teacher, who "gives grades based on the size of the improvement score" (p. 31). Grades based on the amount of improvement cannot reflect an absolute level of performance, but the idea of mastery learning implies that absolute level of (final) performance is the appropriate basis for grades. (The benefit to those who started out ahead of the pack resides in reaching the mastery level in shorter time than their colleagues require, or with less effort, or both.)

B. Maximizing Individual Potential
If the underlying goal is to help each student go as far as possible—that is, to maximize each individual's potential for learning—one would expect both the instruction and the tests to be much wider ranging than in Case A: The instruction must both allow opportunities for advanced students to advance further and provide opportunities for the less skilled to become more skilled, and the tests must provide opportunities for the gamut of students to demonstrate the proficiency they have attained. Under these conditions, it may be reasonable to base grades on gain scores (although it would not be fair if either (a) the advanced students could not learn much from the instruction or could not show their learning on the test for reasons of faulty test design or (b) less skilled students found no instruction at their level of discourse or couldn't show improvement because the posttest items were largely too difficult for them). In such circumstances, posttest variance might well be greater than pretest variance, and positive correlation between pretest and gain scores would be expected, so that those scoring high on the pretest would not be prevented from earning high gains.

Thus, the common belief that the correlation between pretest and gain score must always be negative is not correct, except under circumstances that are unlikely to obtain in the situation described in this scenario.

Dennis M. Roberts
Penn State University
Donald F. Burrill
The Ontario Institute for Studies in Education

References