Stability of School-Building Accountability Scores and Gains
Robert L. Linn and Carolyn Haug

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2002 24: 29
DOI: 10.3102/01623737024001029

The online version of this article can be found at:
http://epa.sagepub.com/content/24/1/29

Published on behalf of

American Educational Research Association

and

SAGE

http://www.sagepublications.com

Additional services and information for Educational Evaluation and Policy Analysis can be found at:

Email Alerts: http://eepa.aera.net/alerts

Subscriptions: http://eepa.aera.net/subscriptions

Reprints: http://www.aera.net/reprints

Permissions: http://www.aera.net/permissions

Citations: http://epa.sagepub.com/content/24/1/29.refs.html
Stability of School-Building Accountability Scores and Gains

Robert L. Linn  
Center for Research on Evaluation, Standards, and Student Testing, University of Colorado at Boulder

Carolyn Haug  
Colorado Department of Education

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 schools in Colorado for four 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 one to two with those from years three to four. Some suggestions for minimizing volatility in change indices for schools are provided.

Keywords: accountability, assessment, changes in performance, school performance, stability of test results, testing

Most state accountability systems that report school-building current status based on aggregate student assessment results also includes some basis for rating improvement in achievement. A few states base their estimates of improvement on longitudinal results that track individual students from year to year, as is done for example in North Carolina and Tennessee. The most common way of monitoring improvement, however, is through the comparison of successive groups of students. For example, the performance of students in fourth grade in one year may be compared to the performance of fourth grade students in that school the previous year.

In the past decade or so quite a few states have moved away from norm-referenced tests and the use of norms to report school results. Instead they have set performance standards and associated cut scores on their tests and started reporting results in terms of those performance standards. Standards-based reporting is generally done in two ways: (a) by reporting the percentage of students scoring in each score region defined by the cut scores (e.g., advanced, proficient, partially proficient, and below partially proficient) and (b) by reporting the percentage of students who score at or above the cut score, sometimes referred to as Percent Above Cut or PAC corresponding to “proficient” or “meets the standard.” In some states an index score is also reported based on the distribution of student scores in the various performance categories.

The work reported herein was supported under the Educational Research and Development Centers Program, PR Award Number R305B600002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The finding and opinions expressed in this report do not reflect 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.
Although means or medians based on scale scores were more familiar ways of reporting in the past and have somewhat better statistical properties than PAC scores, the PAC scores have been widely used in recent years. Clearly there is some loss of information when scale scores are reduced to a dichotomy, or even to one of four performance categories, however, PAC scores still contain a good deal of information and have reasonably good statistical properties.

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 in the same school. Unlike the aggregate performance for a single year, the school-level changes from one year to the next provide a means of recognizing that schools serve students that start at different places as reflected by the different performance levels in the base year. These comparisons of student performance at a grade level in different years, however, rest on the implicit assumption that the characteristics of the students that affect achievement levels are relatively stable from year to year for the 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 error. Second, despite the fact that all, or almost all students are tested each year, they are subject to sampling error, because, as Cronbach, et al, (1997), have argued, for an assessment to be used as the basis for concluding “that a school is effective as an institution requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population” (Cronbach, et al., p. 393).

Third, difference scores tend, with some exceptions (see, for example, Rogosa & Willet, 1983), to be less reliable than the scores used to compute differences (e.g., Cronbach & Furby, 1970; Linn & Slinde, 1977). Fourth, the between-school variability, in the data reported by the authors, of change scores (in Table 2) is considerably smaller than the between-school variability of the scores for a given year (Table 1). Finally, as Kane and Staiger (2001) have shown, a substantial part of the variability found in change scores for schools is due to nonpersistent factors (e.g., an extended sick leave of a teacher, a teacher strike, a small group of disruptive students, or changes in the inclusion rules for tested students) that influence scores in one year but not the other.

Using data from the state of North Carolina, Kane and Staiger estimated that, for the quintile made up of the smallest 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 quintile made up of the largest schools was only slightly smaller (73%). In other words, only about one fifth to one fourth of the observed between-school variability in school change scores was attributable to persistent factors having to do with the school.

In the remainder of this article we use data from four years of administration of the Colorado Student Assessment Program (CSAP) fourth grade reading assessments to investigate the level of uncertainty in school building results for a given year and change in school building results from one year to the next. We conclude with some suggestions for minimizing the level of volatility in change indices for schools.

**Colorado Student Assessment Program**

Colorado introduced a new statewide assessment system called the Colorado Student Assessment Program (CSAP) in 1997. The CSAP tests are designed to measure student performance relative to the Colorado Model Content Standards, to the extent possible with a large-scale, paper and pencil assessment. CSAP tests consist of constructed-response and multiple-choice items. There is one form of each CSAP test, and 25% of the items on each test are released and replaced with new items annually.

In 1997 the 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 the spring 2002 administration, reading and writing will be assessed at each grade from 3 through 10, mathematics will be assessed at each grade from 5 through 10,
and science will be assessed at grade 8. Since fourth grade reading and writing tests were introduced first, trends in student performance can be tracked for the greatest number of years in those two subjects at that grade. Here we will focus on fourth grade reading. Through the spring 2000 administration CSAP grade fourth grade results were available for schools for four years.

The fourth grade-reading test is administered in three 50-minute sessions. In 2000, the fourth grade-reading test consisted of 70 items, 53 multiple-choice and 17 constructed-responses, with a total of 91 points. The 17 constructed-response items included two 1-point, eleven 2-point, two 3-point and two 4-point items. From 1997–2000, the scale for the fourth grade-reading test ranged from 300 to 720. The 2000 fourth grade-reading test has a mean of 506 and standard deviation of 46.0. This test has a reliability index (coefficient alpha) of 0.93. [Additional information about the technical aspects of the CSAP tests, can be found in the CSAP technical report (Colorado Department of Education, 2000).]

Three performance standards have been set for reporting CSAP results that 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 levels. Although few schools are at that level now, the 80% figure provided a goal for the future. Schools with percentages below the 80% figure could still be accredited if there was a 25% increase over the base-line percentage in a three-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 percent of students in the advanced category plus 1.0 times the percent proficient plus 0.5 times the percent partially proficient minus 0.5 times the percent unsatisfactory minus 0.5 times the percent of students with no test scores. Because both the percent of students in the proficient or advanced performance levels 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

The number of schools, the unweighted means, the standard deviations of the percentage of students scoring in the proficient or advanced levels, and of the weighted index scores on the fourth grade reading assessment are shown for each of the four years from 1997 to 2000 in Table 1. As can be seen, on average, slightly over one 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 18.51 to 19.26 over the four years.

The weighted index score started at 68 in 1997 and increased each of the following three years, albeit only slightly from 1997 to 1998. The gains in percentage of students in the proficient or advanced performance levels 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

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of schools</th>
<th>Percent Proficient or Advanced</th>
<th>Weighted Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>1997</td>
<td>757</td>
<td>56.8</td>
<td>18.73</td>
</tr>
<tr>
<td>1998</td>
<td>770</td>
<td>56.7</td>
<td>18.51</td>
</tr>
<tr>
<td>1999</td>
<td>788</td>
<td>59.2</td>
<td>19.26</td>
</tr>
<tr>
<td>2000</td>
<td>802</td>
<td>61.6</td>
<td>18.72</td>
</tr>
</tbody>
</table>
the percentage proficient or advanced equal to one standard deviation above the mean difference for all schools for the pair of years in question 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 for all schools 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 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. The number of schools is different in Tables 1 and 2 because of some schools being new in a given year or closing in a given year and the school had to be open both years to compute the differences reported in Table 2. The correlations of the school percentages for the four years are shown in Table 3. Table 3 reveals that there is a relatively strong relationship between the percentage of students in the proficient or advanced level in one year with the corresponding percentage in another for the four years under study. 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 for the percentages in 1997 with those in 1998. As shown, 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 levels (See Table 4). Although it is clear from the magnitude of the standard deviations of the year-to-year differences in school percentages of proficient or advanced as 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 of the percentage or the weighted index score in one year can be predicted accurately from knowledge of the percentage or weighted index score in another year.

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. Negative correlations of change with initial status are not unexpected, unless the variability of the scores increases substantially from year one to year two. As was shown in

### TABLE 2

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of schools</th>
<th>Percent Proficient or Advanced</th>
<th>Weighted Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>1998–1997</td>
<td>744</td>
<td>-0.2</td>
<td>11.91</td>
</tr>
<tr>
<td>1999–1998</td>
<td>763</td>
<td>2.4</td>
<td>10.75</td>
</tr>
<tr>
<td>2000–1999</td>
<td>776</td>
<td>2.5</td>
<td>11.03</td>
</tr>
</tbody>
</table>

### TABLE 3

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>.796</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>.816</td>
<td>.837</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>.797</td>
<td>.824</td>
<td>.830</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 1, the between-school standard deviations of percentages of students who are proficient or above are relatively stable from year to year. Hence, the correlations of gains with initial status are negative. The change from 1997 to 1998 is correlated $-0.35$ 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 $-0.23$ and $-0.34$, respectively.

The negative correlations of initial status with gains mean that 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%. Clearly, the expected change depends on the starting percentage, and more evidence of the regression effect observed when two variables are not perfectly correlated.

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 two-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 four years was $-0.03$ for the percentage of students in the proficient or advanced levels and $-0.05$ for the weighted index score. In other words, there is a complete lack of stability in the two-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.

Variability in school change scores is generally interpreted to be the result of real changes in the quality of education that is provided by a school. There are, however, many other factors that contribute to changes. Measurement error, differences in the student body from year to year, and nonpersistent factors, such as changes in

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>0.785</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>0.821</td>
<td>0.835</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.803</td>
<td>0.817</td>
<td>0.846</td>
<td>1.000</td>
</tr>
</tbody>
</table>
the teaching staff, contribute to the variability in school change scores. The above results indicate that these other factors make up a substantial part of the overall variability in school change scores. Hence, it should not be surprising that schools identified in one change cycle as outstanding 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 to need assistance in one cycle because of falling short of their change target, or even showing 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 category 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

Factors other than real change in the quality of education provided can be characterized as noise that interfere with a clear reading of the signal, i.e., the real changes in quality. The noise in year-to-year changes in percentage of students scoring at the proficient or advanced levels 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 the school 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

![Figure 1](http://example.com-figure1.png)

**FIGURE 1.** Plot of Changes in Percentages of Students in A School at the Proficient or Advanced Levels From 1997 to 1998 as a Function of School Size in 1997.
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 gains for all five clusters of schools based on school size is close to zero.

The spread of positive and negative difference scores tend 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. This is so because large schools are less affected than small schools by differences in the student body from year to year; that is, the large schools have less sampling variability than the small schools have due to factors other than real changes in the quality of education provided by a school. 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, however, are quite volatile. 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” will generally show increases in scores the year after they are identified simply because of the noise in the estimates of improvement and 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 (AERA, APA, & NCME, 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) or those described 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. 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 accountability purposes. 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. The No Child Left Behind Act of 2001 that was recently signed into law by President Bush calls on states to develop targets for “Adequate Yearly Progress” (AYP) and to identify schools as needing improvement that fail to meet their AYP targets two years in a row. Requiring more than the gains in a single year for such identification is better than doing so after the first round of gains are reported, because many of the schools that fail to meet the
target in the first cycle will meet it in the second. The longer time frame allows for the accumulation of more data that will yield more trustworthy results.

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.

The accumulation of more data or by increasing precision through more sophisticated analytical techniques will enhance the trustworthiness of accountability systems using the successive cohort approach. As Carlson (2000) has shown, however, the successive cohort approach will still provide a poor approximation to measures of change based upon longitudinal data. Although longitudinal tracking of students presents logistical problems and conceptual ones, such as the appropriate accounting for and attribution of results for students who change schools, longitudinal analyses are the most direct and valid way to account for changes in student achievement.

Note

1 The fourth grade-reading test was rescaled in 2001 when a vertical scale was established for all CSAP reading tests from grades 3 through 10.

References


http://www.cde.state.co.us/cdeasses/pubassess.htm


Authors

ROBERT L. LINN is Distinguished Professor, University of Colorado and Co-Director of the Center for Research on Evaluation, Standards, and Student Testing, Campus Box 249, Boulder, CO 80309-0249. He specializes in educational measurement.

CAROLYN HAUG is Director of Student Assessment at the Colorado Department of Education, 201 E. Colfax Avenue, Room 104, Denver, CO 80203. She specializes in assessment and educational policy.

Manuscript received July 7, 2001
Revision received January 24, 2002
Accepted February 12, 2002