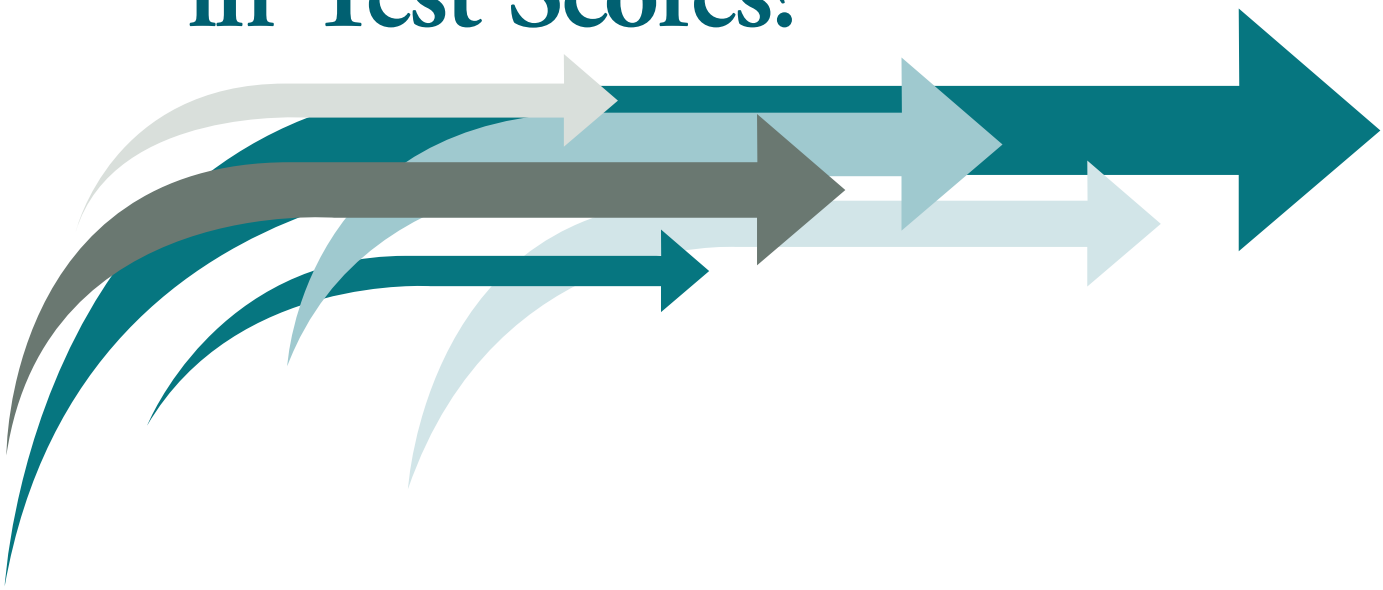
CENTER ON EDUCATION POLICY CEP

State Test Score Trends Through 2007-08, Part 2

Is There a Plateau Effect in Test Scores?



Center on Education Policy

July 2009

Table of Contents

List of Figures and Tables

Is There a Plateau Effect in Test Scores?

Main Findings

"Maryland Test Scores Rise But Near a Plateau" was the headline of a 2006 *Washington Post* report on student performance on that state's assessments (Anderson, 2006). In the third year of the Maryland School Assessment program, students made gains, but not as large as those in the previous year. In the article, a state education official speculated that perhaps initial test score gains were "low hanging fruit," meaning that in the early years of Maryland's testing program, students and teachers became more familiar with the state exam's format and scores skyrocketed, but after those first few years scores leveled off. This phenomenon is often known as a "plateau effect."

A slowing down of increases in test scores has been noted in other states as well (Center for Mental Health in Schools, n.d.), and the plateau effect is often cited as a reason. A report by Massachusetts education leaders on trends in schools making adequate yearly progress under the No Child Left Behind Act (NCLB) stated, "At some point, test scores plateau . . . Straight line projections may look nice, but they do not occur in real life" (MassPartners for Public Schools, 2005, p. 1). In a commentary on California test results, Fuller (2004) wrote that "a majority of California's schools have hit a plateau or worse."

If a plateau effect truly exists, it would certainly cast doubt on the ability of schools to meet ever-increasing performance targets, such as the targets for percentages of students scoring proficient under NCLB. The implication is that under test-based accountability systems, initial gains in test results observed during the first few years are inflated and difficult to sustain, and education leaders should therefore expect to see student test results level off over time.

But does performance data from the state tests used for NCLB accountability support the notion that a plateau effect is widespread—and perhaps inevitable?

The Center on Education Policy (CEP), which for three years has been studying student achievement trends, examined the extent of the plateau effect using our database of test results in reading and mathematics from all 50 states. In particular, we analyzed 55 trend lines from 16 states showing the percentages of students scoring at the proficient level on state tests between 1999 and 2008. Not all states had data covering the entire ten-year period. To ensure that trend lines were long enough to allow time for a plateau to emerge, we only included states with at least six years of data.

Our analysis revealed several main findings:

- **In the current testing context, one cannot assume the existence of a plateau effect when trying to predict state test score trends.** Although this study found instances of plateaus in test score trends in the 16 states analyzed, they were not as pervasive as may be commonly assumed. Percentage proficient trends followed a wide variety of trajectories, including some plateau patterns. Of the 55 trend lines we examined from various states and different grade levels in reading and math, 15 exhibited a plateau pattern. We also found 21 trend lines with steady increases in the percentage proficient over time and 19 more with fluctuating "zigzag" patterns that still moved in an overall upward direction.

- **The largest gains did not consistently show up in the early years of a testing program.** In many of the trend lines, the largest gains occurred between the first and second years of administering a new test. But just as often, the largest gains appeared between the third and fourth years, or between the fifth and sixth. Thus, we concluded that the largest gains are just as likely—and sometimes more likely—to occur after four or even six years of a testing program as they are in the first few years.

- **A clear upswing in test results was apparent after the enactment of the No Child Left Behind Act (NCLB).** In many states, the largest gains in percentages proficient occurred between 2003 and 2004, two years after NCLB took effect. But the early years of NCLB were not always concurrent with the first few years of a state testing program. In several states, the tests used for NCLB had already been in place for some years, and a bump in scores still appeared after NCLB. This pattern suggests that test results can increase substantially even after a test has been in place for several years if higher stakes are introduced in the accountability system.

- **In the three states with the longest trend lines, gains generally did level off after nine or ten years, but the data were too limited to know whether this is a consistent pattern in state test performance.** One complicating factor in studying test score trends is that states tend to change their tests quite frequently, so long trend lines are rare. It may well be that by the time a state would start to show a plateau effect, it changes its tests.

Purpose of the Study

In the fall of 2006, the Center on Education Policy began tracking state test score trends going back as far as 1999 in some states. This work on achievement trends is an extension of CEP's broader ongoing study of the implementation and effects of NCLB.

This report on the plateau effect is the second in a series, entitled *State Test Score Trends Through 2008,* that describes findings from year three of our analysis of achievement trends. Part 1 of the series (*Is the Emphasis on "Proficiency" Shortchanging Higher- and Lower-Achieving Students?*) found that student achievement, as measured by state tests, has generally improved since 2002, not only at the proficient level but also at the basic and advanced levels (CEP, 2009). Other reports in this series will examine trends through 2007-08 in overall performance for racial-ethnic subgroups and low-income students, as well as achievement gaps for these subgroups; will discuss achievement trends for students with disabilities, English language learners, and male and female students; and will explore the policy implications of our findings about achievement from the other parts of the study.

The plateau effect is relevant to the issue of whether the gains described in part 1 of our achievement study are sustainable or whether they are largely a result of score inflation—misleadingly high scores, produced by teachers adjusting instruction to the content of a particular test, that do not necessarily translate into higher scores on other tests or broader improvements in student learning. This analysis of the plateau effect seeks to determine whether we should expect the gains we found to dissipate eventually, and if so, when.

As with any discussion of test scores, one should keep in mind that state tests are not perfect measures of student achievement. A recurring criticism of tests used for high-stakes accountability is that they can lead to score inflation, especially in the early years of a testing program. As educators become more familiar with a particular test, they may narrowly

focus instruction on specific content that is likely to appear on that test and give students practice questions with the same format as test questions. The idea underlying the plateau effect is that once "easy" methods for increasing test scores are exhausted, it is difficult to show more gains, so scores level off. Another possible explanation for a plateau is that once a large percentage of students have reached the state's proficient level, it may be difficult to bump up the remaining students, who often have the greatest learning challenges.

There is evidence that test scores do increase quite a bit in the first few years after implementation of a new test; this has been shown in the cases of Kentucky (Koretz et al., 1991), Texas (Klein et al., 2000), Chicago (Jacob, 2005), Arkansas (Fuller et al., 2007), and New Jersey (Fuller et al., 2007). However, very few studies have explored whether a plateau (leveling off) takes place *after* the first few years of increasing scores.

The main evidence supporting the plateau effect comes from Florida's test results from 1977 to 1997. There was a clear jump in the percentage passing for white, Latino, and African American students between 1977 and 1980, and another between 1983 and 1984. After that, a long plateau ensued. The overall percentage passing remained stagnant until 1997, and scores for African American students declined slightly from 1984 to 1997 (Linn, 1998). It should be noted, however, that when the plateau effect was first observed and discussed, it was based on data from the 1980s and 1990s, when it was more common for states to use the exact same test form with the same questions for many years in a row. This was the case with the Florida test examined by Linn (1998). Since then, state testing programs have changed test forms much more frequently—far fewer test items are "recycled" from year to year.

In any case, the plateau effect seems to be commonly accepted among education researchers, policymakers, and the media, but the evidence for it is based on a small number of states and on test data from the 1980s and 1990s. No study has yet used a large number of cases to see if the plateau effect is a common outcome of current test-based accountability systems. This study investigates the extent of the plateau effect in the current accountability context, using state test data from 16 states during the period between 1999 and 2008.

Study Methods

As part of an ongoing study of NCLB implementation, the Center on Education Policy maintains a database of state test results from all 50 states, going back as far as 1999. These data have been collected with the indispensable assistance of our contractor, the Human Resources Research Organization (HumRRO). For each state, the database includes data on the percentage of students reaching three achievement levels (basic, proficient, and advanced), at three grade levels (elementary, middle, and high school), for two subjects (reading and math). All of the state test data are posted on CEP's Web site at www.cep-dc.org.

In this study, we focused on the percentage of students reaching the proficient level and above because that is the primary indicator of student achievement reported for NCLB. For purposes of this study, a "trend line" refers to the movement in percentages proficient for one grade level and one subject in a single state. For example, the pattern of changes between 1999 and 2008 in percentages proficient in reading for Louisiana 4th graders represents one trend line.

We examined trend lines consisting of six to ten years of test data in order to have enough years of data to detect plateaus. In fact, even longer trend lines may be needed, as discussed later. Some researchers have suggested that a plateau usually occurs between the third and the seventh year of a test-based accountability program (Goldschmitt, Boscardin, & Linn, 2006).

For several reasons, this study does not use data from all 50 states. Many states have introduced new tests in recent years, and their trend lines are too short to see whether longer-term patterns may appear. In other cases, states made revisions in their testing programs, such as changing the cut score for proficiency on an existing test or adopting a new scoring scale. When states took these actions, they created a break in test data, which made it inappropriate and inaccurate to compare test results after the change with results from before the change. To identify when a trend line had been broken, we gathered information from states about changes in their testing programs and limited our analyses of trends to only those states with at least six consecutive years of comparable test data. We also excluded trend lines that began before 1999 because we only collected data back that far. In states that initiated tests during the pre-1999 period, we lacked the data needed to determine whether test scores jumped during the first few years after the test was introduced—a key characteristic of the plateau effect.

In the end, we were able to analyze 55 trend lines from 16 states, a much larger data set than used in any previous studies of plateau effects. These states included Arkansas, Colorado, Florida, Georgia, Indiana, Louisiana, Kentucky, Maryland, Massachusetts, Mississippi, New Hampshire, New Jersey, Oklahoma, Pennsylvania, South Carolina, and Washington.

A panel of five nationally known experts in educational testing or education policy provided advice on aspects of the study design, reviewed data, and commented on drafts of this report. The panel consisted of the following members:

- Laura Hamilton, senior behavioral scientist, RAND Corporation
- Eric Hanushek, senior fellow, Hoover Institution
- Frederick Hess, director of education policy studies, American Enterprise Institute
- Robert L. Linn, professor emeritus, University of Colorado
- W. James Popham, professor emeritus, University of California, Los Angeles

Although the panel members, as well as HumRRO staff, provided input on this report, we did not ask them to endorse it, so the findings and views expressed here are those of CEP.

Frequency of the Plateau Effect

We examined each state trend line to look for plateau-like patterns—large gains in the percentage of students reaching the proficient level on state tests in the first few years, followed by small gains, no gains, or declines in subsequent years. In particular, we looked for 1) jumps in the percentage proficient during the first few years of the trend line; 2) trend lines where the percentage proficient showed larger gains during the first half than in the second half; and 3) a clear leveling or decrease in the percentage proficient for at least two years at the end of the trend line. If trend lines met these three criteria, we labeled them as plateaus.

Of the 55 trend lines examined, we found 15 with a fairly clear plateau effect. Some examples are illustrated in **figure 1**. Gains were made in the percentage proficient during the first half of the trend line and then tapered off. For example, in the case of Arkansas high school reading test results, increases in percentages proficient were larger during the first half of the trend line than in the second half. From 2001 to 2005, the percentage proficient jumped from 22% to 45%, but from 2005 to 2007, it rose only to 51% and remained there in 2008. The Massachusetts trend line exhibited a similar pattern. Washington saw a drop in the percentage proficient during the last three years of its trend line in high school math; the percentage proficient was 44% in 2008, the same as in 2004.

Figure 1. Plateau Effects

Other Patterns

In all but one of the trend lines we examined, the percentage proficient tended to move upward overall; that is, the percentage proficient in the last year of the trend was always greater than in the first year. Aside from the plateau patterns discussed above, the rest of the trend lines fell into two broad categories. The first category, which we called "steady increases," comprised trend lines where gains in the second half of the period analyzed were comparable to or exceeded gains in the first half, and where no more than one decline occurred in the percentage proficient over the entire course of six to ten years. Out of the 55 trend lines, we found 21 that exhibited this pattern. Some examples are shown in **figure 2.**

Note that the trend line in figure 2 for Louisiana elementary reading was flat between 2007 and 2008. We still characterized this as a steady increase rather than a plateau because growth in the percentage proficient accelerated in the second half of the trend line; it increased five percentage points between 1999 and 2004, but jumped another nine percentage points by 2008. Because test scores may naturally fluctuate a bit from year to year for reasons unrelated to student learning, a flat trend over a single year (from 2007 to 2008) is too short of a period to tell whether gains have actually reached a plateau.

Figure 2. Steady Increases

The last category consisted of "zigzag" patterns, where the percentage proficient goes up and down multiple times—in other words, where there are two or more year-to-year declines mixed in among the increases. Some of these trends started with initial decreases rather than increases. We uncovered 19 instances of zigzag patterns. Some examples are shown in **figure 3.**

Figure 3. Zigzag Patterns

Initial Increases in Percentages Proficient

According to the plateau scenario, the largest gains come in the first few years after the introduction of a test-based accountability program. Once the "easy" gains are made, trend lines level off. To see if this was indeed happening, we examined each trend line to see when the largest percentage point jumps occurred.

The results are presented in **table 1.** Of the 55 trend lines examined, 12 showed the largest gains between year 1 and year 2 of a new testing program. In 13 other instances, the largest gains occurred between years 3 and 4 of a testing program, and in 14 cases the largest gains appeared between years 5 and 6. Thus, we found that the largest gains were at least as likely to occur after four or even six years of a testing program as they were in the first few years. We identified fewer instances in which the largest gains occurred in years 9 and 10, but this is complicated by the far smaller data set—only seven trend lines from three states included data for nine or ten years. But the limited evidence suggests that plateaus may become more apparent as trend lines get longer.

Table 1. Largest Increases in Percentage Proficient by Year in Testing Trend Line

Years of Annual Change	Number of Trend Lines With Greatest Gain Between Those Years*	Total Number of Trend Lines with Data for That Period
Year 1 – 2	12	55
Year 2 – 3	9	55
Year 3 – 4	13	55
Year 4 – 5	9	55
Year 5 – 6	14	55
Year 6 – 7	6	46
Year 7 – 8	4	34
Year 8 – 9	1	7
Year 9 – 10	0	7

* The sum of the numbers in this column is greater than 55 because, within a few trend lines, two years had large jumps of identical size. For example, in Kentucky middle school reading, the largest percentage point jumps of three percentage points occurred twice in the trend line, between years 2 and 3 of the testing program and again between years 5 and 6.

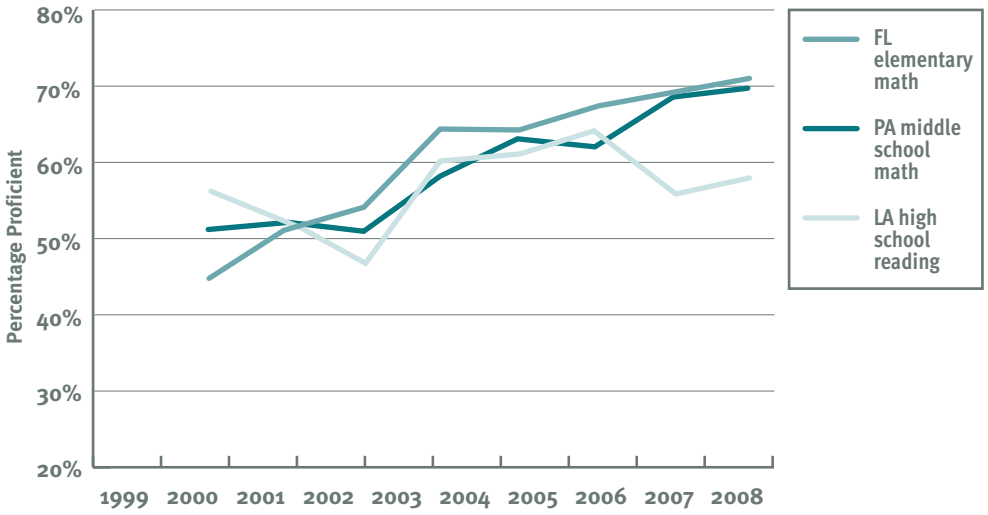
Possible NCLB Effect

To see whether any patterns emerged by calendar year, rather than year of the testing program, we also organized the trend lines according to calendar year. We found that for 20 of the 55 trend lines—more than a third of the trend lines analyzed—the largest jump in percentages proficient occurred between 2003 and 2004. The first full school year of testing under NCLB was 2002-03, and the second was 2003-04. The implications of making adequate yearly progress as defined by NCLB were fully evident by 2003-04. The data suggest that once higher stakes were attached to existing state accountability systems, great gains were often made, even

with tests that had already been in place for four or five years. This is consistent with NCLB having an effect, although it is somewhat difficult to establish clear causation because of other policies being implemented at the state and local levels at the same time.

A few examples of trend lines showing possible NCLB effects are Florida elementary math, Pennsylvania middle school math, and Louisiana high school reading, displayed in **figure 4.** In these cases, as well as 17 more, the largest jump in the percentage proficient occurred between 2003 and 2004.

Figure 4. NCLB Effect



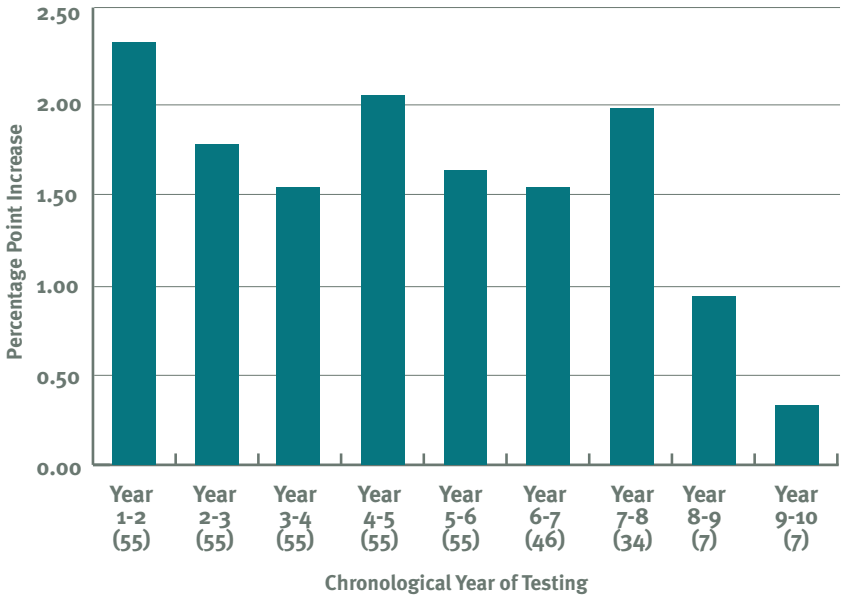
Trends Based on Averages

The analyses described above looked at individual trend lines. To see whether our focus on individual trend lines had caused us to miss any overall tendency, we also analyzed the plateau effect using a different approach, one that averaged the increases in percentages proficient across all of the state trend lines.

The percentage proficient data were arranged according to the number of years a test had been in place, shown in **figure 5**. The year spans in the horizontal axis, such as year 1-2, represent the period for which we calculated an average change across all trend lines in all states with data. For example, the first bar shows the average percentage point gain in the percentage proficient between the first year of testing and the second. The numbers in parentheses in the horizontal axis indicate the number of trend lines that were averaged to produce the value in each bar.

Large gains, on average, occurred soon after a new test was introduced. The average difference in the percentage proficient between year 1 and year 2 of a test was 2.35 percentage points. However, states continued to post large gains even after tests had been in place for several years. Average gains of about 2 percentage points also occurred between years 4 and 5, and between years 7 and 8. According to the criteria developed with advice from our expert panel, an annual

Figure 5. Average Annual Gains in the Percentage Proficient by Number of Years Test Has Been in Place

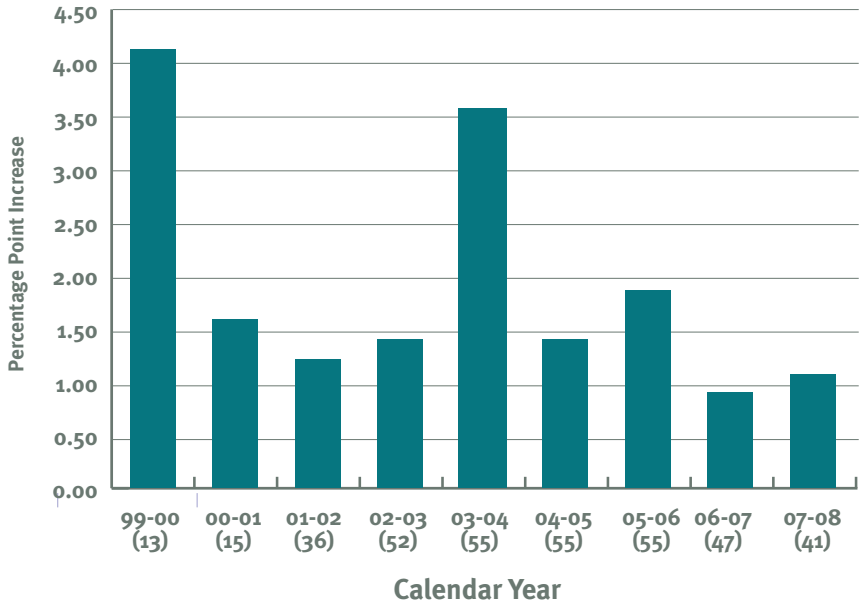


increase in the percentage proficient of greater than 1.0 percentage point constitutes a "moderate-to-large" gain, so a 2 percentage point increase is substantial. The drop-off evident in the last two bars for year 8-9 and year 9-10 may reflect a general leveling off, but as mentioned above, only seven trend lines from three states included nine or ten years of data.

When we averaged the gains by calendar year (**figure 6**), we found that the largest average jump in percentages proficient took place between 1999 and 2000—an average gain of 4.13 percentage points. However, this was based on only 13 trend lines from four states (Louisiana, Washington, New Jersey, and Kentucky), and some of these trend lines demonstrated unusually large jumps ranging from 7 to 11 percentage points in the early years.

The average gain between 2003 and 2004 is the next largest—3.63 percentage points. Again, we suspect this is largely attributable to NCLB. The average for this year span includes data from all 55 trend lines. At first, we thought the 2003-04 jump might be partly explained by the gains typically made in the first year of testing, if a large number of states had adopted new tests right after NCLB was enacted. But in fact, only two of the states we analyzed introduced new tests in 2003, so this spike cannot be attributed simply to gains following the introduction of a new test. The percentage proficient gains tapered off somewhat after 2006, but states still showed average increases of about one percentage point per year during those later years, rather than an actual plateau.

Figure 6. Average Annual Gains in the Percentage Proficient by Calendar Years



Conclusion

Should state officials and other education leaders who are concerned with test results expect to see test scores level off over time? In the current accountability context, and based on the limited data available, the plateau effect should not be assumed. Only 15 of the 55 trend lines in our data set exhibited a plateau pattern. Instead, percentage proficient trend lines followed a wide variety of patterns, and the only predictable pattern was an overall upward trajectory. We found 21 instances of steady increases in test scores over time, and 19 instances where the percentage proficient zigzagged up and down unpredictably. All but one trend line had an overall increase.

We did find that large gains often emerged between the first and second years of testing, as expected in a plateau scenario, but they also often occurred between the third and fourth years, or between the fifth and sixth. In addition, we identified big jumps when higher-stakes policies were added to existing state accountability systems, as was the case between 2003 and 2004, the first two full years of testing after NCLB was enacted. But sizable gains were often made in other years, as well.

The patterns we found in this set of data have two implications. First, it is possible (and common) to see large gains using a test that has been in place for a long period. Second, raising the stakes attached to test results can lead to a substantial increase in performance, even on tests that have already been in place for four or five years.

When data were averaged across states, some limited evidence showed gains tapering off in the ninth or tenth year after the introduction of a test. However, even in these later years, the percentage proficient continued to increase slightly. This finding was based on limited data because few states had the same test in place for more than eight years. If growth really

slows, as the data for these few states suggest, we may see more evidence of a plateau effect in the future. Linn's original description of the plateau effect was based on a 20-year trend line in one state.

The existence of a plateau effect probably also depends largely on the nature of the test being used. For some tests, it may be easy for teachers to predict what types of questions are likely to appear because the test forms may contain very similar (though not exactly the same) items from year to year. Some tested skills may be more easily and quickly taught, such as math computation skills tested with multiple-choice items. Other tested skills may require more cumulative, long-term instruction, such as complex problem-solving where students must show their work or a reading comprehension question based on a literary passage. A simple test that focuses on lower-order skills may be more susceptible to the plateau effect, as it may be easier to adjust instruction relatively quickly to prepare students for the test.

The less frequent reuse of the same test items may help explain why the plateau effect is less common now than it might have been in the past. Most states are using new test forms each year that include a mostly new set of test items; the test forms are carefully developed to be equated with earlier forms in terms of content and difficulty. In the 1980s and 1990s, when the plateau effect was first discussed, it was more common for states to use the same test form with the exact same questions for several years in a row. Changing test forms may inhibit, but not do away with, the kinds of narrow teaching to the test or outright cheating that would cause large initial gains. It may also be more difficult to identify plateaus simply because states are also periodically changing their entire testing programs, which causes breaks in trend lines. As CEP has found in maintaining its database of state test scores, only a very limited number of states have administered the same test without changes for more than a decade. State and federal testing policies have been in transition, largely due to NCLB. States also alter tests when they revise their content standards, make their tests more or less difficult, or move to different testing formats. It may well be that by the time a state would start to show a plateau effect, it changes its tests.

References

Anderson, N. (2006, June 21). Maryland test scores rise but near plateau. *Washington Post*. Retrieved October 2, 2008, from www.washingtonpost.com/wp-dyn/content/article/2006/06/20/AR2006062000558.html

Center on Education Policy. (2009). *State test score trends through 2007-08, part 1: Is the emphasis on "proficiency" shortchanging higher- and lower-achieving students?* Washington, DC: Author.

Center for Mental Health in Schools. (n.d.). Data on the plateau or leveling off effect of achievement test scores. Los Angeles, CA: University of California at Los Angeles. Retrieved March 23, 2009, from http://smhp.psych.ucla.edu/pdfdocs/plateau.pdf

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher (36)*5, 268-278. Retrieved on May 1, 2009 from http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3605/07EDR07_268-278.pdf.

Fuller, B. (2004, August 22). Accountability rises, scores fall. *Los Angeles Times*. Retrieved April 29, 2009, from http://articles.latimes.com/2004/aug/22/opinion/oe-fuller22

Goldschmidt, P., Boscardin, C. K., & Linn, R. (2006, September 30). *CRESST analysis of California Accountability Progress Report*. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Jacob, B. A. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of Public Economics, 89*(5-6): 761-796.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000, Oct. 24). *What do test scores in Texas tell us?* Santa Monica, CA: RAND Corp.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991, April). *Effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association and the National Council of Measurement in Education, Chicago.

Linn, R. L. (1998). *Assessments and accountability.* CSE Technical Report 490. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

MassPartners for Public Schools. (2005). *Facing reality: What happens when good schools are labeled "failures?" Projecting adequate yearly progress in Massachusetts schools.* Retrieved May 1, 2009, from www.massteacher.org/news/headlines/aypreport.pdf

Credits and Acknowledgments

This report was written and researched by Naomi Chudowsky and Victor Chudowsky, CEP consultants. Nancy Kober, a CEP consultant, edited the report. A team from the Human Resources Research Organization (HumRRO) under the direction of Sunny Becker collected the test data analyzed in the report with support from CEP. Diane Stark Rentner, CEP's director of national programs, oversaw the study project for CEP, communicated with the expert panel, and provided advice and assistance for all aspects of the study. Jack Jennings, CEP's president and CEO, provided advice and assistance.

We would like to thank our panel of expert advisors—Laura Hamilton, Eric Hanushek, Frederick Hess, Robert Linn, and W. James Popham—for their invaluable advice. Additionally, we are grateful to the chief state school officers and state assessment personnel for their cooperation in providing information on state testing programs and student achievement data.

Based in Washington, D.C., and founded in January 1995 by Jack Jennings, the Center on Education Policy is a national independent advocate for public education and for more effective public schools. The Center works to help Americans better understand the role of public education in a democracy and the need to improve the academic quality of public schools. We do not represent any special interests. Instead, we help citizens make sense of the conflicting opinions and perceptions about public education and create the conditions that will lead to better public schools.

The Center on Education Policy receives nearly all of its funding from charitable foundations. We are grateful to the William and Flora Hewlett Foundation and the Smith Richardson Foundation for their support of the student achievement study. The George Gund Foundation and the Phi Delta Kappa International Foundation also provide the Center with general support funding that assisted us in this endeavor. The statements made and views expressed are solely the responsibility of the Center.

Center on Education Policy
1001 Connecticut Avenue, NW, Suite 522
Washington, D.C. 20036
tel: 202.822.8065
fax: 202.822.6008
e: cep-dc@cep-dc.org
w: www.cep-dc.org