Stability of School Building Accountability Scores and Gains

CSE Technical Report 561

Robert L. Linn
CRESST / University of Colorado at Boulder

Carolyn Haug
University of Colorado at Boulder

April 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.2 Comprehensive Systems for Accountability and the Measurement of Progress
Robert L. Linn, Project Director, CRESST/University of Colorado at Boulder

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

STABILITY OF SCHOOL BUILDING ACCOUNTABILITY SCORES AND GAINS

Robert L. Linn
CRESST/University of Colorado at Boulder

Carolyn Haug
University of Colorado at Boulder

Abstract

A number of states have school building accountability systems that rely on comparisons of achievement from one year to the next. Improvement of the performance of schools is judged by changes in the achievement of successive groups of students. Year-to-year changes in scores for successive groups of students have a great deal of volatility. The uncertainty in the scores is the result of measurement and sampling error and nonpersistent factors that affect scores in one year but not the next. The level of uncertainty was investigated using fourth-grade reading results for 4 years of administration of the Colorado Student Assessment Program. It was found that the year-to-year changes are quite unstable, resulting in a near-zero correlation of the school gains from Years 1 to 2 with those from Years 3 to 4. Some suggestions for minimizing volatility in change indices for schools are provided.

Most state accountability systems that report school-building current status based on aggregate student assessment results also include some basis for rating improvement in achievement. A few states base their estimates of improvement on longitudinal results obtained either by tracking individual students from year to year, as is done, for example, in Tennessee, or by comparing the performance of students attending a school in a given year at, say, Grade 5, with the performance of students attending that school the previous year in Grade 4, as is done, for example, in North Carolina. The most common way of monitoring improvement, however, is through the comparison of successive groups of students. For example, the performance of students in Grade 4 in one year may be compared to the performance of Grade 4 students in that school the previous year.

A substantial number of states, including California, Colorado, Kentucky, Maryland, and Washington, use the successive groups approach to compare the

achievement of students at selected grades in a given year or biennium with that of students from previous years at the same grade level in the same school. The school-level changes that are found provide a means of recognizing that schools serve students who start at different ability levels. These comparisons of student performance at a grade level in different years rest on the implicit assumption that student characteristics that affect achievement levels are relatively stable from year to year for students attending a given school. This assumption is questionable for schools serving neighborhoods whose demographic characteristics are changing rapidly, but is a reasonable approximation for most schools.

Unfortunately, changes in scores for the students tested at a given grade from one year to the next can be quite unreliable. There are several sources of the unreliability. First, the school summary scores for each year are subject to measurement and sampling error. Second, difference scores tend to be less reliable than the scores used to compute differences. Third, the between-school variability of change scores is considerably smaller than the between-school variability of the scores for a given year. Fourth, as Kane and Staiger (2001) have shown, a substantial part of the variability found in change scores for schools is due to nonpersistent factors that influence scores in one year but not the other.

Using data from the state of North Carolina, Kane and Staiger (2001) estimated that, for the smallest quintile of schools, 79% of the between-school variability in year-to-year changes in fourth-grade reading plus math scores was due to a combination of sampling variability and other nonpersistent factors. The corresponding percentage for the largest 20% of the schools was only slightly smaller (73%). In other words, only about a fifth to a fourth of the observed between-school variability in school change scores was attributable to persistent factors having to do with the school.

Colorado Student Assessment Program

Colorado introduced a new statewide assessment system in 1997 called the Colorado Student Assessment Program (CSAP). In that year, CSAP was limited to tests in reading and writing administered to fourth-grade students. Since that time, additional subjects and grades have been added. In 2001, reading, writing, and mathematics were assessed in Grades 5 through 10, reading and writing were assessed in Grade 4, and reading was assessed in Grade 3. Since Grade 4 reading and writing tests were introduced first, trends in student performance in those two

subjects at Grade 4 can be tracked for the greatest number of years. Here we will focus on Grade 4 reading. Through the spring 2000 administration, CSAP Grade 4 reading results were available for schools for 4 years.

Three performance standards have been set for reporting CSAP results. The standards divide the test scores into four regions that are labeled unsatisfactory, partially proficient, proficient, and advanced. Colorado school district accreditation rules in place prior to June 2001 set a target for schools to have at least 80% of their students in the proficient or advanced performance level. Although few schools are at those levels now, the 80% figure provided a goal for the future. Schools with percentages below the 80% figure could still be accredited if there were a 25% increase over the base-line percentage in a 3-year period.

In June 2001, a new approach to the use of CSAP results for school district accreditation was adopted that makes use of a weighted index of all performance levels. Specifically, the weighted index is equal to 1.5 times the percentage of students in the advanced category plus 1.0 times the percentage who are proficient plus 0.5 times the percentage who are partially proficient minus 0.5 times the percentage in the unsatisfactory category minus 0.5 times the percentage of students with no test scores. Because both the percentage of students in the proficient or advanced performance level and the new weighted index are apt to be important for accountability purposes in the future, we use both in the analyses reported below.

CSAP results. Table 1 shows the number of schools and the unweighted means and standard deviations of the percentage of students scoring in the proficient or advanced level on the fourth-grade reading assessment for each of the 4 years from 1997 to 2000. It also shows the means and standard deviations for the weighted index scores. As can be seen, on average, slightly more than half of the students scored at the proficient level or higher each year. The mean percentage was essentially unchanged from 1997 to 1998 but then increased by 2.5% from 1998 to 1999 and by another 1.4% from 1999 to 2000. The standard deviations of the school percentages were relatively stable, ranging from 18.51 to 19.26 over the 4 years. The weighted index score started at 68 in 1997 and increased each of the following 3 years, albeit only slightly from 1997 to 1998.

Table 1

Descriptive Statistics for Percentage of Students Scoring at the Proficient or Advanced Level and for the Weighted Index Score (Grade 4 Reading)

		Percent proficient or advanced		Weighted index	
Year	Number of schools	Mean	Standard deviation	Mean	Standard deviation
1997	757	56.8	18.73	68.0	23.00
1998	770	56.7	18.51	68.3	21.08
1999	788	59.2	19.26	71.0	21.72
2000	802	61.6	18.72	74.4	21.30

The gains in percentage of students in the proficient or advanced performance level or in the weighted index score from one year to the next, of course, varied from one school to another. The differences in percentages and in the weighted index scores were computed for each school from 1997 to 1998, from 1998 to 1999, and from 1999 to 2000. Means and standard deviations for those differences are reported in Table 2. Schools with differences in the proficient or advanced level one standard deviation above the mean difference gained 11.7% from 1997 to 1998, 13.2% from 1998 to 1999, and 13.5% from 1999 to 2000. On the other hand, schools with differences a standard deviation below the mean declined by 12.1% from 1997 to 1998, by 8.3% from 1998 to 1999, and by 8.5% from 1999 to 2000. Using the weighted index scores, schools one standard deviation above the mean gained 14.9 points from 1997 to 1998, 14.9 points from 1998 to 1999, and 15.2 points from 1999 to 2000. The corresponding losses for schools one standard deviation below the mean in change in index scores were 14.2, 9.7, and 8.4 points.

Table 2

Descriptive Statistics for Year-to-Year Differences in the Percentage of Students Scoring in the Proficient or Advanced Level and in the Weighted Index Scores on the CSAP (Grade 4 Reading)

		Percent proficient or advanced		Weighted index	
Year	Number of schools	Mean	Standard deviation	Mean	Standard deviation
1998–1997	744	-0.2	11.91	0.3	14.58
1999–1998	763	2.4	10.75	2.6	12.27
2000–1999	776	2.5	11.03	3.4	11.82

As can be seen in Table 3, there is a relatively strong relationship between the percentage of students in the proficient or advanced level in one year and the corresponding percentage in another year during the 4 years under study. The correlations of the school percentages for the 4 years are shown in Table 3. The number of schools for these correlations ranged from a low of 744 for the correlation of 1997 results with 1998 results to a high of 776 for the correlation of 1999 results with 2000 results. As can be seen, the lowest correlation was .796, between the percentages in 1997 and those in 1998. All of the correlations are at least .80 or higher when rounded to two decimal places.

The correlations of the weighted index scores for schools from year to year were similar in magnitude to those obtained for the percentage of students in the proficient or advanced level (see Table 4).

As is clear from the magnitude of the standard deviations of the year-to-year differences in school percentages of proficient or advanced shown in Table 2, there is substantial between-school variability in the changes in both the percentage proficient or advanced and the weighted index scores. Nonetheless, the magnitude

Table 3

Correlations of the Within-School Percentages of Students in the Proficient or Advanced Level on the CSAP Across Years (Grade 4 Reading)

Year	1997	1998	1999	2000
1997	1.000			
1998	.796	1.000		
1999	.816	.837	1.000	
2000	.797	.824	.830	1.000

Table 4

Correlations of the School Weighted Index Scores for the Grade 4 Reading CSAP Across Years

Year	1997	1998	1999	2000
1997	1.000			
1998	.785	1.000		
1999	.821	.835	1.000	
2000	.803	.817	.846	1.000

of the percentage or the weighted index score one year can be predicted relatively accurately from knowledge of the percentage or the weighted index score in another year.

As would be expected, the difference in percentage from one year to the next, however, is negatively related to the magnitude of the percentage proficient or advanced in the first year. The change from 1997 to 1998 is correlated -.35 with the percentage proficient or advanced in 1997. Corresponding correlations of the changes from 1998 to 1999 and from 1999 to 2000 with the percentage proficient or advanced in the first year are -.23 and -.34, respectively. Thus, schools with a relatively high percentage of students scoring proficient or advanced in the base year are likely to have smaller gains than schools with a relatively low percentage proficient or advanced in the base year. For example, a school with 10% of its students scoring proficient or advanced in 1997 typically doubled that percentage in 1998, whereas a school that started with 80% proficient or advanced in 1997 typically had a decline in percent proficient or advanced of about 8% in 1998. Clearly, the expected change depends on the starting percentage. Moreover, regardless of starting position, schools that gain a lot from Year 1 to Year 2 generally will show a decline in Year 3, while those that show a decline from Year 1 to Year 2 generally will show a gain in Year 3. The change in percent proficient or advanced from 1997 to 1998 has a correlation of -.49 with the corresponding change from 1998 to 1999. Similarly, the change from 1998 to 1999 has a correlation of -.49 with the change from 1999 to 2000. Negative correlations between changes from Year 1 to 2 and changes from Year 2 to 3 are to be expected, of course, since the score for Year 2 has a plus sign in the first difference and a minus sign in the second difference.

The weighted index scores have similar properties. The correlation of the change in index scores from 1997 to 1998 with the change from 1998 to 1999 was -.51, and the latter change was correlated -.45 with the change from 1999 to 2000. Thus, it should not be surprising that schools that show outstanding gains using either the percentage of students who are proficient or advanced or the weighted index score from one year to the next do not look so good with respect to their gains the following year. Conversely, a school that loses ground from Year 1 to Year 2 and might be identified as in need of assistance will likely rebound with a gain in Year 3.

Volatility of change scores. The change scores are also much less stable than the scores for a single year. To investigate this lack of stability of change scores with the CSAP data, we computed change scores based on 2-year intervals. That is, we

subtracted the percent proficient or advanced in 1997 from the corresponding percent in 1999 (Change 97 to 99). Similarly, we subtracted the 1998 percent proficient or advanced from the corresponding percent in 2000 (Change 98 to 00). In this way, we created two change scores that did not share a percent for a given year. The correlation between change 97 to 99 and change 98 to 00 for the 734 schools with scores in all 4 years was -.03 for the percentage of students in the proficient or advanced level and -.05 for the weighted index score. In other words, there is a complete lack of stability in the 2-year change scores. Knowing the magnitude of the gain or loss in percent proficient or advanced from 1997 to 1999 tells you essentially nothing about the change from 1998 to 2000.

Because so much of the variability in school change scores is attributable to noise, it should not be surprising that schools identified as outstanding in one change cycle for achieving a large change in achievement are unlikely to repeat that performance in the next cycle. The converse is also true. Thus, schools that are identified as needing assistance in one cycle because they fell short of their change target, or even showed a decline, are unlikely to fall in that category the next change cycle. A consequence of this random fluctuation from one change cycle to the next is that the actions taken to assist schools in the latter situation may appear to be more effective than they actually are. Moreover, it is likely to be a mistake to assume that the practices of the schools recognized as outstanding are ones that should be adopted by other schools.

School Size Effects

The noise in year-to-year changes in percentage of students scoring at the proficient or advanced level is quite large in comparison to the between-school variability in change scores for all schools. The magnitude of the noise is especially large for small schools. This is illustrated in Figure 1. The box plots in Figure 1 show the distribution of the differences in percentages from 1997 to 1998 for schools that have been divided into five groups according to number of fourth-grade students in 1997. The first box plot on the left shows the distribution for the 79 schools with 30 or fewer fourth-grade students. The next three box plots display the distributions for the 227 schools with between 31 and 60 fourth-grade students, the 287 schools with between 61 and 90 fourth-grade students, and the 127 schools with 91 to 120 fourth-grade students. The box plot to the far right displays the results for the 24 schools with 121 or more fourth-grade students. As can be seen, the median gain for all five

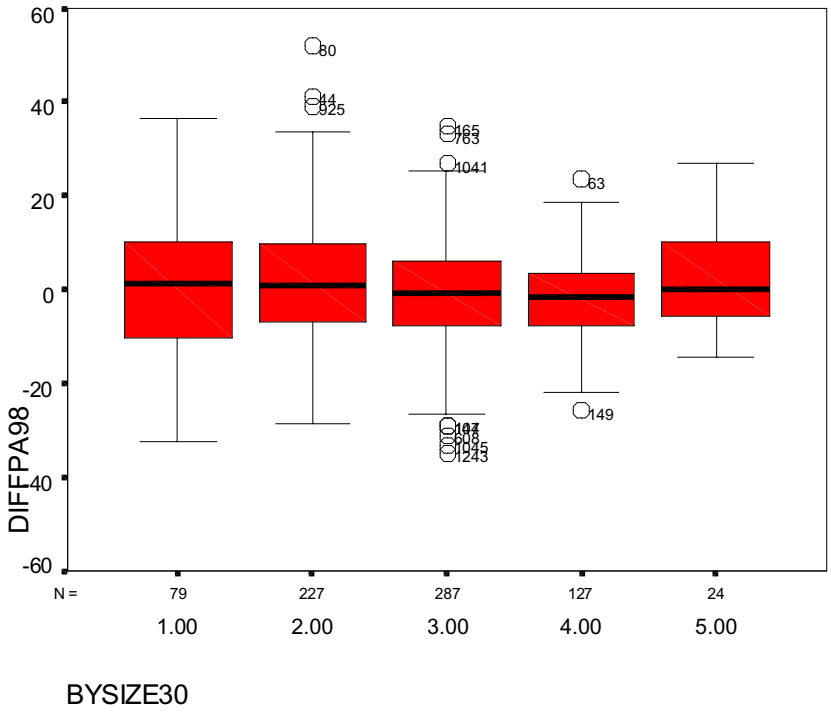

Figure 1. Plot of changes in percentages of students in a school at the proficient or advanced level from 1997 to 1998 as a function of school size in 1997.

clusters of schools based on school size is close to zero. The spread of positive and negative difference scores tends to decrease from left to right in the figure, corresponding to the fact that the variability of school changes in percentages is larger for small schools than for large schools. Large schools are less likely to be found to have either extreme increases or extreme declines in the percentage of students scoring proficient or advanced than are small schools. Thus, one would expect to find a disproportionate number of small schools that are found to be most wanting as well as those that are found most praiseworthy in terms of the changes in percentage of students who are proficient or advanced from one year to the next. A similar pattern was found for the weighted index scores.

Conclusion

The performance of successive cohorts of students is used in a substantial number of states to estimate the improvement of schools for purposes of accountability. The estimates of improvement are quite volatile, however. This volatility results in some schools being recognized as outstanding and other schools

identified as in need of improvement simply as the result of random fluctuations. It also means that strategies of looking to schools that show large gains for clues of what other schools should do to improve student achievement will have little chance of identifying those practices that are most effective. On the other hand, schools that are identified as in need of improvement generally will show increases in scores the year after they are identified simply because of the noise in the estimates of improvement—not because of the effectiveness of the special assistance provided to the schools or pressure that is put on them to improve.

The lack of precision in estimates of school improvement based on comparisons of successive groups of students presents a major challenge. Several ways of dealing with this challenge seem worthy of consideration. At a minimum, reports of accountability results for schools need to be accompanied by information about the dependability of those results as required by the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). This might best be done where schools are placed into graded performance categories by reporting information about the accuracy of classifications. Procedures for evaluating school-building misclassification probabilities are described by Rogosa (1999) and by Hoffman and Wise (2000).

Improvements in the accuracy of results can be achieved by combining data across multiple grades, multiple subject areas, and/or multiple years. Combining across either grades or years increases the precision of results by increasing the number of students used to estimate school results. Combining across grades has the added advantage of increasing the number of teachers who are teaching students whose performance directly contributes to the accountability results for the school and thereby may increase the sense of shared responsibility of results. Although combining across subject areas and grades glosses over relative strengths and weaknesses by subject area and grade level, it is a reasonable approach for obtaining an overall school accountability index and does not preclude the separate reporting of results by grade and subject area, or diminish the importance of the separate reports. Combining across several years lengthens the accountability cycle, but produces results that are more trustworthy and therefore more likely to lead to real long-term improvements and to the identification of exemplary practices as well as enhancing fairness. As is true in a variety of other states, Colorado will combine

across multiple grades and subjects in the computation of the weighted index scores for accreditation purposes.

The precision of estimates also can be improved by the use of more sophisticated analytical techniques. For example, Kane and Staiger (2001) demonstrated this by using “filtered” estimates of school gains. The filtered estimates, which are based on an application of empirical Bayes procedures, are more complicated and therefore less transparent than estimation procedures commonly in use. The loss of transparency seems a good tradeoff for the gain in precision that Kane and Staiger have demonstrated.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Hoffman, R. G., & Wise, L. L. (2000). *School classification accuracy final analysis plan for the Commonwealth accountability and testing system*. Alexandria, VA: HumRRO.

Kane, T. J., & Staiger, D. O. (2001). *Volatility in school test scores: Implications for test-based accountability systems.* Paper presented at a Brookings Institution conference.

Rogosa, D. (1999). *Reporting group summary scores in educational assessments: Properties of proportion at or above cut-off (PAC) constructed from instruments with continuous scoring* (Draft deliverable). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.