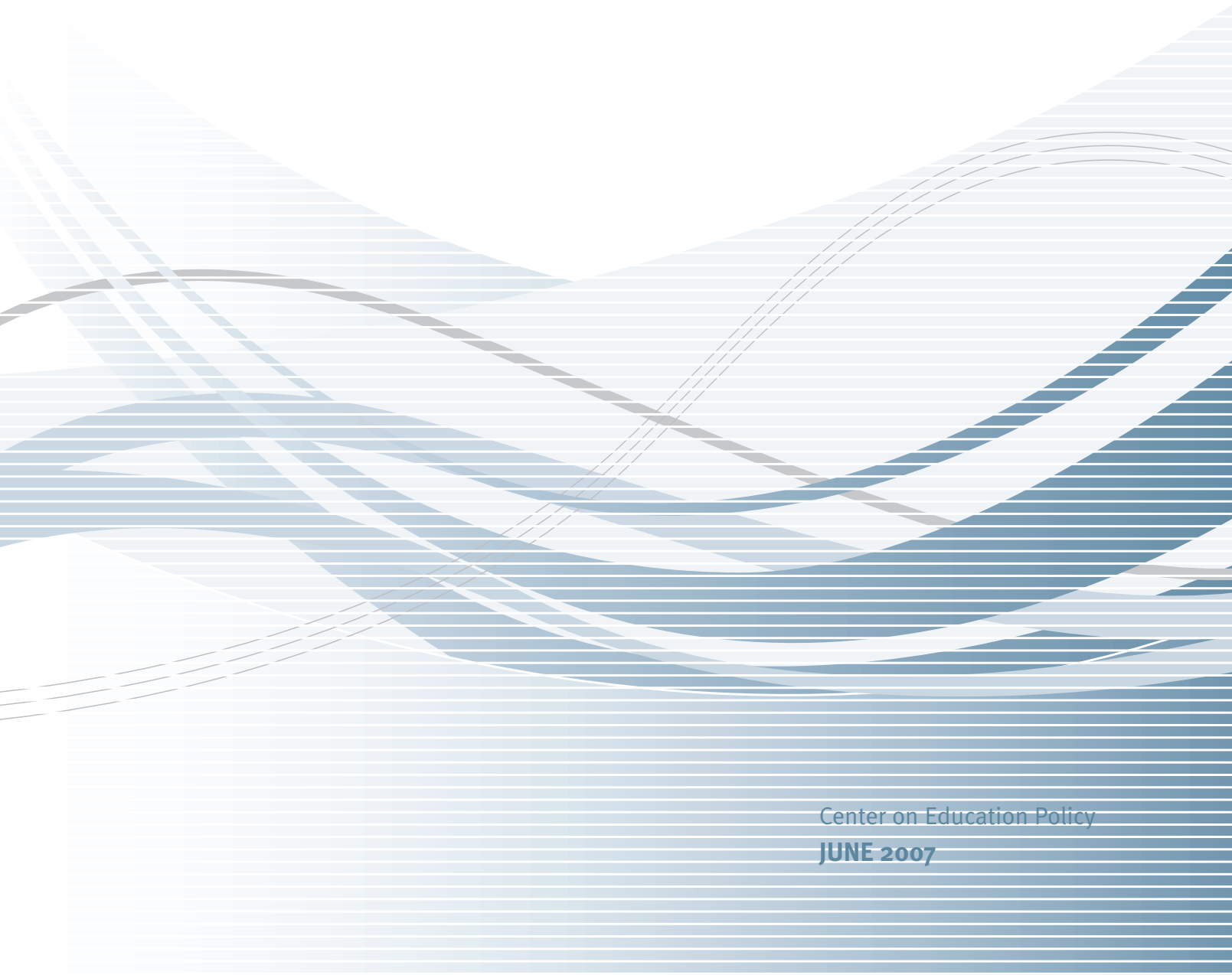
CENTER ON EDUCATION POLICY CEP

ANSWERING THE QUESTION THAT MATTERS MOST

# HAS STUDENT ACHIEVEMENT INCREASED SINCE NO CHILD LEFT BEHIND?



Center on Education Policy

JUNE 2007

CENTER ON EDUCATION POLICY CEP

ANSWERING THE QUESTION THAT MATTERS MOST

# HAS STUDENT ACHIEVEMENT INCREASED SINCE NO CHILD LEFT BEHIND?

Center on Education Policy

**JUNE 2007**

# Table of Contents

# Chapter 1
# Summary of Key Findings



Since 2002, the No Child Left Behind Act (NCLB) has spurred far-reaching changes in elementary and secondary education, all aimed at accomplishing the same fundamental goal—to improve students' academic achievement. As the Congress prepares to reauthorize the Act, two related questions matter most:

1. Has student achievement in reading and math increased since NCLB was enacted?

2. Have achievement gaps between different subgroups of students narrowed since NCLB was enacted?

To answer these questions, the Center on Education Policy (CEP), an independent nonprofit organization, conducted the most comprehensive study of trends in state test scores since NCLB took effect. We carried out this study with advice from a panel of five nationally known experts in educational testing or policy research, and with extensive technical support from the Human Resources Research Organization (HumRRO). Although we collected data from all 50 states, not every state had enough consistent data to do a complete analysis of test score trends in reading and math before and after 2002. Based on the data that states did provide, we reached five main conclusions.

## Main Conclusions

1. In most states with three or more years of comparable test data, student achievement in reading and math has gone up since 2002, the year NCLB was enacted.

2. There is more evidence of achievement gaps between groups of students narrowing since 2002 than of gaps widening. Still, the magnitude of the gaps is often substantial.

3. In 9 of the 13 states with sufficient data to determine pre- and post-NCLB trends, average yearly gains in test scores were greater after NCLB took effect than before.

4. It is very difficult, if not impossible, to determine the extent to which these trends in test results have occurred *because* of NCLB. Since 2002, states, school districts, and schools have simultaneously implemented many different but interconnected policies to raise achievement.

5. Although NCLB emphasizes public reporting of state test data, the data necessary to reach definitive conclusions about achievement were sometimes hard to find or unavailable, or had holes or discrepancies. More attention should be given to issues of the quality and transparency of state test data.

The study that produced these conclusions had several unique features, designed to address the limitations of past research on achievement since 2002. We went to great lengths to gather the most current results on state reading and mathematics tests from all 50 states and to have all states verify the accuracy of their data. Within each state, we limited our analyses to test results that were truly comparable from year to year—in other words, that had not been affected by such factors as the adoption of new tests or changes in the test score students must reach to be considered proficient. We also compared trends before and after 2002 to see whether the pace of improvement has sped up or slowed down since NCLB took effect. We supplemented our analyses of the percentage of students scoring at or above the proficient level—the "magic number" for NCLB accountability—with analyses of effect size, a statistical tool based on average (mean) test scores that addresses some of the problems with the percentage proficient measure. And we analyzed all of the data—which in a typical state included as many as 16,000 individual numbers—as objectively as possible, using a consistent set of rules that were developed without regard to whether they would lead to positive or negative findings.

The rest of this chapter summarizes the findings that led us to the five main conclusions. Additional key findings can be found at the beginning of the other chapters.

## Gains in Reading and Math Since 2002



To reach national conclusions about reading and math achievement, we first determined the test score trends in each state, looking at both the percentages of students scoring proficient and effect sizes where available. The state trends were then aggregated into a national picture of achievement that included these and other findings (chapter 4):

- The number of states showing gains in test scores since 2002 is far greater than the number showing declines. For example, of the 24 states with percentage proficient and effect size data for middle school reading, 11 demonstrated moderate-to-large gains (average gains of at least 1 percentage point per year) in middle school reading, and only one showed a moderate or larger decline.

- Five of the 22 states with both percentage proficient and effect size data at the elementary, middle, and high school levels made moderate-to-large gains in reading and math on both measures across all three grade spans. In other words, these five states showed gains according to all of the indicators collected for this study. In reading alone, seven states showed moderate-to-large increases across all three grade spans on both measures. In math alone, nine states showed similar gains across all three grade spans on both measures. The rest of the states had different trends at different grade spans.

- Elementary school math is the area in which the most states showed improvements. Of the 25 states with sufficient data, 22 demonstrated moderate-to-large math gains at the elementary level on both the percentage proficient and effect size measures, while none showed moderate or larger declines. Based on percentages proficient alone, 37 of the 41 states with trend data in elementary math demonstrated moderate-to-large gains, while none showed moderate or larger declines.

- More states showed declines in reading and math achievement at the high school level than at the elementary or middle school levels. Still, the number of states with test score gains in high school exceeded the number with declines.

- Analyses of changes in achievement using effect sizes generally produced the same findings as analyses using percentages proficient. But in some cases, the effect size analysis showed a different trend. In Nevada, for instance, the percentage proficient in high school math decreased, while the average test score increased. In New Jersey the percentage proficient in middle school reading rose slightly, while the average test score dropped.

- When the percentage of students scoring at the proficient level on state tests is compared with the percentage scoring at the basic level on the National Assessment of Educational Progress (NAEP), states show more positive results on their own tests than on NAEP. Moreover, the states with the greatest gains on their own tests were usually not the same states that had the greatest gains on NAEP. The NAEP tests, however, are not aligned with a state's curriculum as state tests are, so NAEP should not be treated as a "gold standard" to invalidate state test results but as an additional source of information about achievement.

## Narrowing Achievement Gaps

We analyzed trends in test score gaps for major racial-ethnic subgroups of students, low-income students, students with disabilities, and limited-English-proficient (LEP) students. We looked at both percentages proficient and effect size data where available; effect size data were harder to come by for subgroups than for students overall. We considered a narrowing or widening of the achievement gap to be a trend for a specific subgroup if it occurred in the same subject (reading or math) across all three grade spans (elementary, middle, and high school). We compiled trends from the 50 states to arrive at these and other national findings (chapter 5):

- Among the states with sufficient data to discern trends by subgroup, the number of states in which gaps in percentages proficient have narrowed since 2002 far exceeds the number of states in which gaps widened.

- For the African-American subgroup, 14 of the 38 states with the necessary data showed evidence that gaps have narrowed in reading across all three grade spans analyzed, while no state had evidence that gaps have widened. In mathematics, 12 states showed these gaps narrowing, while only one state showed the gaps widening. Results were similar for the Hispanic and low-income subgroups.

- As with the percentage proficient, the states in which effect size gaps have narrowed outnumbered the states in which effect size gaps have widened. However, for states with both types of data, there were a number of instances where gap closings in terms of percentages proficient were not confirmed by effect size. Effect sizes seem to give a less rosy picture of achievement gap trends.

- Even for subgroups that showed evidence of gaps narrowing, the gaps in percentages proficient often amounted to 20 percentage points or more, suggesting that it will take a concerted, long-term effort to close them.

## Gains Before and After NCLB

Many states had reforms well underway before NCLB, so it is useful to know whether the pace of improvement has picked up since NCLB took effect (chapter 4). Only 13 states supplied enough years of data to make this determination—too few to know whether the findings for this sample represent a true national trend. In nine of these states, test results improved at a greater average yearly rate after 2002 than before. In the other four states, the pre-NCLB rate of gain outstripped the post-NCLB rate.

## Difficulty of Attributing Causes for Gains

This report focuses on whether test scores have gone up *since the enactment* of NCLB. We cannot say to what extent test scores have gone up *because* of NCLB (chapter 2). It is always difficult to tease out a cause-and-effect relationship between test score trends and any specific education policy or program. With all of the federal, state, and local reforms that have been implemented simultaneously since 2002, it becomes nearly impossible to sort out which policy or combination of policies is responsible for test score gains, and to what degree. In a similar vein, this report does not take a position on how well specific components of NCLB are working or whether the requirements in the current law are the most effective means to raise achievement and close test score gaps.

One more caveat should be emphasized: test scores are not the same thing as achievement. Although tests are often viewed as precise and objective, they are imperfect and incomplete measures of how much students have learned. Still, state tests are the primary measure of achievement used in NCLB and are the best available standardized measures of the curriculum taught in classrooms.

## Need for More Transparency in Test Data

The No Child Left Behind Act requires states to report a massive amount of test data and attaches serious consequences to these data for districts, schools, and educators. But the data on which so much rests are not easy to access in some states and are sometimes inconsistent, outdated, or incomplete (chapter 7). Moreover, the data needed to calculate effect sizes or determine which subgroups were small or rapidly changing were unavailable in some states, even though these data are integral to all testing systems. Reasons for these shortcomings include overburdened state departments of education, ongoing corrections in test data, and technical or contractual issues with test contractors. These shortcomings are not necessarily the fault of state officials—who were generally cooperative in providing or verifying data when asked—but these problems complicated our efforts to reach definitive conclusions about student achievement.

It took many months of effort to gather all the data needed for this study and have state officials verify their accuracy. Our experience suggests how difficult it would be for the average citizen to get information about test score trends in some states, and points to the need for greater transparency in state test data. States could improve transparency by taking the following steps:

- Posting test data in an easy-to-find place on state Web sites
- Providing clear information and cautions about breaks in the comparability of test data caused by new tests or changes in testing systems
- Reporting standard deviations, mean scale scores, numbers of test-takers, and other important information listed in chapter 7

## State-By-State Achievement Trends on the Web

The trends highlighted in this report have been drawn from an extensive set of data on each state. Complete profiles of test results and other information for individual states can be accessed on the CEP Web site at www.cep-dc.org/pubs/stateassessment. We encourage anyone who is interested in trends for a specific state to visit the Web site and find that state's profile.

## Future Phases of This Study

This report describes the findings from phase I of what will be a three-phase study of student achievement. Phase II, which will be completed this summer, involves on-site interviews with state officials in 22 states. Phase II investigates in more detail the trends uncovered during phase I of the study and the factors that affect comparability or availability of test data; it also reports information from state officials about how well specific requirements of NCLB are working and how the law could be improved. Phase III, which will be carried out in the fall and winter of 2007-08, examines student achievement at the school district level in three states.

# Chapter 2
# What Makes This Study Unique



## Key Findings

- This study was designed to be the most comprehensive and thorough study to date on trends in state test scores since NCLB was enacted. Instead of just looking at test results in a limited number of states, the study analyzed results from all 50 states. And instead of taking for granted that the data reported on state Web sites were accurate, the study asked states to verify the accuracy of the test data collected. The process of gathering, verifying, and analyzing test results from all states turned out to be an arduous task that involved a great deal of cross-checking and depended on cooperation from state officials.

- This study included other unique elements intended to address limitations of past research on achievement. To determine whether the rate of improvement has changed since NCLB was enacted, the study compared achievement trends before and after 2002. Within each state, the study omitted test results that were not comparable because the state had made changes to its testing program. Finally, the study used measures in addition to the percentages of students scoring proficient on state tests.

- Although test scores have gone up since the enactment of NCLB, it is difficult to say whether or to what extent they have gone up because of NCLB. It is nearly impossible to isolate a cause-and-effect relationship between NCLB and test score trends when states, school districts, and schools have simultaneously implemented many different yet interconnected reforms.

- Tests scores are not synonymous with achievement. Tests are imperfect and incomplete measures of how much students have learned. But for a wide-scale study of achievement, test scores are still the best means available to draw inferences about student learning.

## Background on the Study

Since the No Child Left Behind Act was enacted more than five years ago, it has spurred as many changes in elementary and secondary education as any federal policy in U. S. history. Most states have revamped and expanded their testing and accountability systems, and some have created these systems where none existed before. Districts and schools have revised their curricula, expanded programs for struggling students, and reorganized instructional time to meet the law's adequate yearly progress (AYP) requirements. Teachers have changed how they teach. And students continue to take more tests than ever.

A great deal hinges on the state reading and mathematics tests that NCLB requires students to take yearly in grades 3 through 8 and once during high school. The results of these tests are used to determine whether a school or district must take serious steps to raise achievement because it has been identified for improvement under NCLB; whether students are eligible for subsidized tutoring or public school choice; and, if low performance persists, whether teachers and administrators will be replaced or whether a school will be dramatically reorganized, converted into a charter school, operated by an outside contractor, or taken over by the state.

All of the sanctions in NCLB, and all of the changes brought about by the law, are aimed at accomplishing the same fundamental goal—to improve the academic achievement of all students, including those from historically underperforming subgroups. So in 2007, the year that NCLB is up for reauthorization by the Congress, the question that matters most is whether student achievement has gone up since the law took effect.

This report is the first product of a three-phase study of student achievement before and after NCLB. The study is being conducted by the Center on Education Policy, an independent nonprofit organization. For phases I and II, CEP has received invaluable advice from a panel of five nationally known experts in educational testing or policy research, and extraordinary technical support from our contractor, the Human Resources Research Organization.





## STUDY QUESTIONS, PURPOSES, AND DESIGN

The study on which this report is based aims to answer two big questions, to the extent they can be answered now:

1. Has student achievement in reading and math increased since No Child Left Behind was enacted?

2. Have achievement gaps between different subgroups of students narrowed since No Child Left Behind was enacted?

To explore these questions, CEP designed a three-phase study, with the help of the expert panel and HumRRO on phases I and II. During phase I, which lasted 14 months, HumRRO staff collected various types of test data and other information from every state. CEP and HumRRO analyzed these data to determine trends in overall student test scores and achievement gaps in states with sufficient data. Phase II of the study, which will be completed this summer, involves on-site interviews with state officials in 22 states. The goal of these interviews is to investigate further the trends uncovered during phase I, learn more about changes in state testing systems and factors affecting availability of test data, and gather information about how well specific requirements of NCLB are working and how they could be improved. Phase III, which CEP has designed and will carry out in the fall and winter of 2007-08, examines student achievement at the school district level in three states.

This report has two main purposes, one informational and one educational. The first purpose is to document our findings from phase I in response to the two study questions. With this report we have put the most comprehensive and current data available about student test results in reading and math from all 50 states into the hands of policymakers to inform their discussions about reauthorization. A special benefit is the series of 50 online profiles, one for each state, developed by CEP and HumRRO to accompany the report. Each profile contains a rich store of test results and other information for that state. The profiles can be accessed on the CEP Web site at www.cep-dc.org/pubs/stateassessment.

In this report, we have also provided our own analyses of the data, which we conducted as objectively as possible, with support from HumRRO and independent of any special interests.

Although many states test other subjects, this study focuses on reading and math achievement because these are the only two subjects that states are required to test for NCLB accountability purposes. Although NCLB requires states to test science by school year 2006-07, the science results are not used for accountability.

The second purpose of the report is to educate policymakers and others on what can and cannot be known about student achievement, based on available data. With the reauthorization of NCLB underway, people will use test scores to tell the story they want to tell. Everyone interested in NCLB needs to be very careful about reaching conclusions based on flawed or simplistic interpretations of data, or believing claims that go beyond what the data can support. Positive trend lines in test results may indicate that students have learned more, but they may also reflect easier tests, lower cut scores for proficiency, changing rules for testing, or overly narrow teaching to the test. Similarly, flat-line results could signal no change in achievement, or they could mean that the test is not a sensitive measure of the effectiveness of the instruction students are receiving. And not all states have sufficient, comparable data to allow valid conclusions to be drawn about trends in overall student achievement or performance gaps before and after NCLB took effect.

In this climate, it is critical that policymakers and the public understand the quality and limitations of the available test data, the types of data that are not routinely available, and the factors that could distort trends in test results. To fulfill this educational purpose, we have included information about these issues in chapter 7.



## CEP'S EXPERIENCE WITH NCLB RESEARCH

CEP is uniquely positioned to lead a study of student achievement since enactment of the No Child Left Behind Act. This special report on achievement trends represents a continuation—and in some ways the pinnacle—of a broader national study of federal, state, and local implementation of NCLB that CEP has been conducting since 2002. This broader study has been based on data from an annual survey of all 50 states, an annual survey of a nationally representative sample of between 274 and 349 responding school districts per year, and annual case studies of up to 43 school districts and 33 schools.

Since 2002, we have issued annual reports on our findings from this broader work. This year, we are publishing separate reports addressing different aspects of NCLB. Several have been released and more will be published in the coming weeks.[1] This report on achievement is part of that set. All of our past and current NCLB reports are available at www.cep-dc.org.

Since 2004, we have included questions about achievement in our state and district surveys. In separate questions about language arts and math, we asked state and district officials whether student achievement was improving, declining, or staying the same, based on the assessments used for NCLB. We also asked a series of questions about achievement gaps for specific subgroups of students. Whenever we have asked these achievement questions, the majority of state and district officials have responded that student achievement is improving

[1] In 2007, CEP has already published reports on NCLB state accountability plans, school restructuring in California and Michigan, state monitoring of supplemental educational services, and state capacity to administer NCLB. Forthcoming CEP reports on NCLB will deal with teacher quality requirements, assistance to schools in improvement, curriculum and instructional changes, school restructuring in Maryland, and Reading First.

and gaps are narrowing. In this year's state survey, for example, 25 states reported that student achievement was improving in language arts, based on test results from 2005-06; 15 states said that achievement in this subject was staying the same, and 3 reported that it was declining. In math, 27 states reported improvements in achievement, 11 reported flat achievement, and 4 reported declines (CEP, 2007a). Compared with last year's survey responses, fewer states reported improvements this year and more reported flat achievement.

These survey results were based on self-reports and had other limitations. Recognizing these limitations, we designed this achievement study, which builds on our prior NCLB research but goes well beyond it by examining test scores directly.

### ROLES OF THE EXPERT PANEL AND HUMRRO

To develop a sound methodology and provide expert advice at all stages of the study, CEP assembled a panel of some of the nation's top scholars on education testing and education policy issues. Panel members included the following:

- 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

Jack Jennings, CEP's president and CEO, chaired the panel. The panel met four times in Washington, D.C.: March 2006, September 2006, January 2007, and April 2007. At these meetings, the panel developed the study methodology, reviewed the initial state data, developed procedures for analyzing state data, and reviewed drafts of this report. CEP staff and consultants and HumRRO staff also attended the panel meetings. In addition, the panel held formal telephone conferences, reviewed study documents, and provided informal advice by e-mail and phone. The panel members' wealth of knowledge has contributed immeasurably to the quality of this study.

CEP contracted with HumRRO, a nonprofit research organization with considerable experience in evaluating testing systems and standards-based reforms, to collect, compile, and vet the quality of the enormous amount of data required for this achievement study. HumRRO also did the initial analysis of trends in each state. CEP is deeply grateful to the HumRRO staff, whose tireless and capable efforts have been essential to this study.

Although the panel members and HumRRO staff reviewed all drafts of this report, we did not ask them to endorse it, so the findings and views expressed here are those of CEP.

## Unique Features of This Study

From the first meeting of the expert panel, we set out to design the most comprehensive and thorough study of state test scores since the passage of NCLB. We wanted the study to be methodologically sound, feasible, and useful to policymakers and the public, and to build on previous research on this issue, including CEP's past research. We believe this study has met those objectives but, as explained below, it involved a much more intensive effort than we initially realized.

### COMPREHENSIVE SCOPE

A study of test score trends that focused on a limited number of states would immediately raise concerns about how the states were selected and whether conclusions were biased. Therefore, this study aimed to include various kinds of test results from as many states as possible. Ultimately, we obtained data from all 50 states. The effort involved, however, made us appreciate why this type of study has not been done before and why it cannot be easily replicated.

Most other studies of NCLB-related achievement use published state or local data on the percentages of students scoring at the proficient level on state tests, and take for granted that these data are accurate. Early in our research, however, we found holes and discrepancies in the published data, described in detail in chapter 7. Therefore, our study made an extra effort to have states verify the accuracy of the test data we gathered. Using the process outlined in the appendix, we sent state officials a complete file of the data we had collected for their state and asked them to check their accuracy, make corrections, and fill in missing information. State officials were also asked to sign a verification checklist. This verification process was long and complicated, as noted in chapter 7; most states did make modifications in their data.

A study of this breadth and depth would not have been possible without the cooperation of the states. We appreciate greatly the considerable efforts made by officials from individual states and from the Council of Chief State School Officers to voluntarily provide us with and verify their data.

Analyzing the data we collected was also a complicated and time-consuming process. In a typical state, the data tables developed by HumRRO included as many as 16,184 individual numbers, which HumRRO staff and a CEP consultant scrutinized to determine achievement trends.

The data used to arrive at the findings in this report represent the best information we could obtain by the mid-January collection deadline for phase I of the study. Still, the information in this report and the accompanying state profiles represents a snapshot in time.

### OTHER UNIQUE FEATURES OF THE STUDY

Working together, CEP, the expert panel, and HumRRO designed the phase I study to include the following unique elements:

- ***State-centered approach.*** Because each state has its own assessment system, aligned to different content standards and using different definitions of proficiency, it can be perilous to combine test results across states when analyzing achievement trends. Still, state tests are the main measure of achievement under NCLB and the best available standardized measures of the curriculum being taught in classrooms. This study makes separate judgments about achievement in each state, based on that state's test data, and then summarizes those judgments across states to arrive at a national picture.

- ***Pre- and post-NCLB trends.*** Many states had gotten serious about education reform years before NCLB took effect. A study that did not look at trends before NCLB would raise questions about whether gains in achievement after NCLB were just a continuation of previous trends. Furthermore, a study that did not include results from the most recent round of state testing would not accurately reflect current progress in achievement. To the extent possible, this study looks at test score trends before and after 2002, the year NCLB was enacted, to determine whether the trends have changed direction and whether the pace of improvement has sped up or slowed down since NCLB. On the advice of the expert panel, we tried to obtain test data from 1999 through 2006, or for whichever of these years states had data available. In nearly all states, the most recent data available during phase I of this study were from tests administered in 2005-06.

- ***Breaks in data.*** Often the test results reported for NCLB are treated as one long, uninterrupted line of comparable data. But as explained in chapter 3, many states adopted changes in their assessment systems since 2002 that created a "break" in test data—meaning that results after the change are not comparable to results before the change. This study only analyzed test results for years with an unbroken trend of comparable data.

- ***Additional analyses beyond percentages proficient.*** To make judgments about student achievement, NCLB relies mainly on a single indicator, the percentage of students scoring at or above the proficient level on state tests. Like every measure of achievement, this one has its limitations. As explained in chapter 3, we supplemented our analysis of percentages proficient, where possible, with rigorous alternative analyses based on effect sizes, which are derived from mean, or average, test scores. (Definitions of these and other technical terms can be found in the glossary at the end of this report.)



## Limitations of This Study

Even with the steps described above, this study makes judgments about student achievement based on less than perfect information. In addition to the test construction issues discussed in chapter 3, two broader types of limitations, described below, are particularly noteworthy:

### DIFFICULTY OF ATTRIBUTING CHANGES TO NCLB

This report focuses on whether student achievement has improved since the enactment of NCLB. It is very difficult to determine whether students are learning more because of NCLB. Isolating the cause-and-effect relationship of any education policy is often impracticable. With a policy as far-reaching as NCLB, it becomes nearly impossible when states, districts, and schools are simultaneously implementing so many different yet interconnected policies and programs. If student achievement has risen since NCLB took effect, is this due to NCLB, or to state or local reforms implemented at roughly the same time, or to both? If both, how much of the improvement is attributable to state or local policies and how much to federal policies? Using multiple methods of analyzing achievement will not tease out the causes of gains or declines.

In a similar vein, this study does not take a position on how well specific components of NCLB are working or whether the requirements in the current law are the most effective means to raise achievement and close test score gaps.

## AN IMPERFECT MEASURE OF ACHIEVEMENT

Like virtually all studies of achievement, this one relies on test scores as the primary measure of how much students are learning. But test scores are not synonymous with achievement. Although tests are often viewed as precise and very objective, they are imperfect and incomplete measures of learning. Only certain types of knowledge and skills get tested on large-scale state tests—generally those that can be assessed with questions that are easy to administer and score.

In addition, test scores can go up over time without actually indicating that students have learned more; for example, several researchers have observed a "bump" in scores in the first few years after a test has been introduced, as students and teachers become more familiar with its format and general content (Hamilton, 2003; Linn, Graue & Sanders, 1990; Koretz, 2005). Moreover, tests vary in their instructional sensitivity—in other words, how well they detect improvements due to better teaching (Popham, 2006).

Still, tests are the best means available to draw inferences about student learning, especially across schools, districts, and states. That is why test results, despite their limitations, are the focus of this study.

# Chapter 3
# Achievement Measures Used in This Study



## Key Findings

- The percentage of students scoring at the proficient level on state tests—the "magic number" used for NCLB accountability and the only measure of achievement the Act requires states to collect and report—may appear to be accurate, objective, and consistent, but in some cases it can be misleading. Definitions of "proficient" performance vary so much from state to state that the percentage proficient should not be used for comparisons between states. Even within the same state, percentages proficient may not be comparable from year to year due to federal and state policy changes. Moreover, the percentage proficient provides no information about progress in student achievement that occurs below or above the proficient level.

- In this study, we used the percentage proficient as one measure of achievement. To address its limitations, however, we used a statistical tool called effect size as a second measure. Because effect sizes are based on mean, or average, test scores in conjunction with the dispersion of scores, they capture changes in achievement below and above the proficient level. They also avoid a problem that arises when percentages proficient are used to analyze achievement gaps for student subgroups—namely, that the gaps between higher- and lower-achieving subgroups can look different, depending on where a state has set its cut score for proficient performance on the scoring scale for the test.

- This study used a set of rules, applied consistently across all states, to determine such issues as when breaks had occurred in the comparability of test data, when subgroup scores should be approached with caution, and what constitutes a trend in achievement.

- Even if an ideal amount of test score data had been available for every state, policymakers and others should still be cautious when interpreting test score trends because of the many ways that a state's test can change from year to year. There is a certain degree of "fuzziness" or potential distortion in state test results that is derived from the tests themselves and from the way they are created, administered, and scored.

## Complexities of Measuring Achievement

Measuring student achievement is a much more complex proposition than measuring a child's height with a yardstick. Although the tests used for accountability under the No Child Left Behind Act are a logical starting point for a study of achievement since the law took effect, there are different ways of looking at test data and different ways of defining improvement.

In this chapter, we review the limitations of the primary measure of achievement used by NCLB—the percentage of students scoring at or above the proficient level on state tests. We describe how we addressed many of these limitations by taking into account "breaks" in comparable data, supplementing the percentage proficient with additional measures, and flagging data that should be approached with caution. In addition, we explain the rules developed by CEP, HumRRO, and the expert panel for deciding which specific data to include in our analysis and determining whether improvement has occurred. We add some cautions about why test results—even if carefully analyzed in multiple ways—may still not provide a completely accurate picture of student performance trends. We conclude with a list of the detailed information available in the state-by-state profiles posted on the CEP Web site.

## Limitations of Percentages Proficient

The main measure used to gauge student achievement under the No Child Left Behind Act is the percentage of students scoring at or above the proficient level on state tests. The federal law does not define proficient performance. Instead, NCLB directs each state to set its own proficiency standard and measure student progress toward it with its own tests. Consequently, "proficient" means different things in different states. States vary widely in curriculum, learning expectations, and tests, and they have defined proficiency in various ways, using different cut scores. (The cut score is the score students must meet or exceed on a test to be counted as proficient.) States have also developed different rules for how to calculate the percentage proficient.





Even with this variety, the percentage proficient is NCLB's magic number—it determines whether schools and districts are making adequate yearly progress toward the goal of 100% proficiency in 2014 or whether they are "in need of improvement." It is also the only form of test data that states are required by the Act to collect and report, which they must do for the state as a whole and for school districts, schools, and subgroups of students.

On one hand, the percentage proficient measure has the advantage of being easily understood by policymakers, the media, parents, and the public. It also addresses the concern that large numbers of students are not achieving at an adequate level by giving a snapshot of how many students have met the performance expectations set by their state. On the other hand, the percentage proficient has limitations as a measure of whether student achievement has increased. People assume that this measure is accurate, objective, and consistent, but in reality it can sometimes be misleading. Three limitations of the percentage proficient are particularly problematic in studies of achievement trends over time: a lack of comparability within the same state, a lack of comparability across states, and omission of progress above and below the proficient level.

### THE PROBLEM OF COMPARABILITY WITHIN STATES

Since NCLB was first enacted, states have made policy changes over the years that have affected calculations of the percentage proficient. Although these changes have been made with the approval of the U.S. Department of Education (ED), they can influence the comparability of percentages proficient from one year to the next in the same state. In essence, certain changes have made it easier for some students to be deemed proficient even if they haven't learned more. As a consequence, 62% proficient in 2006 may not mean the same thing as it did in 2002. Similarly, an increase from 62% to 72% proficient between 2002 and 2006 does not necessarily mean that students' raw test scores have gone up a proportionate

amount or that students have learned that much more. Rather, an indeterminate amount of this increase may be due to policy changes, including some of the changes described in depth in a recent CEP report on state accountability plans (CEP, 2007c).

One notable state change that affects the percentage proficient involves retesting—students retaking a state test (typically a different form of the same test) that they had not passed the first time. Initially, ED held to the "first administration" rule for tests used for NCLB—the score that counted was the one a student earned the first time the test was taken. Many states, particularly those with high school exit exams, allow students multiple opportunities to pass, which conflicted with ED's rule. In 2004, ED revised its policy and began permitting states to count proficient scores on retests for AYP purposes, and to "bank" the scores of students who pass the exams early and count these scores as proficient in a subsequent year.

In another relevant policy change, a few states put a "standard error of measurement" of plus or minus a few points around an individual student's test score.[2] This practice is intended to address measurement error that occurs in test scores due to differences in the sample of questions that appear on different forms of the same test, student guessing, students' physical condition or state of mind, distractions during testing, less than perfect agreement among raters who score written responses to open-ended test questions, and other factors unrelated to student learning. In states that use this type of standard error, some students are counted as proficient even though their scores fall slightly below the proficiency cut score. This has the effect of inflating the percentage proficient figure.

Changes in federal regulations and guidance have also had an impact on percentage proficient calculations. Most notably, ED issued major rule changes that affected which students with disabilities and limited-English-proficient (LEP) students are tested for NCLB accountability purposes, how they are tested, and when their test scores are counted as proficient under NCLB (see **box A**). Ultimately, these adjustments have made it easier for some students in these subgroups to be counted as proficient, which in turn has affected the comparability of test results for these subgroups over time. The impact has been significant enough to make it inadvisable to draw comparisons of the performance of these two subgroups between 2002 and 2006.

The comparability of the percentage proficient measure within the same state can also be affected by significant shifts in subgroup demographics and numbers. Many states have experienced rapid growth in the Hispanic and LEP subgroups. For example, in just two years (2004 to 2006), Tennessee saw a 46% increase in the number of students in the Hispanic subgroup in 4th grade, as measured by the number of students taking the state reading test.

Rapid changes in the number of students tested can affect achievement trends in ways that do not reflect the effectiveness of an educational system, complicating efforts to determine trends across years for the same subgroup or to compare trends in gaps between different subgroups.

[2] Many more states use a related method referred to as a "confidence interval," which puts a band of plus or minus a few percentage points around a school's or subgroup's percentage proficient for the purpose of determining adequate yearly progress. However, this technique does not affect the percentage proficient that is reported at the state level. For more on confidence intervals, see CEP, 2005.

## Box A. Students with Disabilities and Limited-English-Proficient Students: A Moving Target

Since the No Child Left Behind Act took effect, states and school districts have encountered continuous problems in attempting to square the law's testing and accountability requirements with the unique needs of students with disabilities and limited-English-proficient students. In response, the U.S. Department of Education made several policy changes to accommodate these subgroups while still holding states, districts, and schools accountable for these students' performance in the same way as other subgroups. Over the past few years, these policy decisions have affected which students are counted in these subgroups, which students are tested, how they are tested, and how their test scores are counted.

Before NCLB, it was not uncommon for students with disabilities and LEP students to be exempted from standardized testing altogether or given different tests than other students (National Research Council, 1997). NCLB included a requirement for students in these two subgroups to be tested with the same tests and standards as other students, but in the early years of the law, some school districts were unsure how to implement this requirement, and states and districts had various policies for how to test students in these subgroups (CEP, 2003; 2004). Some districts gave the regular state test with no modifications, which made it difficult for students with cognitive or learning disabilities to score at the proficient level. Other districts made liberal use of test accommodations or modifications and tested some students with disabilities with assessments geared to their learning level (alternate standards) rather than their grade level—practices that likely helped some students reach a proficient score.

Experience has shown that it is very difficult for the subgroup of students with disabilities to score at the proficient level on regular state tests. In 2003, ED issued regulations that allowed states to give students with significant cognitive disabilities an alternate assessment geared to alternate standards. However, the number of scores from these alternate assessments that were counted as proficient for AYP purposes could not exceed 1% of all tested students. Another policy change in April 2005 expanded the opportunities for students with disabilities to take alternate assessments by allowing additional students to be tested against "modified" standards, with a cap of 2% of all students. The modified tests allowed more students with disabilities (and to a lesser extent, all students) to be counted as proficient, but it also meant that the percentages proficient for the disabilities subgroup would not be truly comparable from year to year and would not be a reliable measure of achievement trends.

Federal policy changes have also affected the comparability of test results for LEP students. Under ED's initial, strict interpretation of NCLB, this subgroup as a whole could not, by definition, achieve proficiency in English because once a student became proficient in the English language, he or she was moved out of the LEP subgroup, and those remaining were not proficient. In February 2004, ED issued a policy allowing states to exempt immigrant students in their first year of enrollment in a U.S. school from taking the regular state English language arts test. States could also include former LEP students in the LEP subgroup for two years after they reached English proficiency, which of course increased the percentage of LEP students scoring proficient in reading or English language arts. But again, this policy change means that the percentage proficient for this subgroup would not be a reliable indicator of trends in achievement.

For these reasons, we do not include trends for the students with disabilities and LEP subgroups in the national summary on achievement gaps in chapter 5. We do, however, report performance for these subgroups within the individual state profiles.

*Source: Center on Education Policy, 2007c.*

## THE PROBLEM OF COMPARABILITY BETWEEN STATES

Because the definition of "proficient" varies so much from state to state, it is inadvisable to use the percentage proficient measure to compare student achievement in one state with student achievement in another. A seemingly identical outcome—say, 75% of high school students scoring proficient on a math test—will mean two different things in two different states in terms of what students actually know and can do in math. For this reason, we have avoided making these sorts of state-to-state comparisons in this report, and we strongly urge others to avoid doing so.

Many of the policy changes described above that affect the comparability of the percentage proficient measure within the same state also affect its comparability between states. Between-state comparisons are further confounded by decisions about where to locate the cut score on the scoring scale for a particular test. States can set a low cut score or a high one, so that more students or fewer students are deemed proficient, and states have made very different choices. In Tennessee, 88% of 3rd grade students reached the proficient level in math last year; in Hawaii, the figure was only 30% proficient. It is unlikely that there are such huge discrepancies in student achievement between these states. It is more likely that these results largely reflect differences in the difficulties of the tests and the location of proficiency cut scores.

The location of the cut score also creates problems in comparing trends over time in the percentage proficient across states. This is because the same amount of percentage point increase means different things at different points on the score distribution. A 10% increase in the percentage proficient is much more difficult to achieve when it involves an improvement from 85% to 95% than when it involves a gain from 50% to 60%. Moreover, the location of the proficiency cut score can affect how large achievement gaps between subgroups appear to be, and can make it difficult to accurately compare progress in narrowing these gaps (see **box B**).

## THE PROBLEM OF ONE LEVEL OF ACHIEVEMENT

A persistent criticism of the percentage proficient measure raised by educators is that it provides a picture of student test performance that is limited to just one level of achievement—proficient—and provides no information about achievement above or below that level.

For example, in a school with large numbers of low-performing students, teachers and administrators may be working very hard to improve achievement and may be making progress in boosting students from the "below basic" to the "basic" levels but raising fewer students to the higher level of "proficient." It is possible for test scores to increase without that increase being reflected in the percentage proficient if the increase occurs below the proficient level. Despite progress at the lower achievement levels and increasing test scores, a school or district would fail to make adequate yearly progress under NCLB and would be subject to the law's sanctions. Similarly, schools do not receive credit for gains by students who are already performing at or above the proficient level. In response to this problem, ED has recently allowed states to experiment with "growth models" to calculate adequate yearly progress, but only a few states have received permission to use these methods so far.

## Box B. Cut Scores and Gaps

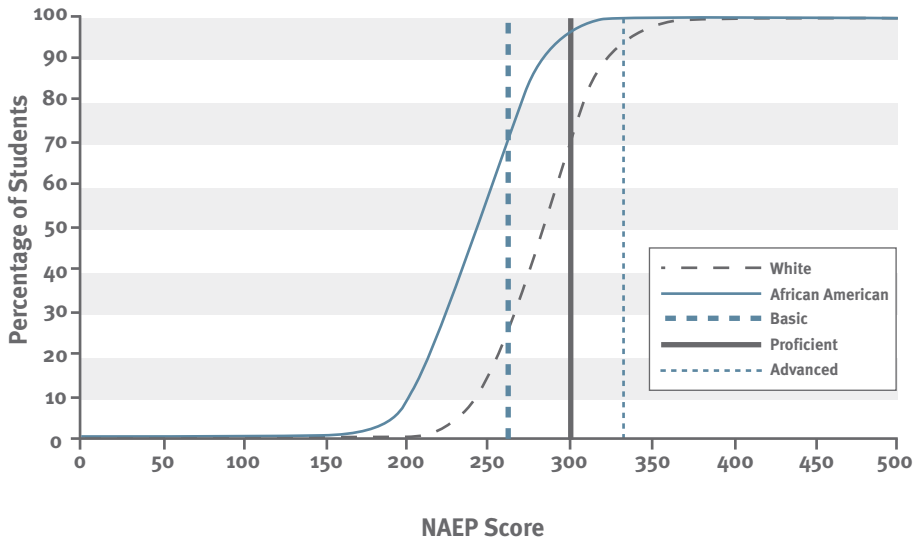
Some states appear to have shown more progress than others in narrowing gaps in percentages proficient between, for example, African American and white students. Can one conclude that educational practices aimed at reducing these gaps are working better in one state than another? Not necessarily.

Changes in instruction can affect the size of achievement gaps, but so can other factors. An important issue to consider when looking at achievement gaps is the location of the proficiency cut score—that is, the test score students must reach or exceed to be considered "proficient" on a test. Research shows that where the proficiency cut score is set makes a difference in the apparent size of the gap. If a proficiency cut score is very high or low, so that almost everyone reaches the cut score or almost nobody reaches it, there is little apparent gap. A cut score closer to the mean test score will be more sensitive to the achievement gap.

This was illustrated graphically by Paul Holland (2002). **Figure 1** shows the results of the 2000 administration of the math portion of the National Assessment of Educational Progress for 8th grade African American and white students. The test was scored on a scale of 0-500, with the cut score for the "basic" level of achievement set at 262, "proficient" at 299, and "advanced" at 333.

The graph shows the percentage of students in each group that scored *at or below* a certain level on NAEP. The x axis is the score, and the y axis is the percentage of students achieving at or below that score. So, about 25% of white students scored at or below 262 (basic)—marked with a dashed vertical line in figure 1—while 75% exceeded this score. About 70% of African American students scored at or below 262, while about 30% exceeded this score. Therefore, at the basic level, the achievement gap between African American and white students is about 45 percentage points—quite large.

**Figure 1. African American/ White Achievement Gap**
**NAEP Mathematics 2000 Grade 8**




However, the achievement gap picture changes as one moves along the score scale. At the "proficient" level of 299—marked with a solid vertical line in figure 1—the black/white gap shrinks to about 30 percentage points. As one moves toward the advanced cut score of 333 (shown in figure 1 as a dotted vertical line), the gap continues to shrink until it reaches about 6 percentage points at the advanced level. The same is true at the low end of the scale, where the gap is also a lot smaller.

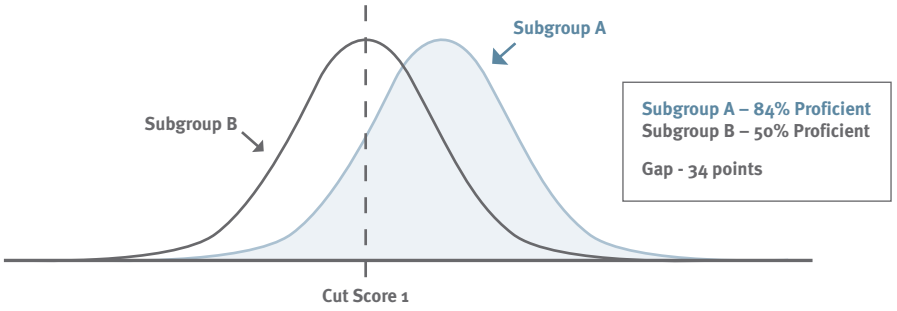
This NAEP illustration shows that focusing on a cut score of 262, 299, or 333 will have a dramatic impact on the apparent size of the achievement gap between African American and white students. The gap is larger at the middle of the NAEP score scale than at the extremes.

## Box B. (continued)

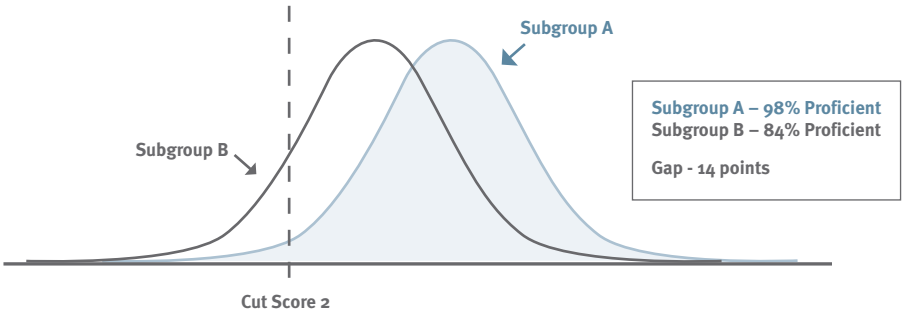
Another way of illustrating this phenomenon is provided in **figure 2**, which consists of two normal distributions of test scores for two subgroups of students, subgroup A and subgroup B. A normal curve graphically represents the typical way that students' scores are distributed on a test. Most scores are fairly close to the middle or average, and fewer scores are at the very low or very high ends. A hypothetical example is displayed in figure 2: the initial cut score (cut score 1) is set so that 84% of the students in subgroup A score at or above the cut score, compared with 50% of the students in subgroup B. (The areas to the right of the cut score under both curves represent the students who pass.) Therefore, the gap in percentages proficient between the two groups is 34 percentage points.

If a state were to set an easier cut score, represented by cut score 2 in figure 3, more students would meet or exceed it. At that point, 98% of subgroup A students and 84% of subgroup B students would pass, and the achievement gap would be reduced to 14 percentage points.

### Figure 2. Size of Gaps in Percentages Proficient with a Cut Score at the Mean




### Figure 3. Size of Gaps in Percentages Proficient with a Lower Cut Score




Therefore, anyone examining achievement gaps must take into account the location of the proficiency cut score. If a cut score is very high or low, there is little apparent gap. A cut score closer to the mean test score is a more sensitive measure of the achievement gap. In addition, discussions of changes in achievement gaps ideally should take into account any possible changes in cut scores.

*Source: Holland (2002), and Center on Education Policy.*

## How We Addressed the Limitations of the Percentage Proficient

Our analyses began with the percentage proficient because this is the primary measure used to track progress under NCLB and is available from every state. Then, to address the limitations of the percentage proficient measure described above and other factors that could skew achievement trends, we took two additional steps: carefully identifying breaks in testing programs that limit trend analyses, and analyzing student achievement using different measures. These other measures by themselves are not perfect either. They cannot take into account differences between states in the standards and tests. They do provide us, however, with alternatives or additional data points that can support or contradict percentage proficient trends. In some cases, limited comparisons can be made between states using these alternative measures.

### IDENTIFYING BREAKS IN TESTING PROGRAMS

Many states have changed their testing systems since 1999 as a result of NCLB and/or their own state policies. Often these changes create a break in the test data that makes it inadvisable to compare test results before and after the change. For instance, if a state introduced a new test in 2004, the results should not be compared with results from the old test given in 2003 (unless special equating procedures were conducted). Similarly, when a state introduces new state content standards that outline what students are expected to learn at different grades, usually the state must also redesign its testing program to ensure the tests are aligned with the new standards; this situation also results in a break in comparability. Chapter 7 describes the specific reasons we found for breaks in data.

Major changes, such as the adoption of a new test, are usually announced and explained to the public. But not all changes are publicized. Sometimes states change their cut scores, including the proficient score for NCLB purposes—a process that may or may not be done quietly. Once new cut scores are set, the percentage proficient results cannot responsibly be compared with those from earlier years. (Mean scale scores would still be comparable if the tests themselves had not been changed in other ways).

There are many educationally sound reasons why states make changes to their testing programs, such as better aligning tests with the curriculum taught in classrooms. Nevertheless, this situation makes it difficult or even impossible to track long-term trends in achievement. Ideally, the best data on trends come from states that had the same (or equated) assessments in place through all or most of the period from 1999 to 2006.

To determine whether states made alterations to their testing programs that could affect the comparability of test results during our period of analysis (1999 through 2006), we collected various descriptive information from each state, including major changes in testing programs. (The specific information collected is listed in the appendix.) Data from a state that introduced a new test or changed its proficiency cut score had to be examined closely, because often these data were not suitable for trend analyses. Identifying breaks in testing programs helped to address the problem described above of year-to-year incompatibility of test results in the same state. After identifying the breaks, we limited our analysis to those years that had an unbroken line of comparable test results.

## COLLECTING MEANS AND STANDARD DEVIATIONS

In addition to gathering data on percentages proficient, we also collected mean scale scores and standard deviations, explained below. These indicators give a more complete picture of the entire distribution of student performance, including performance above and below the proficient level. Examining these indicators helped us address the differences in state definitions of proficient performance and capture improvements across the entire range of high- and low-performing students.

Mean test scores provide a different perspective on student performance than the percentage proficient. The mean is the same as an "average" and is figured by adding up all the individual test scores in a group and dividing them by the number of people in the group. Mean test scores are expressed on an interval (numerical) scale and permit more rigorous quantitative analysis than a simple determination of whether a student falls into the proficient or not proficient category. Mean test scores also provide a more accurate measure of achievement gaps because, as explained in box B, the size of the gap depends highly on where the proficiency cut score is set.

When considered along with the percentage proficient, means provide additional information about the overall distribution of test scores. Consider a situation in which the percentage of students scoring at or above the proficient level in a particular state remains at 40% in both 2005 and 2006, suggesting no improvement. However, the state's mean test score might have gone up during that same period if students who were performing above the proficiency cutoff score achieved higher scale scores on the test in 2006. Or, to present another scenario, the mean could also rise if many students who scored below the proficient level earned higher scale scores but not enough to reach proficiency. In either case, the mean score might show progress not captured by the percentage proficient measure. Using mean scores also removes the uncertainty about comparability that arises when proficiency cut scores change. However, mean scores would not help to reveal trends in overall achievement trends or gap trends if the test itself has been changed.

The standard deviation is a measure of how spread out or bunched together test scores are within a data set. It is a way to measure the distance of all scores from the mean score. This statistic gives more information about the entire distribution of test scores than the percentage proficient does. A standard deviation can be calculated for any batch of test scores. If test scores are bunched close together (meaning all students score close to the mean), then the standard deviation will be small. Conversely, if test scores are spread out (meaning that many students score far from the mean), then the standard deviation will be large. **Box C** provides further explanation of standard deviations.

## ANALYZING CHANGES IN EFFECT SIZES

Using means and standard deviations, we were able to compute an effect size statistic called Cohen's D (Cohen, 1988; Willingham & Cole, 1997). Effect sizes provide a standard index to gauge achievement trends and gap trends; simply put, they are a measure of growth compared to a standard deviation.

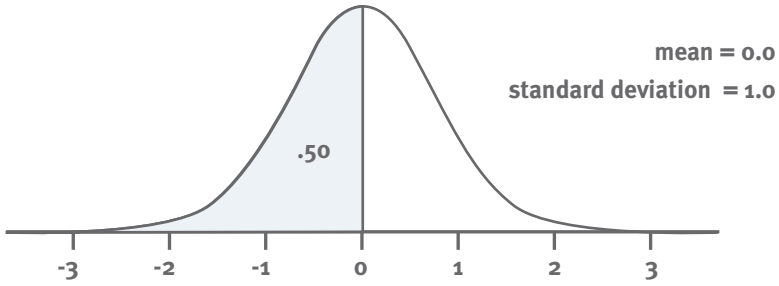
An effect size is computed by subtracting the year 1 mean test score from the year 2 mean score and dividing by the average standard deviation of the two years. Where there has been no change in the average score, the effect size is 0. An effect size of +1 indicates a shift upward of 1 standard deviation from the previous year's mean test score (in practice, effect sizes tend to be much smaller than 1 for mean changes from year to year). Even if two states have widely varying score scales and proficiency cut scores, the effect size statistic describes annual changes in the mean in terms of the tests' standard deviations.

Effect size results are a little more difficult for many readers to interpret than the percentage proficient. What does it mean, for example, that the reading score of Delaware 4th graders went up by a total of 0.19 of one standard deviation between 2002 and 2006? Is it a big

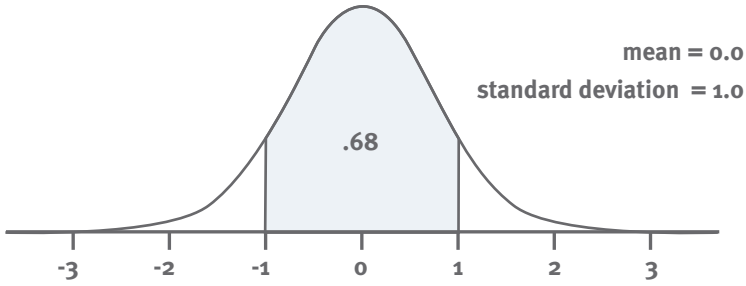
## Box C. What Are Standard Deviations?

Curve figures like the ones below are used to graphically represent the distribution of scores on any administration of a test. The largest numbers of test-takers' scores cluster close to the middle or high point of the curve, while fewer scores are situated at the low and high extremes.

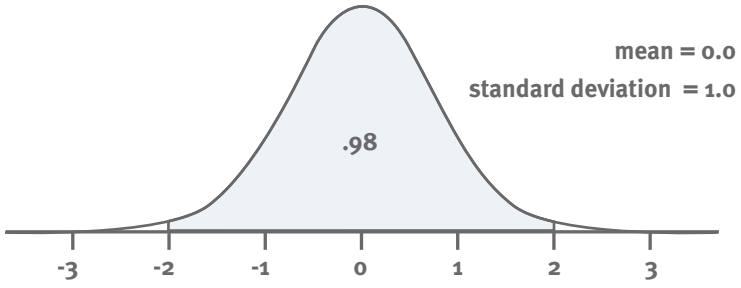
Three areas on a standard normal curve are useful for interpreting test scores. The first is at point 0, which is the mean, or average, test score. Fifty percent of the scores are below the mean and 50% are above.




The second area is within +1 or -1 standard deviation from the mean; 68% of the scores on a given test fall within this area. One standard deviation above the mean captures 34% of scores (half of 68%).




The third area of interest is between +2 or -2 standard deviations and accounts for 95% of the scores on a given test.




Continued on page 25

**Box C. (continued)**

Let's say that a test is scored on a scale from 1 to 1000 and that the mean score is 500 and the standard deviation is 80. This means that the scores of 68% of test-takers are between 420 (500 - 80) and 580 (500 + 80). Similarly, 95% of the scores would fall between 430 and 660.

Since the percentage of test-takers who score within one or two standard deviations of the mean is always the same for any test, the standard deviation is a common unit of measurement that can be used to make limited comparisons of groups of test-takers. In this study, effect size is used, which is the proportion of the difference between two years of test data or two subgroups of students in standard deviation units.

*Source: Stockburger (1996), and Center on Education Policy.*

improvement or a small one? It is still somewhat of a subjective figure, but broad comparisons are possible. Two Harvard University researchers, in an earlier review of NCLB, noted that between 1967 and 1982, scores of U.S students on the SAT college admissions test fell by 0.3 standard deviations. Between 1970 and 1982, high school science scores on the National Assessment of Educational Progress fell by 0.4 standard deviations. These trends were considered alarming at the time, and were among the factors that gave rise to the accountability movement in education (Peterson & West, 2003, pp. 3-5).

Let's say a hypothetical state showed an improvement of +1.0 standard deviations between 2002 and 2006. This would constitute a huge leap in student performance. Assuming a normal curve, as shown in box C, that gain would be the equivalent of an increase from 50% of students performing at the proficient level or above in 2002 to 84% in 2006.

Because effect sizes are based on mean test scores, they capture changes in student achievement above and below the proficient level. This helped us address one of the limitations of the percentage proficient measure. Effect sizes also have an advantage over mean scale scores alone in that they provide a standard index for gauging achievement and gap trends across states. Since state tests use different scoring scales (such as 1-100 or 1-500), it is difficult to interpret changes in mean test scores from one state to another. Effect sizes allow researchers to make some limited comparisons between states.

Effect sizes do have limitations. They do not enable comparisons within the same state if that state had a change in its testing program. To compare statistics within the same state, the test data for various years must be expressed on the same continuous scale because the computation involves subtracting one year's mean from another year's mean. If, for instance, a state used one scale for tests in 2001 and 2002, and then changed the scale for 2003 and 2004, one could compute an effect size for the differences in means from 2001 to 2002 and from 2003 to 2004. But one could not compute a comparable effect size from 2002 to 2003. Thus, one cannot use effect sizes to examine achievement trends unless a state has maintained a continuous scale that allows for these comparisons.

Effect sizes also do not take into account the relative difficulty of tests and standards from state to state. They only allow one to compare improvement or decline on each state's respective tests, such as comparing how much students in Minnesota have improved on Minnesota's state test with how much students in Vermont have improved on Vermont's state test. One still doesn't know which state's test is more difficult, or which state's students are achieving more. Furthermore, effect size is a measure of relative growth, not absolute performance. For

these reasons, we do not compare or rank states based on effect size results. Instead, we use the effect size as an additional source of information to examine achievement changes within each state, and then we summarize changes in each state to arrive at a national picture of student achievement since 2002.

We make no claim that effect size data are going to provide us with a perfect picture of student achievement since the inception of NCLB. Effect sizes are just another piece of information to supplement the percentage proficient measure, so we are drawing conclusions based on two sources of information about achievement rather than just one.

### FLAGGING SMALL AND CHANGING SUBGROUPS

To address the problem of changing composition of some subgroups, we collected the number of students tested (often referred to as N-counts) in reading and math at each tested grade and for each subgroup tracked under NCLB. In our analysis of achievement gaps, the N-counts were used to consider whether a subgroup of students was large and stable enough to draw valid inferences about achievement trends.

First, we flagged subgroups that were small with a footnote, because small groups are especially susceptible to year-to-year fluctuations in test scores that do not reflect actual changes in student achievement (see CEP, 2002 for a fuller discussion). We defined small as less than 5% of the total state student population tested at a grade, or fewer than 500 students.





Second, we flagged with a footnote any subgroups that had changed substantially in size during the period studied. When the size of a subgroup increases or decreases very rapidly in the course of a few years, this complicates trend analyses. Changes in test results may be due to changes in the composition of the subgroup as well as changes in achievement. We defined rapidly changing subgroups as those that changed by at least 25% up or down during the years analyzed. For all flagged subgroups, we have reported test results in the tables in the state profiles, along with specific cautions about interpreting them.

Third, we addressed the problem of changes in policies for students with disabilities and LEP students by deciding not to present trends for these two groups in the national summary of achievement gaps in chapter 5. There was no way to arrive at valid and reliable trends in achievement for these two subgroups for the reasons described in box A. We also noted in our analysis that trends for these two subgroups should be interpreted with caution. We did record test results and some trends for these two subgroups in the online state profiles, with reminders about the need for caution when drawing conclusions.

## Problems with Tests That We Did Not Address

Even in states that could provide data on percentages proficient, means, standard deviations, and numbers of students tested, one must still be cautious when interpreting test score trends. There is a certain degree of "fuzziness" and distortion in state test results that is derived from the tests themselves and the way they are created, administered, and scored. For instance, student performance may be affected by changes in the specific test questions that appear on each year's version of the test and by scoring procedures and cut scores. A state can adjust its test from year to year in ways that can affect the validity of inferences about changes in test scores over time.

In addition to the major changes noted above that create obvious breaks in data, test results can be affected by less explicit or unintentional changes. There can still be subtle manipulation of tests through a series of small decisions made by test administrators—tinkering rather than momentous changes. Following are some test construction issues and decisions that can affect the comparability of test scores from year to year:

- ***Test equating.*** To guard against breaches of test security, such as teachers and students memorizing test questions and answers, states use different forms of a test each year, composed of partially or entirely different test questions. Test developers try to make each test form similar in terms of the general content covered and level of difficulty. In addition, they often use a statistical technique called equating to make it possible to compare multiple forms of the same test with each other. Technical factors or the use of an incorrect equating methodology can introduce various types of errors into the equating process, producing forms that are not truly comparable (Kolen, 1988). A member of our expert panel has observed that typical equating procedures used by states can cause annual fluctuations in the percentage proficient of plus or minus 2% or more (Linn, n.d.)

- ***Weighting of test questions.*** Many state tests use a combination of multiple-choice, short-answer, and essay questions. Usually the questions that require students to write out their responses are worth more points on a test than multiple-choice questions; that is, they are more highly weighted. If the relative proportion and weighting of different types of test questions changes from year to year on a state's test, this can affect the comparability of scores.

- ***Changes to scoring procedures.*** Short-answer and essay questions must be scored by hand by trained scorers. If the scoring guidelines (called rubrics) or training procedures change even slightly from year to year, this can affect the comparability of test results.

- ***Re-use of test questions.*** For cost effectiveness and equating purposes, states often re-use some test questions across years. When entire test forms or large numbers of items are used repeatedly, students and teachers tend to become familiar with the questions and memorize the answers. While test scores may go up, the trend can be misleading if students have simply learned the answers to particular test questions but have not truly learned more about the larger subject being assessed.

These are just some examples of factors that can affect the comparability of test results from year to year in the same state. Even when accurate and complete test data are obtained, more subtle changes in state testing systems of the type described above can affect results. In an atmosphere of intense pressure to demonstrate achievement gains, administrators might err on the side of leniency when making these types of decisions. Based on test information that states make publicly available, it is often difficult to tell whether or how much any of the factors mentioned above actually distort the picture of student achievement derived from test scores in a state. These issues will be further explored during in-depth state interviews in phase II of this study.

## Rules for Analyzing Data

To analyze the state achievement data collected during phase I of this study, we took pains to develop consistent rules for analysis that would weed out incompatible data; identify trends that were consistent enough across grades and years to indicate a clear pattern of improvement; avoid "cherry picking" years, grades, or subgroups with the best or worst performance; and treat all states similarly and fairly. With extensive input from the expert panel, CEP and HumRRO arrived at the following rules for reporting and analyzing data:

- ***Grades analyzed.*** We looked separately at the elementary, middle, and high school levels for all of the achievement analyses. In states that tested multiple grades within each of these spans, we made decisions about which specific grades to report on and analyze based on a fixed set of criteria that were applied consistently across all states and developed without regard to whether achievement was high or low at a given grade. Generally, the grades selected were those with the longest trend lines. For analyses of effect sizes and achievement gaps, we selected one representative grade at the elementary, middle, and high school levels from among the grades included in the overall percentage proficient analysis. The first choices for these analyses were grades 4, 8, and 10, but if trend data were not available for these grades in a specific state, an adjacent grade was used in a fixed order. The detailed criteria for selecting grades are explained in the appendix.

- ***Years analyzed.*** If the state introduced a new test in 2005 or earlier, the analysis used that test and ended in 2005. If the state introduced a new test in 2006, the prior test was used in our analysis. Because many states introduced tests at different times in different grades, the years covered by our analyses sometimes varied at the elementary, middle, or high school level. For example, the analysis in a state might span 1999-2004 at the elementary and middle school levels but cover 2005-2006 at the high school level.

- ***Separate analyses for reading and math.*** Changes in achievement were analyzed separately for reading and math, since student performance in these subjects is often very different.

- ***Trend determinations.*** Differences involving just two years of data were referred to as "changes" in achievement rather than trends, since two years are too short of a period to discern whether a change is an actual trend or simply a normal year-to-year fluctuation in test results. On a similar note, we based our determinations of achievement trends on a broad pattern across multiple years, disregarding the kinds of small year-to-year fluctuations that typically occur in test results. For our findings about achievement gaps, we considered an increase or decrease in the gap for a specific subgroup to be a trend if it occurred in the same subject across all three grade spans analyzed (elementary, middle, and high school).

- ***Emphasis on average yearly gains.*** To even out the normal year-to-year fluctuations that occur with any test, we averaged gains or declines in test results across a period of years and focused on these average yearly gains in our analyses.

- ***"Slight" increases or decreases.*** We characterized an average change in achievement of less than 1 percentage point per year as a "slight" increase or decrease. This is because test scores are not perfect and include some measurement error resulting from factors unrelated to student learning, such as those listed earlier in this chapter in the discussion of standard error. "Slight" increases or decreases should be interpreted with caution because they may reflect measurement error rather than real changes in student achievement.

- ***Subgroups analyzed.*** The subgroups included in the achievement gap analyses were those tracked for accountability purposes in the NCLB law: major racial/ethnic groups in the state, low-income students, students with disabilities, and limited-English-proficient students. Within the state tables, we used the labels for these subgroups used by that particular state, so the subgroup names vary among states. When reporting subgroup trends, we did not mention subgroups that performed roughly the same as or higher than the comparison group of white students, as the Asian subgroup did in most states.

- ***Subgroup comparisons.*** For subgroups other than racial/ethnic groups, we compared the achievement of the subgroup of interest with the universe of other students who were not in that subgroup, whenever possible. For example, we compared low-income students with students who were not low-income when these comparison data were available. When test results for the comparison group were unavailable, we compared the group of interest to the state as a whole—for example, we compared low-income students with all students in the state. Although this latter approach is not the optimum one, it was the best option available in some states.

- ***Small or changing subgroups.*** As noted above, we flagged results for subgroups that were small or had changed significantly in size and included notes about interpreting results for these subgroups with caution.

- ***Special caution for students with disabilities and LEP students.*** As explained above, we avoided reaching national conclusions about these subgroups and cautioned people not to put too much stock in apparent trends for these subgroups.

Detailed information about other methods we used can be found in the appendix.

## Individual State Profiles on the Web

The findings in this report are based on state-by-state analyses of achievement data. These state analyses, along with the detailed data tables and figures on which they are based, have been packaged into profiles for every state. Individual state profiles can be viewed and downloaded from the CEP Web site at www.cep-dc.org/pubs/stateassessment. We encourage all readers who are interested in trends for a specific state to visit the Web site and look at the profile for that state. **Box D** lists the information contained in the state profiles.

## Box D. Contents of Profiles for Individual States

The state profiles available on the CEP Web site (www.cep-dc.org/pubs/stateassessment) contain the following descriptive information:

- ***Test characteristics.*** A list of the key characteristics of the reading and math tests used in the state for NCLB accountability, including the test name, grades tested, time of year when the test is administered, first year the test was administered, and major changes in the testing system since 2002.
- ***Summary of findings.*** The most important trends emerging from our analyses of state achievement data.
- ***Achievement trends.*** Findings from our analyses of overall trends in student achievement based on percentages proficient and effect sizes where available.
- ***Gap trends.*** Findings from our analyses of trends in achievement gaps based on percentages proficient and effect sizes where available.

Each profile also contains the following data figures and tables, based on the data available by the deadline for phase I of the study:

- ***Overall percentages proficient.*** Figures and tables for reading and math showing the percentages of students scoring at or above the proficient level for various grades at the elementary, middle, and high school levels. These figures and tables cover all of the years from 1999 through 2006 for which comparable data were available. The tables also show the average yearly gains or declines in percentages proficient before and after 2002, when NCLB took effect.
- ***Overall effect size data (where available).*** Figures and tables for reading and math displaying mean test scores, standard deviations, and effect sizes for one grade at each of three grade spans (elementary, middle, and high school). These figures and tables cover all of the years from 1999 through 2006 for which comparable data were available. The tables also show the average yearly gains or declines in effect size before and after NCLB.
- ***Gaps in percentages proficient.*** Tables for reading and math showing percentages proficient by student subgroup at three different grade levels for 2002 and 2006 (or for whichever adjacent years comparable data were available). Subgroups displayed on these tables include all the major racial/ethnic subgroups in the state, plus low-income students, students with disabilities, and limited-English-proficient students. These tables also show the percentage point gaps between various subgroups at the selected grades, changes in achievement gaps during the period analyzed, and average yearly gains or declines in gaps.
- ***Gaps by effect size (where available).*** Tables for reading and math showing gaps by effect size for subgroups of students for 2002 and 2006 (or for whichever adjacent years comparable data were available). Effect size data are included for three different grade levels for the subgroups of students listed above for which data are available. These tables also indicate changes in effect size gaps over the years analyzed and average yearly gains or declines in the effect size gap.
- ***Supplemental tables.*** Additional tables intended primarily for researchers. These include overall percentages proficient in reading and math converted to z-scores (defined in the glossary at the end of this report); gaps in percentages proficient converted to z-scores; and, where available, data on the number of test-takers in each subgroup for the period analyzed.

# Chapter 4
# Trends in Overall Achievement





## Key Findings

- The weight of evidence indicates that state test scores in reading and mathematics have increased overall since NCLB was enacted. All of our analyses—including percentages of students scoring proficient, effect sizes (a measure based on average, or mean, test scores), and pre- and post-NCLB trends—found substantially more states with gains in student test results than with declines since 2002.

- Regardless of whether one analyzes percentages proficient or effect sizes, the number of states showing achievement gains since 2002 is far greater than the number showing declines. (The subset of states with sufficient data varies, depending on the particular analysis.) For example, of the 24 states with both percentage proficient and effect size data for middle school reading, 11 states demonstrated moderate-to-large gains in this subject and grade span, while only one state exhibited a moderate or larger decline. Using percentage proficient data alone, 20 of the 39 states with this type of data showed moderate-to-large gains in middle school reading, while only one state showed a moderate or larger decline.

- Of the 22 states with both percentage proficient and effect size data, 5 made moderate-to-large gains in reading and math across all grade spans (elementary, middle, and high school) according to both measures. In other words, these five states showed gains according to all of the indicators collected for this study, allowing one to conclude with some confidence that achievement has gone up in those states. In reading, seven states showed moderate-to-large increases across all grades analyzed, according to both the percentage proficient and effect size measures. In math, nine states showed similar gains across all grades analyzed on both measures. (The group of seven and the group of eight states include the five states that made gains in both subjects.) The rest of the states had different trends at different grade spans.

- Elementary-level math is the area in which the most states showed improvements. Of the 25 states with both percentage proficient and effect size data in elementary math, 22 demonstrated moderate-to-large math gains at the elementary level on both measures, while none showed moderate or larger declines. Based on percentages proficient alone, 37 of the 41 states with trend data in elementary math demonstrated moderate-to-large math gains, while none showed declines of that magnitude.

- More states showed declines in reading and math achievement at the high school level than at the elementary or middle school levels. Still, the number of states with test score gains in high school exceeded the number with declines.

- Since many states had reform efforts well underway before NCLB, it is useful to know whether the pace of improvement has picked up or slowed down since NCLB took effect in 2002. Only 13 states supplied enough years of data for us to make this determination. In nine of these states, test results improved at a greater yearly rate after 2002 than before. In the other four states, the pre-NCLB rate of average yearly gain outstripped the post-NCLB rate.

- Analyzing changes in achievement using effect sizes generally produced the same findings as analyzing achievement using percentages proficient. But in some cases, the effect size analysis showed a different trend. For instance, in Nevada the percentage proficient in high school math decreased while the average test score increased. Conversely, in New Jersey the percentage proficient in middle school reading increased slightly, while the average test score dropped.

## How We Analyzed Overall Test Score Trends

Improving the academic achievement of all public elementary and secondary school students is a primary goal of the No Child Left Behind Act, along with closing achievement gaps. This chapter describes our findings about trends in overall achievement. We looked at two measures of achievement, where available:





- The percentages of students scoring at or above the proficient level on state tests—the primary measure of adequate yearly progress under NCLB

- Effect sizes, which are based on mean test scores, and standard deviations give a more complete picture of the entire distribution of student performance

We focused mainly on achievement trends since NCLB took effect in 2002. In states with available data, we also compared trends before and after NCLB.

To arrive at the findings in this chapter, we produced very detailed data profiles for each of the 50 states, consisting of up to four figures and 13 tables per state and narrative descriptions of that state's achievement trends. The figures and tables included a host of data on percentages proficient, mean scores, effect sizes, and other information described in chapter 3. (The state profiles can be found online at www.cep-dc.org/pubs/stateassessment/.) Using the data in the profiles, we closely analyzed achievement trends in reading and math within individual states, looking grade by grade across all the years. This was an enormous undertaking due to the amount of data involved. We then coded and compiled our findings from the 50 states to produce the tables in this chapter and develop a national picture of achievement trends. The appendix provides more detailed information about study methods.

During phase I of our study, we could not obtain from every state all of the data necessary to do pre- and post-NCLB analyses of percentages proficient and effect sizes. Therefore, the total number of states with sufficient data is different for each type of analysis:

- In 30 states, we obtained both percentages proficient and effect size data for at least some of the years between 2002 and 2006. In general, these are the states in which we have the most confidence about post-NCLB trends because we could use the two types of analyses as cross-checks. In 8 of these 30 states, data were missing for a particular grade level or did not cover enough years at a grade level to constitute a three-year trend. The number of states with sufficient data—as well as the specific subset of states—differs by grade span and subject.

- In all 50 states, we obtained percentages proficient for at least some of the years between 2002 and 2006. This group includes the 30 states described above, plus 20 states that did not make available effect size data by the phase I deadline but did have percentages proficient. In 16 of the 50 states, data were missing for a particular grade level or did not cover enough years at a grade level to constitute a three-year trend.

- Only 13 states provided sufficient achievement data for the years before 2002 to enable us to compare achievement trends before and after NCLB took effect.

When data were available for less than three years, this was mainly due to breaks in comparability caused by the introduction of new tests or changes in existing testing systems, such as changes in cut scores or content standards. Often these changes were made partly in response to the testing and accountability requirements of NCLB. Chapter 7 explains in detail which states provided various types of data and why data are limited or have breaks in comparability.

## Findings about Overall Test Score Trends Since 2002

Below we describe our findings about broad trends in achievement since 2002. We also spotlight trends by subject in reading and math and offer possible explanations for the trends we found.

### TRENDS BASED ON PERCENTAGES PROFICIENT AND EFFECT SIZES

In states with both percentages proficient and effect size data, we used both of these measures to analyze achievement trends.

#### Trends Across Three Grade Spans

**Table 1** displays the number of states with achievement gains in reading and mathematics at all three grade spans (elementary, middle, and high school) for the states with both percentage proficient and effect size data. Some states showed consistent gains across grade spans in both subjects while others states made consistent gains in just one subject. The rows of the table indicate the size of the gains. (As noted in chapter 3, we classified an average annual gain or decline in the percentage proficient as moderate-to-large if it equaled one percentage point or more and as slight if it equaled less than one percentage point.) None of the states with both percentage proficient and effect size data showed declines across all grade spans, so there are no rows to indicate them. There is also a row for states with mixed positive results, referring to a mixture of slight gains and moderate-to-large gains. For example, a state in this row may have had moderate-to-large gains at the elementary and middle levels, but only a slight gain in high school; or the state might have shown gains in two grade spans according to both percentages proficient and effect sizes but in the remaining grade span according to percentages proficient only.

**Table 1. Number of States with Various Test Score Trends Since 2002 Across All Three Grade Spans**
(Includes states with both percentage proficient and effect size data)

| Type of Trend | Both Reading & Math All Grade Spans | Reading All Grade Spans | Math All Grade Spans |
|---|---|---|---|
| Moderate-to-large gains in both percentage proficient and effect size | 5 | 7 | 9 |
| Slight gains[1] in both percentage proficient and effect size | 0 | 1 | 0 |
| Mixed slight and moderate-to-large gains in percentage proficient and/or effect size | 4 | 4 | 6 |
| Number of states with sufficient trend data for this analysis[2] | 22 | 22 | 22 |

Table reads: Since NCLB was enacted in 2002, five state have made moderate-to-large gains in both the percentages of students scoring proficient and in effect sizes in reading and math at all three grade spans analyzed (elementary, middle, and high school). Seven states have made gains at all three grade spans in reading, and nine made gains at all three grade spans in math.

[1] A "slight gain" means an average yearly gain of less than 1.0 percentage point.

[2] States with sufficient data have comparable test data for at least three of the years between 2002 and 2006.





Some notable findings from table 1:

- Of the 22 states with sufficient trend data for three grade spans in reading and math, five states—Delaware, Kansas, Kentucky, Louisiana, and Washington—demonstrated gains in both subjects and all three grade spans based on both the percentage proficient and effect size measures. In other words, these five states showed moderate-to-large gains according to all of the indicators collected for this study. Another four states demonstrated gains that varied in magnitude or by measure.

- Seven of the 22 states with sufficient data in reading showed moderate-to-large gains in this subject at all three grade spans, according to both the percentages proficient and effect size measures; these states include the five listed above plus Tennessee and Idaho. Nine of the 23 states with sufficient data in math made moderate-to-large gains in this subject at all three grade spans on both measures; these states include the five listed above plus Mississippi, New Jersey, Utah, and West Virginia.

- No state showed declines of any magnitude across all grade spans in either subject.

### Trends by Subject and Grade Span

**Table 2** summarizes trends in test scores separately by grade span and subject for the 30 states with both percentage proficient and effect size data. The specific number of states with sufficient data to analyze trends in each subject and grade span varied. The rows of the table display different types of achievement trends based on one or both measures, ranging from gains to declines. The top row of the table, for example, displays the numbers of states demonstrating moderate-to-large gains in both percentages proficient and effect sizes. The second row shows the number of states with gains in percentages proficient but a different trend in effect sizes, either a flat trend or a decline.

**Table 2. Number of States with Various Test Score Trends Since 2002 by Grade Span and Subject**
(Includes states with both percentage proficient and effect size data)

| Type of Trend | Reading Elementary | Reading Middle | Reading High | Math Elementary | Math Middle | Math High |
|---|---|---|---|---|---|---|
| Moderate-to-large gain in both percentage proficient and effect size | 14 | 11 | 10 | 22 | 17 | 12 |
| Moderate-to-large gain in percentage proficient only, contradicted by effect size | 1 | 0 | 1 | 0 | 0 | 1 |
| Slight gain[1] in both percentage proficient and effect size | 3 | 5 | 4 | 1 | 3 | 4 |
| Slight gain[1] in percentage proficient only, contradicted by effect size | 4 | 3 | 2 | 0 | 2 | 0 |
| Slight decline[1] in both percentage proficient and effect size | 2 | 2 | 1 | 1 | 2 | 1 |
| Slight decline[1] in percentage proficient only, contradicted by effect size | 0 | 2 | 2 | 1 | 0 | 2 |
| Moderate-to-large decline in both percentage proficient and effect size | 1 | 1 | 2 | 0 | 0 | 1 |
| Moderate-to-large decline in percentage proficient only, contradicted by effect size | 0 | 0 | 0 | 0 | 0 | 1 |
| Total number of states with sufficient data for trend[2] | 25 | 24 | 22 | 25 | 24 | 22 |

Table reads: Since NCLB was enacted in 2002, 14 states have made achievement gains in reading at the elementary level, based on both the percentage of students scoring at the proficient level and effect size.

[1] A "slight gain" or "slight decline" means an average yearly gain or decline of less than 1.0 percentage point.

[2] States with sufficient data have three or more years of comparable test data since 2002.

The columns show achievement trends separately for reading and math at each of the grade spans analyzed. For example, the column for elementary reading indicates that 14 states experienced moderate-to-large gains using both measures, but that 1 state had a gain in the percentage proficient which was not confirmed by its effect size trend. When the same upward trend appears across both measures (percentages proficient and effect sizes), one can conclude with some confidence that achievement has improved since NCLB was enacted. Each column also shows the total number of states with three or more years of comparable test data since 2002 in a particular subject and grade span.

In table 2, as in table 1, the number of states demonstrating gains in student performance since 2002 far exceeds the number showing declines. For instance, table 2 displays the following trends:

- Twenty-two of the 25 states with sufficient data experienced moderate-to-large gains on both measures in elementary math. In general, more states showed improvements at the elementary level than at the middle or high school level.
- The number of states with moderate-to-large gains far exceeded the number with slight gains.
- Only a handful of states—no more than five in a given subject and grade span—showed declines of any magnitude. No state showed a decline at all grade spans in reading or math. More states showed declines at the high school level than at the lower grades.





### Similarity between Percentage Proficient and Effect Size Trends

In most states with effect size data, the effect size analysis confirmed the trends in percentages proficient. This gives us a fair amount of confidence in the results. In some states, however, the findings did not converge, which suggests the importance of conducting both types of analyses. When the two measures diverge, this may be a signal to look carefully at the gains in percentages proficient and be cautious in drawing conclusions about overall achievement trends.

An example of divergent trends occurred in Nevada. **Table 3** shows the percentage of Nevada students performing at or above the proficient level in math. (Data were available in this state only for the years 2004 through 2006.)

**Table 3. Percentage of Nevada Students Scoring at the Proficient Level or Above in Math**

| Grade Level | Reporting Year 2004 | 2005 | 2006 | Post-NCLB Average Yearly Percentage Point Gain[1] |
|---|---|---|---|---|
| Grade 3 | 45% | 51% | 50% | 2.7 |
| Grade 5 | 50% | 51% | 55% | 2.4 |
| Grade 8 | 49% | 49% | 50% | 0.4 |
| Grade 10 | 52% | 51% | 47% | -2.8 |

Table reads: The percentage of Nevada 3rd graders who scored at or above the proficient level on the state math test increased from 45% in 2004 to 51% in 2005, then declined slightly to 50% in 2006. The average yearly gain in the percentage proficient at grade 3 was 2.7 percentage points after NCLB took effect (2004-2006).

[1] Averages are subject to rounding error.

School officials concerned with making adequate yearly progress under NCLB or parents wanting to know about the quality of Nevada schools might look at the percentages proficient in table 3 and see that students in grades 3 and 5 are making progress in math. But they might be worried about students in grade 10, where the percentage proficient in math declined by five percentage points in just two years—from 52% in 2004 to 47% in 2006.

Are Nevada high school students doing worse in math? Maybe not. **Table 4** shows student performance in terms of effect sizes.

Here we see that the mean math score of Nevada 10th graders actually *increased* between 2004 and 2006—from 288.6 to 293.1; in terms of effect sizes this was 8% of one standard deviation. This increase runs counter to the percentage proficient trend. This might have occurred because of improvement in mean scores among students either below or above the proficient level. A far more detailed analysis would be necessary to determine the exact reasons, but in any case, this example illustrates why multiple measures of performance are necessary to determine whether student achievement has increased since NCLB's inception.



**Table 4. Nevada Achievement Trends in Math in Terms of Effect Size**

| Grade Level | Reporting Year 2004 | 2005 | 2006 | Post-NCLB Average Yearly Percentage Point Gain[1] |
|---|---|---|---|---|
| Grade 5 | | | | |
| MSS *(SD)* | 294.7 *(69.1)* | 300.6 *(71.8)* | 302.0 *(70.8)* | |
| AAES | 0.00 | 0.08 | 0.10 | 0.05 |
| Grade 8 | | | | |
| MSS *(SD)* | 291.7 *(97.0)* | 291.9 *(98.2)* | 295.9 *(97.8)* | |
| AAES | 0.00 | 0.00 | 0.04 | 0.02 |
| Grade 10 | | | | |
| MSS *(SD)* | 288.6 *(58.5)* | 289.5 *(57.3)* | 293.1 *(57.7)* | |
| AAES | 0.00 | 0.02 | 0.08 | 0.04 |

Table reads: The mean scale score (MSS) of Nevada 5th graders on the state math test increased from 294.7 in 2004, to 300.6 in 2005, to 302.0 in 2006. The standard deviation (SD) for the mean scale score in 2004 was 69.1 (a statistic needed to calculate effect size). Using 2004 as a starting point (0.00), the accumulated annual effect size (AAES) for grade 5 math totaled 0.10 standard deviation units by 2006. For the period after NCLB (2004-2006), the average yearly gain in effect size for grade 5 was 0.05 standard deviation units (0.10 ÷ 2 years).

Note: Nevada's tests used for NCLB are scored on a scale of 100-500.

[1] Averages are subject to rounding error.

**Table 5. Number of States with Various Trends in Percentages Proficient Since 2002 Across All Three Grade Spans**

| Type of Trend | Both Reading & Math All Grade Spans | Reading All Grade Spans | Math All Grade Spans |
|---|---|---|---|
| Moderate-to-large gains | 7 | 9 | 19 |
| Slight gains[1] | 0 | 2 | 0 |
| Mixed slight and moderate-to-large gains | 8 | 9 | 8 |
| Number of states with sufficient trend data for this analysis[2] | 34 | 34 | 37 |

Table reads: Since NCLB was enacted in 2002, 7 states have made moderate-to-large gains in the percentage of students scoring proficient in reading and math at all three grade spans analyzed (elementary, middle, and high school). Nine states have made gains at all three grade spans in reading, and 19 have made gains at all three grade spans in math.

[1] A "slight gain" means an average yearly gain of less than 1.0 percentage point.

[2] States with sufficient data have comparable test data for at least three of the years between 2002 and 2006.





## TRENDS BASED SOLELY ON PERCENTAGES PROFICIENT

We analyzed trends in percentages proficient alone in all 50 states.

### Trends Across Three Grade Spans

**Table 5** displays the number of states with achievement gains in reading and mathematics at all three grade spans (elementary, middle, and high school), using percentage proficient data only. Table 5 includes those states from table 1 that had effect size data as well, plus additional states that only had percentage proficient data.

Table 5 largely echoes the results displayed in table 1. Seven states out of the 34 with sufficient data showed moderate-to-large gains across both subjects and all three grade spans. Nine states demonstrated moderate-to-large gains across all grade spans in reading, and 19 did so in math; these totals include the 7 states that made gains in both subjects. No state showed declines across all three grade spans in either subject.

### Trends by Subject and Grade Span

**Table 6** summarizes trends in achievement by subject and grade span for the states with percentage proficient data. Although all 50 states were able to supply varying amounts of percentage proficient data, the actual number with sufficient data to analyze trends across three years varied by subject and grade span.

**Table 6. Number of States with Various Trends in Percentages Proficient Since 2002 by Grade Span and Subject**

| Type of Trend | Reading Elementary | Reading Middle | Reading High | Math Elementary | Math Middle | Math High |
|---|---|---|---|---|---|---|
| Moderate-to-large gains | 29 | 20 | 16 | 37 | 32 | 26 |
| Slight gains[1] | 7 | 13 | 12 | 2 | 6 | 6 |
| Slight declines[1] | 3 | 5 | 5 | 2 | 2 | 6 |
| Moderate-to-large declines | 2 | 1 | 4 | 0 | 0 | 2 |
| Total number of states with sufficient date for trend[2] | 41 | 39 | 37 | 41 | 40 | 40 |

Table reads: Since NCLB was enacted in 2002, 29 states have made gains in reading at the elementary level, based on percentages of students scoring proficient.

[1] A "slight gain" or "slight decline" means an average yearly gain or decline of less than 1.0 percentage point.

[2] States with sufficient data have three or more years of comparable test data since 2002.





Table 6 shows the following trends:

- Thirty-seven states demonstrated moderate-to-large gains in math at the elementary level out of 41 states with sufficient trend data for this subject and grade span. In general, more states showed gains in percentages proficient at the elementary level than at the middle or high school levels.
- The number of states with moderate-to-large gains in percentages proficient far outnumbered those with slight gains.
- Declines in percentages proficient were less frequent. No more than nine states showed declines of any magnitude in a particular grade and subject. More states showed declines at the high school level than at the lower grades.

### Average Yearly Gains in Percentages Proficient

How much progress have states made in raising their percentages proficient? Since a certain amount of year-to-year fluctuation in test scores is normal, it is often more meaningful to look at average yearly gains than to simply compare one year's percentage proficient with another's. (The average yearly gain or decline is determined by computing the cumulative change over a period of years and dividing by the number of years.)

For each of the grade spans with two or more years of comparable data (the minimum period needed to compute average yearly gains), we calculated the median of states' average yearly gains since 2002. (The median is a sort of midpoint; an equal number of states fall above or below the median.) The results for reading are displayed in **table 7** and the results for math in **table 8**. In reading, the median of states' average yearly gains in percentage proficient since

**Table 7. Statistics on Average Yearly Gains in Percentages Proficient in Reading Since 2002**

| Statistic | Elementary | Middle | High |
|---|---|---|---|
| Median | 1.8 | 1.0 | 1.0 |
| Minimum | -2.2 | -2.0 | -2.9 |
| Maximum | 10.0 | 8.0 | 18.0 |
| Number of states with valid data | 49 | 49 | 46 |

Table reads: Of the 49 states with at least two years of comparable data, the median average yearly gain in the percentage proficient in reading was 1.8 percentage points per year at the elementary level. Among individual states, the average yearly gains in percentage proficient in elementary reading ranged from a minimum of -2.2 percentage points per year (a decline) to a maximum of +10.0 percentage points per year.

**Table 8. Statistics on Average Yearly Gains in Percentages Proficient in Math Since 2002**

| Statistic | Elementary | Middle | High |
|---|---|---|---|
| Median | 3.0 | 2.1 | 1.8 |
| Minimum | -0.9 | -1.0 | -4.0 |
| Maximum | 11.0 | 11.0 | 7.7 |
| Number of states with valid data | 49 | 49 | 47 |

Table reads: Of the 49 states with at least two years of comparable data, the median average yearly gain in the percentage proficient in math was 3.0 percentage points per year at the elementary level. Among individual states, the average yearly gains in percentage proficient in elementary math ranged from a minimum of -0.9 percentage points per year (a decline) to a maximum of +11.0 percentage points per year.



2002 was 1.8 percentage points per year at the elementary level, and 1.0 percentage points at both the middle and high school levels. In math, the median of states' average yearly gains was notably higher—3.0 percentage points per year at the elementary level, 2.1 at the middle school level, and 1.8 at the high school level. Above and below the median, the average yearly gains in individual states covered a wide spectrum. In elementary reading, the average yearly gains in percentages proficient ranged from a minimum of -2.2 percentage points (in other words, a decline) in one state to a maximum of +10.0 percentage points in another. In high school reading the range of average yearly gains was even broader—from a minimum of -2.9 percentage points (a decline) in one state to a maximum +18.0 percentage points in another.

## TEST SCORE TRENDS IN READING AND MATH SINCE 2002

In addition to analyzing broad trends across grades and subjects, we also took a closer look at separate achievement trends in reading and in math since 2002.

### Reading Trends

In reading, performance has increased since 2002. As already noted, seven states showed gains in reading across all three grade spans in both percentages proficient and effect sizes (table 1). Based on percentages proficient alone, 9 states demonstrated moderate-to-large increases across the three grade spans (table 5).

Within the same grade span, more states demonstrated increases in reading than declines. For example, based on both percentage proficient and effect size data (table 2), 14 states showed moderate-to-large gains at the elementary level, while just one state showed a decline. Based on percentages proficient alone, 29 states experienced moderate-to-large gains in elementary reading, while only 2 states experienced moderate-to-large declines (table 6).

Fewer states made improvements in reading at the high school level than at other grade levels. Based on both percentage proficient and effect size data (table 2), two states showed moderate-to-large declines at high school. Based solely on percentages proficient (table 6), four states experienced high school declines.

### Mathematics Trends

In math, performance has also increased since 2002. As table 5 illustrates, 19 states showed moderate-to-large gains in percentages proficient across all grade spans in math; among them were the 9 states with gains across all grade spans in effect sizes, as well (table 1).

Within the same grade span, more states experienced gains in math than declines. Of the states with both percentage proficient and effect size data (table 2), 22 showed moderate-to-large gains in math at the elementary school level, while none had declines. As with reading, improvements at the high school level were less striking than at other grade spans; only 12 states showed moderate-to-large gains at the high school level.

Results in math were most impressive at the elementary level. Based on proficiency data alone (table 6), 37 of the 41 states with sufficient data showed at least moderate gains in elementary math.

## POSSIBLE EXPLANATIONS FOR TRENDS SINCE 2002

Our evidence shows that gains in state test scores have far outweighed declines since NCLB took effect. Below we offer some possible explanations for the increases. The list is not exhaustive, but these are the explanations most often mentioned in research on test trends. Any or all of these factors in combination may be contributing to these trends. Moreover, different explanations could apply to different states or school districts within states.

- ***Increased learning.*** One likely reason for the upward trends in state test scores is that students are learning more and consequently are doing better on state tests. Administrators and teachers have made major efforts to improve achievement, according to CEP's case studies and nationally representative survey of school districts. According to this year's district survey, the following four strategies were most often considered successful in raising achievement in Title I schools identified for improvement under

NCLB: hiring additional teachers to reduce class size (cited as at least somewhat successful by 97% of districts with such schools), providing assistance through school support teams (95%), increasing the quality and quantity of teacher and principal professional development (92%), and aligning curriculum and instruction with standards and/or assessments (91%) (CEP, 2007a).[3] As noted in chapter 2, however, it is not possible to sort out how much of the impetus for these types of changes has come from NCLB and how much from state and local reforms. In CEP's surveys, roughly 7 of 10 district respondents cited school district programs and policies unrelated to NCLB as important causes of improved achievement in reading and math, and more than a third cited state programs and policies (CEP, 2007a).

- ***Teaching to the test.*** Teaching to the test can be a positive practice when it means aligning curricula to well-designed standards and tests and ensuring that classroom teaching covers the most important knowledge and skills contained in those standards (CEP, 2002). Teaching to the test can have adverse effects, however, if it means narrowing the curriculum to include only the subjects, topics, and skills that are likely to appear on state tests. This latter practice can raise test scores without also improving students' mastery of the broader subject being tested. It can give the false impression that student achievement is rising when students are actually learning the same amount or less; this is sometimes referred to as "score inflation" (Koretz, 2005). CEP's past district surveys and case studies found evidence that many school districts are reducing time in other subjects to allow more time for reading and math (CEP, 2006).[4]

- ***More lenient tests, scoring, or data analyses.*** We were careful not to compare test data when we were aware of breaks in comparability due to major changes in testing systems. But as explained in chapter 3, test results can still be subtly manipulated through a series of small decisions that affect such factors as equating, scoring, and proficiency levels and that amount to tinkering rather than substantial alterations. Faced with intense pressure to show achievement gains, state education officials may be likely to err on the side of leniency when making these types of decisions. It is difficult to find evidence of these types of subtle changes to state testing programs; however, we do know that some of the changes that states have made to their NCLB accountability plans have increased the numbers of students counted as proficient (CEP, 2007c).

- ***Changes in populations tested.*** Changes in the student population tested from year to year can affect aggregate state test scores. To cite one example, if significantly more students are held back in grade, it could appear that achievement in a particular grade has increased from one year to the next; for instance, the students who are retained in 4th grade may do better on the 4th grade tests after repeating a grade, while the cohort of students in 5th grade will not include the lowest-achieving students who had not been promoted. To cite a contrasting example, if one year's cohort of test-takers includes a significantly higher proportion from historically low-performing subgroups, such as limited-English-proficient students, than the previous year's cohort did, achievement may appear to decrease in the aggregate, but the apparent decrease is a consequence of demography rather than learning.

Phase II of this study, which involves actual visits to a subset of states and interviews with state officials, will explore alternative explanations for test score trends in participating states.

[3] More detailed information about strategies for raising achievement in schools identified for improvement is included in a new report from CEP, scheduled for release in June 2007.

[4] Additional information about changes in instructional time, curriculum emphasis, and test preparation related to NCLB is included in a second new report from CEP, also scheduled for release in June 2007.

## Pre- and Post-NCLB Trends

Knowing whether test scores have improved since NCLB took effect is only part of the national picture. Since many states began implementing reform efforts well before NCLB was enacted, it is also important to determine whether the pace of improvement has sped up or slowed down since 2002. To make this determination, we compared average yearly gains in achievement before and after the law's enactment in 2002, using both percentages proficient and effect sizes, where available. In our analysis, the pre-NCLB period ended at 2002, and the post-NCLB period started at 2002.

Only 13 states met the criteria necessary to make pre- and post-NCLB comparisons—at least two years of data before and after 2002.

### STATES WITH GREATER POST-NCLB GAINS

Of the 13 states with sufficient pre- and post-NCLB data, 9 made greater average yearly gains in achievement after NCLB was enacted than before, by most indicators. They include Kansas, Kentucky, Louisiana, New Hampshire, New Jersey, New York, Pennsylvania, Washington, and Wyoming.

**Table 9** compares average yearly gains in achievement before and after NCLB in each of these nine states. Pre- and post-NCLB comparisons are made for every grade span that had sufficient trend data, using percentages proficient and effect sizes where available. In the table, each measure (percentage proficient or effect size) and each grade and subject (such as grade 4 reading or grade 4 math) is counted as one point of comparison. In the far right column is a statement summarizing how many points of comparison showed greater post-NCLB gains than pre-NCLB gains. For example, Kansas had 12 points of comparison—two measures (percentages proficient and effect size) times six grades (three grades in reading and three in math). New York had four points of comparison—one measure (percentages proficient) times four grades (two in reading and two in math). In Kansas, post-NCLB gains exceeded pre-NCLB gains on all 12 points of comparison; in New York, post-NCLB gains exceeded pre-NCLB gains on three of the four points of comparison.

**Table 9. Average Yearly Gains by Subject and Grade for States with Greater Overall Gains After NCLB Than Before**

| State, Years of Data, Subject, Grade Level | Average Yearly Percentage Point Gain | | Average Yearly Effect Size Gain | | Comparisons of Pre-NCLB v. Post-NCLB Gains |
|---|---|---|---|---|---|
| | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | |
| **Kansas 2000-2005** | | | | | |
| Reading 5 | -0.1 | 4.9 | 0.00 | 0.14 | |
| Reading 8 | -0.5 | 3.5 | -0.01 | 0.10 | |
| Reading 11 | -0.9 | 2.9 | -0.03 | 0.08 | Post-NCLB gains exceed pre-NCLB |
| Math 4 | 2.6 | 5.7 | 0.08 | 0.20 | gains on 12 of 12 comparisons |
| Math 7 | 1.1 | 4.1 | 0.03 | 0.11 | |
| Math 10 | 0.7 | 2.6 | 0.02 | 0.06 | |

| State, Years of Data, Subject, Grade Level | Average Yearly Percentage Point Gain | | Average Yearly Effect Size Gain | | Comparisons of Pre-NCLB v. Post-NCLB Gains |
|---|---|---|---|---|---|
| | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | |
| **Kentucky 1999-2006** | | | | | |
| Reading 4 | 1.3 | 2.7 | 0.03 | 0.06 | |
| Reading 7 | 1.7 | 2.0 | 0.03 | 0.05 | |
| Reading 10 | 1.7 | 3.3 | 0.04 | 0.07 | Post-NCLB gains exceed pre-NCLB |
| Math 5 | 2.7 | 5.3 | 0.08 | 0.11 | gains on 11 of 12 comparisons |
| Math 8 | 1.0 | 2.0 | 0.05 | 0.06 | |
| Math 11 | 1.7 | 2.0 | *0.06* | *0.04* | |
| **Louisiana 1999-2006** | | | | | |
| Reading 4 | 0.7 | 1.8 | *0.04* | *0.02* | |
| Reading 8 | 1.7 | 1.8 | *0.06* | *0.03* | Post-NCLB gains exceed pre-NCLB |
| Math 4 | 2.7 | 3.0 | 0.08 | 0.08 | gains on 5 of 8 comparisons |
| Math 8 | 1.0 | 3.0 | 0.04 | 0.05 | |
| **New Hampshire Years vary as shown** | | | | | |
| Reading 3 2000-2004 | 1.5 | 1.0 | | | |
| Reading 6 2000-2004 | -0.5 | 6.0 | | | |
| Reading 10 2000-2006 | 1.5 | 3.0 | No effect size | | Post-NCLB gains exceed pre-NCLB |
| Math 3 2000-2004 | -0.5 | 5.5 | data available | | gains on 5 of 6 comparisons |
| Math 6 2000-2004 | 0.5 | 2.5 | | | |
| Math 10 2000-2006 | -0.5 | 3.8 | | | |
| **New Jersey 1999-2006** | | | | | |
| Reading 8 | -1.4 | 0.3 | *-0.02* | *-0.03* | |
| Math 4 | 2.7 | 3.5 | 0.08 | 0.10 | Post-NCLB gains exceed pre-NCLB |
| Math 8 | -1.0 | 1.6 | -0.04 | 0.03 | gains on 5 of 6 comparisons |
| **New York 1999-2005** | | | | | |
| Reading 4 | 4.3 | 3.0 | | | |
| Reading 8 | -1.3 | 1.3 | No effect size | | Post-NCLB gains exceed pre-NCLB |
| Math 4 | 0.0 | 6.0 | data available | | gains on 3 of 4 comparisons |
| Math 8 | 2.3 | 3.5 | | | |

| State, Years of Data, Subject, Grade Level | Average Yearly Percentage Point Gain | | Average Yearly Effect Size Gain | | Comparisons of Pre-NCLB v. Post-NCLB Gains |
|---|---|---|---|---|---|
| | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | |
| **Pennsylvania**[1] | | | | | |
| Reading 5 | 0.9 | 0.9 | *0.02* | *-0.01* | |
| Reading 8 | -1.3 | 3.0 | 0.00 | 0.12 | |
| Reading 11 | 0.9 | 1.5 | 0.03 | 0.04 | Post-NCLB gains exceed pre-NCLB |
| Math 5 | 0.1 | 3.5 | 0.03 | 0.11 | gains on 8 of 12 comparisons |
| Math 8 | 0.7 | 2.6 | 0.03 | 0.06 | |
| Math 11 | 1.7 | 0.6 | 0.03 | 0.02 | |
| **Washington 1999-2006** | | | | | |
| Reading 4 | 2.2 | 3.8 | 0.05 | 0.13 | |
| Reading 7 | 1.3 | 4.1 | 0.03 | 0.22 | |
| Reading 10 | 2.6 | 5.6 | 0.06 | 0.11 | Post-NCLB gains exceed pre-NCLB |
| Math 4 | 4.8 | 1.8 | 0.14 | 0.08 | gains on 9 of 12 comparisons |
| Math 7 | 2.1 | 4.5 | 0.06 | 0.17 | |
| Math 10 | 1.4 | 3.4 | *0.05* | *0.05* | |
| **Wyoming 1999-2005** | | | | | |
| Reading 4 | 0.0 | 1.0 | | | |
| Reading 8 | -0.7 | 0.3 | | | |
| Reading 11 | -.03 | 1.7 | No effect size | | Post-NCLB gains exceed pre-NCLB |
| Math 4 | -0.7 | 2.0 | data available | | gains on 5 of 6 comparisons |
| Math 8 | 1.0 | 1.7 | | | |
| Math 11 | 2.7 | 2.7 | | | |

Table reads: Kansas, which had six years of comparable percentage proficient and effect size data (2000-2005), had a pre-NCLB average yearly decline in grade 5 reading of 0.1 percentage points and a post-NCLB average yearly gain of 4.9 percentage points. In terms of effect size, Kansas had a pre-NCLB average yearly gain in grade 5 reading of 0.00 standard deviation units and a post-NCLB average yearly gain of 0.14 standard deviation units (14% of a standard deviation). Post-NCLB gains exceeded pre-NCLB gains on all 12 points of comparison in Kansas.

Note: Italics signify that effect sizes showed a different trend than percentages proficient.

[1] Pennsylvania had percentages proficient for 2001-2006 and effect sizes for 1999-2006.

## STATES WITH GREATER PRE-NCLB GAINS

Of the 13 states with pre- and post-NCLB achievement data, 4 states, while not showing overall declines, experienced slower rates of increase by most indicators after NCLB took effect. These states include Delaware, Massachusetts, Oregon, and Virginia.

**Table 10** is similar to table 9, except that it compares pre- and post-NCLB average yearly gains in states that had greater gains before 2002 than after.

## Table 10. Average Yearly Gains by Subject and Grade for States with Greater Overall Gains Before NCLB Than After

| State, Years of Data, Subject, Grade Level | Average Yearly Percentage Point Gain | | Average Yearly Effect Size Gain | | Comparisons of Pre-NCLB v. Post-NCLB Gains |
|---|---|---|---|---|---|
| | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | |
| **Delaware**[1] | | | | | |
| Reading 3 | 3.6 | 1.7 | N/A | N/A | |
| Reading 4/5[2] | 5.1 | 2.4 | 0.13 | 0.05 | |
| Reading 8 | 3.1 | 2.3 | 0.09 | 0.04 | Pre-NCLB gains exceed |
| Reading 10 | 4.2 | 1.2 | 0.08 | 0.04 | post-NCLB gains |
| Math 3 | 2.8 | 2.3 | N/A | N/A | on 14 of 14 comparisons |
| Math 4/5[2] | 3.9 | 3.2 | 0.10 | 0.08 | |
| Math 8 | 4.1 | 1.6 | 0.10 | 0.05 | |
| Math 10 | 4.2 | 3.0 | 0.11 | 0.08 | |
| **Massachusetts (1999-2006)** | | | | | |
| Reading 10 | 8.5 | 2.6 | | | Pre-NCLB gains exceed |
| Math 4 | 1.1 | 0.2 | No effect size | | post-NCLB gains |
| Math 8 | 1.9 | 1.6 | data available | | on 4 of 4 comparisons |
| Math 10 | 6.6 | 5.8 | | | |
| **Oregon (1999-2006)** | | | | | |
| Reading 3 | 1.3 | 0.5 | | | |
| Reading 5 | 3.3 | 1.0 | | | |
| Reading 8 | 2.7 | 0.4 | | | Pre-NCLB gains exceed |
| Reading 10 | 0.3 | 0.5 | No effect size | | post-NCLB gains |
| Math 3 | 2.3 | 2.3 | data available | | on 5 of 8 comparisons |
| Math 5 | 3.0 | 2.4 | | | |
| Math 8 | 1.7 | 2.1 | | | |
| Math 10 | 3.0 | 0.0 | | | |
| **Virginia (1999-2006)** | | | | | |
| Reading 3 | 3.7 | 3.0 | | | |
| Reading 5 | 3.0 | 2.3 | | | Pre-NCLB gains exceed |
| Reading 8 | 0.7 | 2.3 | No effect size | | post-NCLB gains |
| Reading 11 | 3.7 | 1.0 | data available | | on 6 of 7 comparisons |
| Math 3 | 4.0 | 2.5 | | | |
| Math 5 | 6.7 | 3.0 | | | |
| Math 8 | 3.7 | 1.3 | | | |

Table reads: Delaware, which had seven years of comparable percentage proficient data (1999-2005) and eight years of comparable effect size data (1999-2006), had a pre-NCLB average yearly gain in grade 5 reading of 5.1 percentage points and a post-NCLB average yearly gain of 2.4 percentage points. In terms of effect size, Delaware had a pre-NCLB average yearly gain in grade 4 reading of 0.13 standard deviation units (13% of a standard deviation) and a post-NCLB average yearly gain of 0.05 standard deviation units (5% of a standard deviation). Pre-NCLB gains exceeded post-NCLB gains on all 14 points of comparison in Delaware.

[1] Delaware had percentages proficient for 1999-2005 and effect sizes for 1999-2006.

[2] Percentage proficient data are for grade 5, and effect size data are for grade 4.

### MAIN FINDINGS ABOUT PRE- AND POST-NCLB COMPARISONS

Nine of the 13 states with pre- and post-NCLB achievement data made greater average yearly gains after NCLB took effect than before, by most indicators. The other four states showed slower progress after NCLB took effect, by most indicators. It is difficult to say, however, whether the small sample of 13 states represents a true national trend of hastening progress after NCLB. For now, these comparisons should be taken as suggestive.

In most states with complete data, the effect size and percentage proficient comparisons of pre- and post-NCLB trends were consistent. In a few instances, however, effect size data contradicted the percentage proficient data. In Kentucky for grade 11 math and in Louisiana for grades 4 and 8 reading, the percentage proficient results showed greater average yearly gains after NCLB than before, but effect size results at those grades showed greater gains before NCLB than after. In New Jersey, gains in percentages proficient for grade 8 reading were higher after NCLB than before, but effect sizes showed average yearly declines that were larger after NCLB than before. Clearly, one would have more confidence in concluding that gains have been greater after NCLB in a state like Kansas, where 12 of 12 indicators show the same general trend (table 9), than in a state like Louisiana, which showed a mixed pattern.

### POSSIBLE EXPLANATIONS FOR PRE- AND POST-NCLB TRENDS

Possible explanations for higher post-NCLB gains in the nine states are the same as those given in the achievement trends above—greater student learning, more intensive teaching to the test, and changes in test administration, or a combination of these factors.

Changes in the population of students tested may be a particularly important factor for comparisons between pre- and post-NCLB trends. The trends could be skewed in favor of pre-NCLB results because of changes in the number of students included in testing. NCLB requires that 95% of students be tested, including students with disabilities and limited-English-proficient students. Before NCLB, fewer than 95% of students might have been tested. In particular, these two subgroups with special needs may have been left out of testing more often, along with students with attendance problems. This may help explain why four states showed greater gains before NCLB.

Other possible explanations for these trends are being investigated more fully in phase II of this study.

## State-by-State Summary of Overall Achievement Trends

**Table 11** summarizes our findings about overall achievement for each state, displaying the trends in tables 1, 2, 5, and 6 of this chapter in more detail. This table is not intended to support comparisons between states; each state has its own standards, testing systems, and proficiency definitions so comparing one state with another is not advisable or informative.

**Table 11. State-by-State Summary of Achievement Trends Since 2002**

| State | Reading Elementary PP | Reading Elementary ES | Reading Middle School PP | Reading Middle School ES | Reading High School PP | Reading High School ES | Math Elementary PP | Math Elementary ES | Math Middle School PP | Math Middle School ES | Math High School PP | Math High School ES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | ↑ | ○ | ↑ | ○ | ↘ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Alaska | ♦ | ○ | ♦ | ○ | ↘ | ○ | ♦ | ○ | ♦ | ○ | ↑ | ○ |
| Arizona | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ |
| Arkansas | ♦ | ♦ | ♦ | ♦ | ↑ | ✓ | ♦ | ♦ | ♦ | ♦ | ↑ | ✓ |
| California | ↑ | ○ | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Colorado | ↑ | ○ | ↗ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ | ↗ | ○ |
| Connecticut | ↘ | ✓ | ↓ | ✓ | ↗ | ✗ | ↘ | ✓ | ↘ | ✓ | ↗ | ✓ |
| Delaware | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Florida | ↑ | ○ | ↗ | ○ | ↓ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Georgia | ↑ | ✓ | ↑ | ✓ | ○ | ○ | ↑ | ✓ | ↑ | ✓ | ↗ | ○ |
| Hawaii | ↗ | ✗ | ↘ | ✓ | ↗ | ✗ | ↗ | ✓ | ↑ | ✓ | ↘ | ✗ |
| Idaho | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↘ | ✓ |
| Illinois | ↑ | ○ | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↗ | ○ | ↘ | ○ |
| Indiana | ↗ | ✓ | ↗ | ✓ | ↓ | ✓ | ↑ | ✓ | ↑ | ✓ | ↗ | ✓ |
| Iowa | ↗ | ✓ | ↗ | ✓ | ↗ | ✓ | ↑ | ✓ | ↑ | ✓ | ↘ | ✗ |
| Kansas | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Kentucky | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Louisiana | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Maine | ↑ | ○ | ↘ | ○ | ↓ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Maryland | ↑ | ○ | ↑ | ○ | ♦ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Massachusetts | ↘ | ○ | ↗ | ○ | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ |
| Michigan | ↑ | ✓ | ↑ | ✓ | ○ | ○ | ↑ | ✓ | ↑ | ✓ | ○ | ○ |
| Minnesota | ↑ | ✓ | ♦ | ♦ | ♦ | ♦ | ↑ | ✓ | ♦ | ♦ | ♦ | ♦ |
| Mississippi | ↑ | ✓ | ↑ | ✓ | ↗ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Missouri | ↘ | ✓ | ↗ | ✗ | ↘ | ✗ | ↑ | ✓ | ↗ | ✗ | ↑ | ✗ |
| Montana | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Nebraska | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Nevada | ↓ | ✓ | ↘ | ✗ | ↗ | ✓ | ↑ | ✓ | ↗ | ✓ | ↓ | ✗ |
| New Hampshire | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| New Jersey | ↗ | ✗ | ↗ | ✗ | ↗ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| New Mexico | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |

| State | Reading Elementary PP | Reading Elementary ES | Reading Middle School PP | Reading Middle School ES | Reading High School PP | Reading High School ES | Math Elementary PP | Math Elementary ES | Math Middle School PP | Math Middle School ES | Math High School PP | Math High School ES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New York | ↑ | ○ | ↑ | ○ | ○ | ○ | ↑ | ○ | ↑ | ○ | ○ | ○ |
| North Carolina | ↗ | ✓ | ↗ | ✗ | ♦ | ♦ | ↑ | ✓ | ↗ | ✓ | ♦ | ♦ |
| North Dakota | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Ohio | ↓ | ○ | ♦ | ○ | ↑ | ○ | ♦ | ○ | ♦ | ○ | ↑ | ○ |
| Oklahoma | ↑ | ✗ | ↘ | ✓ | ↑ | ✗ | ↑ | ✓ | ↗ | ✗ | ↑ | ✓ |
| Oregon | ↑ | ♦ | ↗ | ♦ | ↗ | ♦ | ↑ | ♦ | ↑ | ♦ | ↘ | ♦ |
| Pennsylvania | ↗ | ✗ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↗ | ✓ |
| Rhode Island | ↑ | ○ | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ | ↘ | ○ |
| South Carolina | ↑ | ✓ | ↘ | ✗ | ↓ | ✓ | ↑ | ✓ | ↗ | ✓ | ↓ | ✓ |
| South Dakota | ♦ | ○ | ♦ | ○ | ♦ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Tennessee | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↗ | ✓ |
| Texas | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ |
| Utah | ↑ | ✓ | ↗ | ✓ | ↘ | ✗ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Vermont | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ | ♦ | ○ |
| Virginia | ↑ | ○ | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |
| Washington | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| West Virginia | ↑ | ✓ | ↗ | ✓ | ↘ | ✓ | ↑ | ✓ | ↑ | ✓ | ↑ | ✓ |
| Wisconsin | ↗ | ✗ | ↗ | ✓ | ↑ | ✓ | ↘ | ✗ | ↘ | ✓ | ↑ | ✓ |
| Wyoming | ↑ | ○ | ↗ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ | ↑ | ○ |

Table reads: In Connecticut, the percentage of students scoring at the proficient level or above on state reading tests has decreased slightly since 2002 at the elementary grade analyzed for this report. When measured in terms of effect size, achievement in elementary-level reading has also decreased during that period, confirming the percentage proficient trend. In high school reading, however, percentages proficient increased slightly but effect sizes declined, contradicting the percentage proficient trend.

**Legend**

PP = Percentage proficient

ES = Effect size

↑ = Moderate-to-large gain

↓ = Moderate-to-large decline

↗ = Slight gain

↘ = Slight decline

♦ = Not enough years of data (only 1-2 years) to determine trend

○ = Data not available

✓ = Effect size confirms percentage proficient trend

✗ = Effect size contradicts percentage proficient trend

# Chapter 5
# Trends in Achievement Gaps



## Key Findings

- Among the states with sufficient data to discern trends by subgroup, the number of states in which gaps in percentages proficient have narrowed since 2002 far exceeds the number of states in which gaps widened.
- For the African-American subgroup, 14 of the 38 states with the necessary data showed evidence that gaps have narrowed in reading across all three grade spans analyzed, while no state had evidence that gaps have widened. In mathematics, 12 states showed these gaps narrowing, while only one state showed the gaps widening. Results were similar for the Hispanic and low-income subgroups.
- As with the percentage proficient, the states in which effect size gaps have narrowed outnumbered the states in which the effect size gaps have widened. However, for states with both types of data, there were a number of instances where gap closings in terms of percentages proficient were not confirmed by effect size. Effect sizes seem to give a less rosy picture of achievement gap trends.
- Even for subgroups that showed evidence of gaps narrowing, the gaps in percentages proficient were often sizeable, suggesting that it will take a concerted, long-term effort to close them.
- Data on achievement gap trends for students with disabilities and limited-English-proficient students are not reliable enough to support solid conclusions because policies affecting how these students are tested and how their scores are counted as proficient have changed since 2002.

## How We Analyzed Achievement Gap Trends

Closing achievement gaps between subgroups of students—such as African American and white students, Hispanic and white students, or low-income and higher-income students—is a second major goal of No Child Left Behind. Therefore, the second major research question of this study focuses on whether achievement gaps between subgroups have narrowed since NCLB was enacted. This chapter reports our findings in response to this question.

We analyzed achievement gaps in terms of percentages proficient; where available, we used effect sizes as a second indicator of changes in gaps. Effect size analyses are especially helpful

when looking at gap trends. As explained in box A of chapter 3, the size of the gap in percentages proficient between a lower-achieving and a higher-achieving subgroup can look different depending on where a state sets its proficiency cut score on the score scale. Effect size analyses avoid this problem by focusing on mean test scores instead of percentages proficient.

Our analyses of achievement gaps did not include a comparison of pre- and post-NCLB trends, as they did with overall achievement trends. Few states produced and publicly reported test scores disaggregated by subgroup before NCLB.

To arrive at the findings in this chapter, we drew from very detailed data profiles for each of the 50 states, which can be accessed and downloaded from www.cep-dc.org/pubs/stateassessment/. The tables in these profiles include a multitude of data on gaps in percentages proficient and effect sizes, where available, for every subgroup with sufficient data. For each such subgroup, the profiles display data at one grade per grade span (elementary, middle, and high school); the specific grades were chosen using consistent rules laid out in chapter 3. The profiles show the extent of the gaps between 2002 and 2006. Not all of the states had data for the entire post-NCLB period; in some states, comparable data disaggregated by subgroup were available only for shorter periods, such as 2003 through 2006 or 2002 through 2005. The profiles also show changes in the gap between the starting and ending years, as well as the average yearly reductions in the gap.





Using the data in the profiles, we closely analyzed gap trends in reading and math for every state, a painstaking task in light of the amount of data that had to be reviewed. Our criteria for determining gap trends at the state level were stringent: we looked for evidence that gaps were indeed changing within a state by weeding out states with inconsistent results across grades. We counted an achievement gap as narrowing or widening only when trends in the same subject were consistent across *all three grade spans* in a given state.

Specific subgroups were included in the gap analysis for a particular state only when a state had three or more years of comparable test data for that group. In addition, subgroups that were small or had changed significantly in size were not counted in the national summary tables in this chapter for reasons explained below. We also omitted Asian students, students with disabilities, and limited-English-proficient students from the national summary for reasons discussed below.

We coded and compiled our findings from the 50 states to produce the tables in this chapter and develop a national picture of gap trends. More detailed information about study methods appears in the appendix.

## Gaps in Percentages Proficient Since 2002

**Table 12** shows the states with sufficient evidence that achievement gaps in percentages proficient have narrowed or widened for specific subgroups since 2002.

### AVAILABILITY OF GAP TREND DATA

The total number of states with sufficient data to analyze gaps in percentages proficient varied by subgroup and subject. As displayed in table 12, 31 states had sufficient data to analyze gaps in reading between low-income and non-low-income students, while 41 states had sufficient data to determine gaps in math between Hispanic and white students.

**Table 12. Number of States in Which Gaps in Percentages Proficient Narrowed or Widened Across Three Grade Spans Since 2002**

| Gap Trends | Black/ white reading | Black/ white math | Hispanic/ white reading | Hispanic/ white math | Am. Indian/ white reading | Am. Indian/ white math | Low-income/ not low-income reading | Low-income/ not low-income math |
|---|---|---|---|---|---|---|---|---|
| Gaps narrowed | 14 | 12 | 13 | 11 | 2 | 2 | 15 | 13 |
| Gaps widened | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| Total number of states with sufficient data[1] | 38 | 38 | 40 | 41 | 38 | 38 | 31 | 29 |

Table reads: Of the 38 states with the necessary data, the gap in percentages proficient between African American and white students has narrowed in reading since 2002 across all grade levels analyzed in 14 states; it has not widened across all grade levels analyzed in any state.

[1] States with sufficient data have three or more years of comparable test data since 2002.





If a state was not included in table 12 in the tallies of gaps that narrowed or widened, it was most likely because the narrowing or widening did not occur across all three grade spans—for example, the gaps for a subgroup may have narrowed at the elementary and middle levels but widened at high school. In this situation, we considered that state to have mixed results for that subgroup. Also not included in the counts of gaps that narrowed or widened were states that had subgroups that were too small to count or that changed significantly in size, as discussed below.

### NARROWING GAPS IN PERCENTAGES PROFICIENT

Based on percentages proficient, many more states showed evidence that gaps have narrowed rather than widened since 2002. For example, the first column of table 12 indicates that in 14 states, gaps in percentages proficient between African American and white students have narrowed in reading at the elementary, middle, and high school levels; no state had reading gaps that widened at all three grade spans.

The results for the African American and Hispanic subgroups turned out to be similar. For the Hispanic subgroup, 13 of the 40 states with sufficient data had evidence of narrowing gaps at all three grades levels in reading, and no state had evidence of widening gaps. In math, 11 of the 41 states with sufficient data showed gaps narrowing for the Hispanic subgroup, while none showed gaps widening.

The gaps between low-income students and other students also narrowed in many states. In reading, gaps for this subgroup narrowed in 15 of the 31 states with sufficient data; in math, gaps for this subgroup narrowed in 13 of the 29 states. Only one state showed gaps widening across grade levels for the low-income subgroup.

Additional trends in achievement gaps for specific subgroups within individual states are discussed in the state profiles available online.

### MAGNITUDE OF PERCENTAGE PROFICIENT GAPS

As we analyzed these data for all the states, one more striking pattern emerged. Even though achievement gaps have narrowed, the gaps for major subgroups remain wide in many states. Gaps of 20 to 30 percentage points are not uncommon, and some states, such as Pennsylvania and Connecticut, show gaps of 30 to 40 percentage points. The magnitude of the gaps suggests that closing achievement gaps is a long-term proposition that will require continued and intensive efforts.

## Effect Size as a Second Indicator

**Table 13** was drawn from the group of 30 states that made available both percentage proficient and effect size data for subgroups. For some of these states, data for a certain grade level might have been missing or did not cover an adequate number of years, so the total number of states with data to analyze gap trends for each subgroup is shown at the bottom of each column.

### EFFECT SIZE TRENDS

Using both measures—effect sizes and percentages proficient—we found that since 2002, many more states showed narrowing gaps at all three grade spans than widening gaps. While numerous states showed gaps narrowing using one or both of these measures, no more than one state (though not always the same state) showed gaps getting wider for any subgroup in either subject. For instance, for the black/white reading gap, the states showed the gap narrowing across all three grade spans according to one or both of the measures, but no states showed these gaps widening (table 13).

Another notable result from table 13 is that for states that showed narrowing gaps, the two measures of percentages proficient and effect size were not entirely in agreement. There were cases where the narrowing was revealed by only one of the two measures. For example, in the black/white mathematics gap column, two states showed gaps narrowing in terms of both percentages proficient and effect size, and an additional four states showed this gap narrowing in terms of percentages proficient only. Therefore, a total of six states with both types of data showed this gap narrowing in terms of percentages proficient, but only two states showed this gap narrowing in terms of effect size. Similar discrepancies exist for the Hispanic and low-income groups. For low-income reading, four states showed gaps narrowing in terms of both measures. Another seven states showed gaps narrowing in terms of percentages proficient only, but only two states showed gaps narrowing according to effect size only.

More states showed gaps narrowing when the percentage proficient measure was used. This result hints at the possibility that if effect size measures, based on mean test scores, were to be used in all 50 states to measure achievement gaps, the gap results might not be as positive as those using percentage proficient. It is certainly a topic for further research.

**Table 13. Number of States in Which Gaps in Percentages Proficient or Effect Sizes Narrowed or Widened Across Three Grade Spans Since 2002**

| Gap Trends | Black/ white reading | Black/ white math | Hispanic/ white reading | Hispanic/ white math | Am. Indian/ white reading | Am. Indian/ white math | Low-income/ not low-income reading | Low-income/ not low-income math |
|---|---|---|---|---|---|---|---|---|
| Narrowed for both percentage proficient and effect size | 5 | 2 | 3 | 3 | | | 4 | 4 |
| Narrowed for percentage proficient only, not confirmed by effect size | 3 | 4 | 3 | 1 | | | 7 | 4 |
| Narrowed for effect size only, not confirmed by percentage proficient | 2 | | | | | | 2 | 3 |
| Widened for both percentage proficient and effect size | | | | | | | | |
| Widened for percentage proficient only, not confirmed by effect size | | 1 | | 1 | | | 1 | 1 |
| Widened for effect size only, not confirmed by percentage proficient | | | 1 | | | | | |
| Total number of states with sufficient data[1] | 24 | 25 | 25 | 26 | 24 | 24 | 21 | 21 |

Table reads: In reading, the gaps in both percentages proficient and effect sizes between African American and white students have narrowed since 2002 in five states at all three grade spans analyzed. In three states, the gaps between African-American and white students narrowed at all three grade spans in terms of percentages proficient only; in two states, the gap narrowed in terms of effect size only. Twenty-four states had the data necessary to analyze both percentage proficient and effect size gaps in reading for African American and white students.

[1] States with sufficient data have three or more years of comparable test data since 2002.

### MAGNITUDE OF EFFECT SIZE GAPS

The magnitude of the gaps in effect size varied among states and subgroups but often fell within a range of 0.30 and 1.00 standard deviation. For example, the effect size gaps between African American and white students, and between Hispanic and white students, were generally smaller than 1.0 standard deviation, usually between 0.30 and 0.99. But is a gap of 0.99 standard deviation units a large one? Peterson and West (2003) noted that the gap in test scores between African American and white students nationally is about one standard deviation, which they call "very large." It is the "performance difference between typical fourth and eighth graders;" it is also the size of the gap in math between U.S. students and their higher-scoring Japanese counterparts (Peterson & West, 2003, p. 4). On a normal curve, such as the one illustrated in box C of chapter 3, a difference of one standard deviation is equivalent to the difference between 50% of students in one group scoring proficient versus 16% of another group scoring proficient.

## Tradeoffs between Overall Gains and Gap Reduction

We were curious about whether there was a tradeoff between overall achievement gains in a state and narrowing of gaps. That is, were states that showed moderate-to-large achievement gains less likely to show gaps narrowing? And were states that showed gaps narrowing less likely to show overall gains in student achievement? We cross-tabulated achievement trends with gap changes by state and did not find any evidence of such a tradeoff. States that made overall achievement gains also tended to show gaps narrowing. This result is partly explained by our findings that few states showed overall declines in achievement or widening of gaps.





## Possible Explanations for Gap Trends

Why have percentage proficient gaps in both reading and math narrowed more often than they have widened since 2002? A primary purpose of NCLB is to highlight and address differences in the achievement of student subgroups. The law has forced states, districts, and schools to disaggregate test scores by subgroup and report them to the public. This, in turn, has raised educators' awareness of the need to better serve historically low-performing subgroups and increased pressure on educators to do so. According to CEP's case studies of school district implementation of NCLB, disaggregation of data has also encouraged some districts to develop special programs or devote more resources to raising achievement for lower-performing subgroups (CEP, 2006).

Interviewees in the majority of our case study districts agreed that high expectations for all students and greater attention to subgroup achievement were among NCLB's most positive aspects. For example, officials in a California district with a Hispanic majority said that data collected to meet NCLB requirements highlighted the underperformance of some subgroups of students and have spurred staff to rethink teaching strategies and redirect resources (CEP, 2006). But it is difficult to determine to what extent achievement gains or gap reductions are a result of real, generalized learning gains rather than overly focused test preparation (the "teaching to the test" described in chapter 4). CEP's case studies of the impact of state exit exams in Austin, Texas, and Jackson, Mississippi, may shed some light on this issue (CEP, 2007b). Teachers and students in these two urban districts reported that test preparation strategies were often used during class time and that test preparation was detracting from other potentially valuable instructional experiences. Moreover, in Austin schools with higher

enrollments of minority or low-income students, state tests had a stronger influence on curriculum; instruction in some subjects was shortened to make time to teach and review material likely to be tested. Still, our NCLB case studies suggest that in some districts, students from underperforming subgroups are receiving more intensive instruction and seem to be learning more (CEP, 2006).

We do know that urban districts are most affected by NCLB. For school year 2005-06, 47% of urban districts reported having at least one school identified for improvement under NCLB, compared with 22% of suburban districts and 11% of rural districts (CEP, 2007a). Urban districts have more subgroups large enough to count toward AYP because they serve higher percentages of minority and disadvantaged students than their suburban or rural counterparts. Therefore, urban districts are under the most pressure and are likely to be reallocating resources to improve their test results.

## Other Subgroups

As mentioned above, we did not include achievement gap data for the following subgroups in the summary tables in this chapter, but we did discuss trends for these subgroups in the individual state profiles, where appropriate.

### STUDENTS WITH DISABILITIES AND LEP STUDENTS

As described in box A of chapter 3, data for students with disabilities and limited-English-proficient students subgroups must be interpreted with caution because changes in federal regulations and guidance and in state accountability plans may have affected which students in these subgroups are tested for NCLB accountability purposes, how they are tested, and when their test scores are counted as proficient under NCLB. All of these factors affect the comparability of the results for these subgroups over time and make trend analyses inadvisable. While we describe the state-by-state findings for these subgroups in the state profiles (with the necessary qualifiers), we do not believe the data are reliable enough to be included in the national summary tables in this chapter.

### SUBGROUPS WITH SMALL OR CHANGING NUMBERS OF STUDENTS TESTED

We also excluded subgroups from the national summary tables if they were small, defined as less than 5% of the state's total population of test-takers in a given grade or fewer than 500 students. We adopted this policy because test scores fluctuate more for small subgroups, so results may not be reliable and should be interpreted with caution. We also excluded subgroups in which the number of students tested changed by at least 25% in either direction over the years reported. That is because year-to-year changes in results may be due to changes in the composition of the subgroup rather than performance on tests, so the results may not be reliable. Subgroups that met these criteria across all grade spans were excluded from our national summary but included in the state reports. However, we did include subgroups in the national summary if they had at least one grade with adequate and relatively stable numbers over time.

Where data were available, we included tables with numbers of test-takers by subgroup at the end of each state profile. These are the data we used to determine which subgroups were small or had changed significantly in size over time.

### ASIAN SUBGROUP

Finally, we did not include the Asian subgroup in our national summary of achievement gaps, although the figures can be found in each state report. That is because Asian students preformed as well as or better than white students in most states, and we focused on subgroups for which achievement gaps have been a major educational hurdle. In the state profiles, we do note gap trends for the Asian subgroup if this subgroup had lower performance than white students and showed evidence of a trend across grades. Alaska and Hawaii are two such states.

## State-By-State Summary of Achievement Gap Trends

**Table 14** summarizes our findings about achievement gap trends for each state. This table is not intended to support comparisons between states; each state has its own standards, testing systems, proficiency definitions, and list of major subgroups, so comparing one state with another is not advisable or informative.

## Table 14. State-by-State Summary of Gap Trends (Percentage Proficient)

| State | Black/White | | Hispanic/White | | American Indian/White | | Low-Income/ Not Low-Income | |
|---|---|---|---|---|---|---|---|---|
| | Reading | Math | Reading | Math | Reading | Math | Reading | Math |
| Alabama | ✱ | ➔➔[2] | ✱ | ✱ | ✱ | ✱ | ✱ | ➔➔[2] |
| Alaska | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Arizona | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Arkansas | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| California | ✱ | ✱[1] | ✱ | ➔➔ | ▲ | ▲ | ✱ | ➔➔[1] |
| Colorado | ✱ | ➔➔ | ➔➔ | ➔➔ | ▲ | ▲ | ❍ | ❍ |
| Connecticut | ✱[1] | ✱[1] | ➔➔[1] | ➔➔[1] | ▲ | ▲ | ✱[1] | ✱[1] |
| Delaware | ➔➔[1] | ➔➔[1] | ▲ | ▲ | ▲ | ▲ | ➔➔[1] | ✱[1] |
| Florida | ➔➔ | ➔➔[1] | ➔➔ | ➔➔[1] | ▲ | ▲ | ✱[1] | ➔➔[1] |
| Georgia | ➔➔[2] | ➔➔[1] | ▲ | ▲ | ▲ | ▲ | ❍ | ❍ |
| Hawaii | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ✱[1] | ➔➔[1] |
| Idaho | ▲ | ▲ | ➔➔ | ➔➔ | ▲ | ▲ | ➔➔ | ✱ |
| Illinois | ✱ | ✱ | ➔➔ | ➔➔ | ✱ | ✱ | ✱ | ✱ |
| Indiana | ✱ | ✱ | ▲ | ▲ | ▲ | ▲ | ✱[1] | ✱[1] |
| Iowa | ▲ | ▲ | ➔➔[1] | ➔➔[1] | ▲ | ▲ | ➔➔ | ✱ |
| Kansas | ➔➔[1] | ✱[1] | ➔➔[1] | ▲ | ▲ | ▲ | ➔➔[1] | ✱[1] |
| Kentucky | ✱ | ✱[1] | ▲ | ▲ | ❍ | ❍ | ✱[1] | ✱[1] |
| Louisiana | ➔➔ | ➔➔ | ▲ | ▲ | ▲ | ▲ | ➔➔ | ➔➔[1] |
| Maine | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ❍ | ❍ |
| Maryland | ➔➔ | ✱[1] | ➔➔ | ✱[1] | ▲ | ▲ | ➔➔ | ✱[1] |
| Massachusetts | ➔➔ | ✱ | ➔➔ | ✱ | ✱ | ➔➔ | ❍ | ❍ |
| Michigan | ➔➔[2] | ➔➔[2] | ▲ | ▲ | ❍ | ❍ | ➔➔[2] | ➔➔[1,2] |
| Minnesota | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ➔➔ | ➔➔ |
| Mississippi | ✱ | ➔➔ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| Missouri | ✱ | ✱ | ▲ | ▲ | ▲ | ▲ | ↔ | ↔ |
| Montana | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Nebraska | ➔➔ | ➔➔ | ➔➔ | ➔➔ | ➔➔ | ✱ | ➔➔ | ➔➔ |
| Nevada | ➔➔ | ✱ | ➔➔[1] | ➔➔[1] | ▲ | ▲ | ✱[1,2] | ↔[1,2] |
| New Hampshire | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ✱ | ✱ |
| New Jersey | ✱ | ✱ | ✱[1] | ➔➔[1] | ▲ | ▲ | ✱ | ➔➔ |
| New Mexico | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |

| State | Black/White | | Hispanic/White | | American Indian/White | | Low-Income/ Not Low-Income | |
|---|---|---|---|---|---|---|---|---|
| | Reading | Math | Reading | Math | Reading | Math | Reading | Math |
| New York | →← [2] | →← [2] | →← [2] | →← [2] | ▲ | ▲ | →← [2] | →← [2] |
| North Carolina | →← | →← | ▲ | ▲ | ▲ | ▲ | →← | →← [1] |
| North Dakota | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Ohio | ✱ | ♦ | ▲ | ▲ | ▲ | ▲ | ✱ [1] | ♦ |
| Oklahoma | →← | ✱ | ✪ | ✱ | →← | ✱ | ✱ | ✱ [2] |
| Oregon | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ❍ | ❍ |
| Pennsylvania | →← [1] | →← [1] | ▲ | ▲ | ▲ | ▲ | →← [1] | →← [1] |
| Rhode Island | ✱ | ✱ | ✱ | ✱ | ✱ | ✱ | →← | →← |
| South Carolina | ✱ | ✱ | ▲ | ▲ | ▲ | ▲ | ✱ | ✱ |
| South Dakota | ♦ | →← | ♦ | →← | ♦ | →← | ♦ | ✱ |
| Tennessee | →← | →← | ▲ | ▲ | ▲ | ▲ | →← | →← [1] |
| Texas | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Utah | ▲ | ▲ | →← | ✱ | ▲ | ▲ | →← [1] | →← |
| Vermont | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ | ♦ |
| Virginia | →← | →← | ✱ | ✱ | ✱ | ❍ | ❍ | ❍ |
| Washington | →← [1] | ↔ [1] | →← [1] | ↔ [1] | ▲ | ▲ | ❍ | ❍ |
| West Virginia | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | →← | →← |
| Wisconsin | ✱ | ✱ | ✱ [1] | ✱ [1] | ▲ | ▲ | ✱ | ✱ |
| Wyoming | ✱ | ✱ | ✱ | ✱ | ❍ | ❍ | →← | ✱ |

Table reads: In Alabama, the gap between black and white students was mixed (gaps narrowed and widened at different grade spans) on the state reading tests. Data were not available for one of the three grade levels analyzed on the state math tests, but in the remaining two grade levels, the gap narrowed between black and white students.

Note: Trends for students with disabilities and limited-English-proficient students are not shown because changes in state and federal policies may have affected the year-to-year comparability of test results for these subgroups. Trends for Asian students are not shown because in most states this subgroup performed as well as or better than white students.

[1] Subgroup was small or changed significantly in size in one or two of the grades analyzed.

[2] Data were not available for one of the three grade levels analyzed but in the remaining two grade levels the gap narrowed, widened, or was mixed (as indicated by the symbol).

**Legend**

→← = Gap narrowed at three grade levels

↔ = Gap widened at three grade levels

✱ = Mixed (gaps narrowed and widened at different grade spans)

♦ = Not enough years of data (only 1-2 years) to determine trend

❍ = Data not available

▲ = Subgroup is small or changed significantly in size at all three grade levels analyzed

✪ = No change in gap

# Chapter 6
# Comparing State Test Score Trends with NAEP Results



## Key Findings

- Since 2002, many states with improved scores on state tests have shown declines or flat results on the National Assessment of Educational Progress. Results on state tests and NAEP diverged more often at the 8th grade level than at 4th grade. The most similar results were in grade 4 math, where almost all states showed gains on both assessments.

- Even when the percentage of students scoring at the *proficient* level on state tests is compared with the percentage scoring at the *basic* level on NAEP—a more equivalent comparison according to many analysts—states show more positive results on their own tests.

- Correlations between average yearly percentage point gains on state tests and gains on NAEP were low. That is, the states with the greatest gains on their own tests were usually not the same states that had the greatest gains on NAEP. The only significant correlation between state and NAEP results was a moderate relationship in grade 4 reading. It is possible that state tests and NAEP tests in grade 4 reading tend to be more closely aligned in content and format than tests in other grades and subjects.

- NAEP results should not be used as a "gold standard" to negate or invalidate state test results; instead they provide an additional point of information about achievement. While state tests are far from perfect, they are the best available standardized measures of the curriculum being taught in classrooms. NAEP provides a useful independent measure, but it also has limitations, such as lack of alignment with state standards, less motivation on the part of students, and a changing population of students tested.

## Background on NAEP

The National Assessment of Educational Progress, also known as "the Nation's Report Card," is the only national assessment of what U.S. students know and can do in various subject areas.[5] NAEP provides a measure of student achievement that is independent of state tests. NAEP differs from state tests—to varying degrees, depending on the state—in the content covered, types of test questions used, and rigor of the achievement levels set, such as the proficient level. NAEP also differs from state tests in that it is based on a matrix sampling design, whereby samples of students take only a portion of the larger assessment rather

[5] In this chapter we refer to the main NAEP assessment, which sometimes includes state-level results. NAEP also has a separate long-term trend assessment that has been in place for more years but is less relevant to this study. The two types of NAEP assessments are administered in different years.

than the entire test. NAEP is given to a representative sample of students in each state so that it will yield both national and state-level results by grade (such as 4th grade) and by subgroup (such as female students or Hispanic students). Because of its sampling design, however, NAEP cannot provide scores for individual students or schools.

## Recent NAEP Trends

Overall, NAEP trends since 2002 show a less positive picture of student achievement than the state test results reported in this study. The most recent NAEP reading and math assessments at grades 4 and 8—the elementary and middle school grades consistently tested by NAEP—were conducted in 2005, and the previous assessments were conducted in 2003. (Although NAEP tested reading and math at grade 12 in 2005, only nationwide results were reported, not the state results needed for this analysis.)

The 2005 NAEP results for grades 4 and 8 indicate that reading achievement has remained essentially flat since 2003. Between 2003 and 2005, the average national NAEP scores in reading rose just 1 point for 4th graders—to 219 on a 500-point scale—and dropped 1 point for 8th graders to 262. In fact, reading scores for both grades have stayed about the same since 1992, the earliest year that data comparable to the recent NAEP reading assessments were available (National Center for Education Statistics, 2006a; 2006b).





During that same period, average NAEP scores also increased in math but at a slower rate than in previous years. Math results for grades 4 and 8 climbed dramatically in the early 1990s, leveled off in the mid-1990s, and then rose again between 2000 and 2003. Scores increased more modestly between 2003 and 2005, when the 4th grade average score increased by 3 points to 238, and the 8th grade average score rose by 1 point to 279.

## How State Test Score Trends Compare with NAEP Trends

To what extent do the trends identified in this CEP study of state test results parallel trends in state-level NAEP results? Our study examined whether achievement has increased since NCLB was enacted in 2002, so to answer this question, we have compared our findings about state test score trends since 2002 with results from the NAEP administrations that most closely match our time frame—the two administrations for 4th and 8th graders in 2003 and 2005.

Our approach is a little different from that used in several other studies, which compared the percentages of students scoring at the proficient level on state tests with the percentage scoring proficient on NAEP tests (Achieve, n.d.; Education Week, 2006; Fuller et al., 2006; Lee, 2006; Peterson & Hess, 2006). Instead we compared the percentage of students reaching the proficient level on state tests with the percentage of students reaching the "basic" level of achievement on NAEP. We chose this approach because the research cited above has shown that the NAEP proficient level is more rigorous than most states' proficiency levels. The NAEP proficient level is even a tough hurdle internationally. The former U.S. Commissioner of Education Statistics, Gary Phillips, recently linked the NAEP scale to the scale of the Trends in International Mathematics and Science Study (TIMSS), an assessment that compares student achievement in 46 countries, including the U.S. He found that the average 8th grader in just five countries (Singapore, South Korea, Hong Kong, Taiwan, and Japan) could reach the equivalent of the NAEP proficient level on TIMSS (Phillips, 2007).

The process used to set NAEP achievement levels has been criticized (National Research Council, 1999; Pellegrino, 2007). Some researchers have suggested (Achieve, n.d.) that the "proficient" level on state tests is closer to the "basic" level on NAEP (one level below "proficient" in the NAEP hierarchy of achievement). U.S. Secretary of Education Margaret Spellings has also indicated that the proficient level on state tests is more appropriately compared to the basic level on NAEP (Dillon, 2005).

## COMPARISON OF STATES SHOWING GAINS ON STATE TESTS VS. NAEP

Tables 15 through 18 compare state-level trends in the percentage scoring proficient on state tests with trends in the percentage scoring basic on NAEP since 2002. They illustrate the extent to which results converge—which states showed gains on both sets of assessments and which had state results that were not supported by NAEP. Overall, we found that many states made gains on state tests but not on NAEP; there was more convergence in 4th grade than 8th grade. In the discussion that follows, we focus on the numbers of states showing moderate-to-large gains (or declines) on state tests because slight changes may simply reflect natural fluctuations in test scores due to measurement error.

### Elementary Level

**Table 15** compares the number of states showing gains and declines in the percentages scoring proficient on state elementary-level reading tests since 2002 with the number of states showing gains and declines in the percentages scoring basic on NAEP grade 4 reading tests between 2003 and 2005. (For the most part, the state test results are for grade 4, but when appropriate trend data were not available for grade 4, an adjacent grade was used, consistent with CEP's rules for analysis discussed in chapter 3.) The top row includes trends for all 29 states that showed moderate-to-large gains on their own elementary reading tests. The second column displays the 18 states (out of the preceding 29) that also showed gains on NAEP. Five states with gains on state tests showed flat results on NAEP, and six had contradictory results—they had gains on state tests but declines on NAEP.

The main NAEP reports express the percentage of students reaching the basic level or higher as a whole number. So in cases where the rounded percentages for 2003 and 2005 were identical, we calculated a difference of 0 and called it a "flat" trend, even though there may have been very small differences between the percentages before rounding.

For elementary reading, state test results are similar to NAEP results in that the number of states with gains exceeds the number with declines and a similar number of states show gains —29 states with moderate-to-large gains on state tests and 26 states with gains on NAEP. However, the two sets of states exhibiting gains did not fully overlap: of the 29 states with gains on state tests, only 18 also showed gains on NAEP, and 6 actually showed declines on NAEP. The NAEP results also revealed a greater proportion of states with declines—15 states with declines on NAEP, versus just 2 states with moderate-to-large declines on state tests.

For elementary math, the match between NAEP and state results appears to be better, as displayed in **table 16**. Here, of the 37 states that demonstrated gains on their state tests, 31 also demonstrated gains on NAEP. Five of these states showed flat scores on NAEP and only one had a decline. One obvious reason for the better match in elementary math is that very few states evidenced declines on either state tests or NAEP assessments.

## Table 15. Comparison of Trends on State Elementary Reading Tests and NAEP Grade 4 Reading Tests[1]

| Trend on State Test | NAEP Gain | NAEP Flat | NAEP Decline | Total of Row |
|---|---|---|---|---|
| Moderate-to-large gain | 18 states– AL, DE, FL, ID, IL, KY, LA, MD, ME, MN, NE, NY, TN, UT, VA, WA, WV, WY | 5 states– CA, CO, KS, OK, RI | 6 states– GA, MI, MS, NH, OR, SC | 29 |
| Slight gain | 1 state– PA | 1 state– HI | 5 states– IA, IN, NC, NJ, WI | 7 |
| Slight decline | 1 state– MA | | 2 states– CT, MO | 3 |
| Moderate-to-large decline | | 2 states– NV, OH | | 2 |
| Insufficient years for trend | 6 states– AR, MT, ND, NM, SD, TX | 1 state– AK | 2 states– AZ, VT | 9 |
| Total of column | 26 | 9 | 15 | 50 |

Table reads: Twenty-nine states have demonstrated moderate-to-large gains on state elementary (usually grade 4) reading tests since 2002. Of these, 18 states also showed gains on NAEP grade 4 reading tests between 2003 and 2005, while 5 states had flat results on NAEP and 6 states showed declines on NAEP.

[1] This table compares trends in the percentage of students scoring proficient on state elementary (usually grade 4) reading tests since 2002 with trends in the percentage scoring basic on NAEP grade 4 reading tests between 2003 and 2005.

**Table 16. Comparison of Trends on State Elementary Math Tests and NAEP Grade 4 Math Tests[1]**

| Change on State Test | Gain on NAEP | NAEP Flat | Decline on NAEP | Total |
|---|---|---|---|---|
| Moderate-to-large gain | 31 states– AL, CA, CO, DE, FL, GA, IA, ID, IL, IN, KS, KY, LA, MD, ME, MI, MN, MS, NH, NJ, NV, NY, OK, OR, PA, RI, SC, SD, TN, UT, WA | 5 states– MO, NE, VA, WV, WY | 1 state– NC | 37 |
| Slight gain | 2 states– HI, MA | | | 2 |
| Slight decline | 2 states– CT, WI | | | 2 |
| Moderate-to-large decline | | | | |
| Insufficient years for trend | 8 states– AK, AR, MT, ND, NM, OH, TX, VT | 1 state– AZ | | 9 |
| Total | 43 | 6 | 1 | 50 |

Table reads: Thirty-seven states have demonstrated moderate-to-large gains on state elementary (usually grade 4) math tests since 2002. Of these, 31 states also showed gains on NAEP grade 4 math tests between 2003 and 2005, while 5 states had flat results on NAEP, and 1 state showed a decline on NAEP.

[1] This table compares trends in the percentage of students scoring proficient on state elementary (usually grade 4) math tests since 2002 with trends in the percentage scoring basic on NAEP grade 4 math tests between 2003 and 2005.





### Middle School Level

The match between state test results and NAEP results is far weaker at the middle school level than at the elementary level. As **table 17** illustrates, there is little convergence between state test results for middle school reading (usually 8th grade) and NAEP results for 8th grade reading. Twenty states demonstrated moderate-to-large gains on their own tests, but only 12 demonstrated gains on NAEP, and the states with gains did not closely overlap. Of the 20 states with gains on state tests, only 5 had their gains confirmed by NAEP, while 11 of these states showed a decline on NAEP. Only one state showed a decline on its own middle school reading test while 27 showed a decline on NAEP.

The fit between results is a little better for 8th grade math, but the picture is still very murky, as **table 18** illustrates. Thirty-two states made gains on their own tests, and 26 made gains on NAEP. But again, the sets of states with gains do not match. Of the 32 states with gains on their own tests, only half showed confirming results on NAEP. Nine of these states demonstrated a decline on NAEP.

As tables 15 through 18 illustrate, trends in state test results are often different from trends on NAEP (more so for 8th grade than 4th). Many states have demonstrated improvements in state test results that are not supported by NAEP. Our findings confirm those of previous studies, which have generally found student achievement to be higher according to

**Table 17. Comparison of Trends on State Middle School Reading Tests and NAEP Grade 8 Reading Tests[1]**

| Change on State Test | Gain on NAEP | NAEP Flat | Decline on NAEP | Total |
|---|---|---|---|---|
| Moderate-to-large gain | 5 states– DE, KS, NE, PA, TN | 4 states– ID, LA, NY, RI | 11 states– AL, CA, GA, IL, KY, MD, MI, MS, NH, VA, WA | 20 |
| Slight gain | 3 states– MA, NJ, WY | 2 states– IA, WI | 8 states– CO, FL, IN, MO, NC, OR, WV, UT | 13 |
| Slight decline | 1 state– ME | 1 state– NV | 3 states– HI, OK, SC | 5 |
| Moderate-to-large decline | | | 1 state– CT | 1 |
| Insufficient years for trend | 3 states– AK, MN, ND | 4 states– MT, NM, OH, SD | 4 states– AR, AZ, TX, VT | 11 |
| Total | 12 | 11 | 27 | 50 |

Table reads: Twenty states have demonstrated moderate-to-large gains on state middle school (usually grade 8) reading tests since 2002. Of these, 5 states also showed gains on NAEP grade 8 reading tests between 2003 and 2005, while 4 states had flat results on NAEP, and 11 states showed declines on NAEP.

[1] This table compares trends in the percentage of students scoring proficient on state middle school (usually grade 8) reading tests since 2002 with trends in the percentage scoring basic on NAEP grade 8 reading tests between 2003 and 2005.





state tests. In this analysis we go one step further by comparing the percentage scoring *proficient* on state tests with the percentage scoring *basic* on NAEP. Therefore, fewer states have demonstrated progress on NAEP than have demonstrated progress on their own tests, even when the NAEP cut score is at the "basic" level.

## CORRELATIONAL ANALYSES

To further compare state test results and NAEP results we used a statistical test of correlation, which measures the extent to which two numeric variables are related (unlike the previous section, which compared gross categories of gains and declines). If the states with the greatest gains on state tests also tend to show the greatest gains on NAEP (and the same for declines), then the degree of correlation, as measured by a statistic called a correlation coefficient, will be positive. The higher the coefficient, the stronger the relationship.

We compared the average yearly increase in the percentage scoring at the proficient level on state tests since 2002 with the average yearly increase in the percentage scoring at the basic level on NAEP between 2003 and 2005. For each state, we computed the average yearly percentage point gains on the state test since 2002 using the method described in

**Table 18. Comparison of Trends on State Middle School Math Tests and NAEP Grade 8 Math Tests[1]**

| Change on State Test | Gain on NAEP | NAEP Flat | Decline on NAEP | Total |
|---|---|---|---|---|
| Moderate-to-large gain | 16 state – CA, DE, FL, GA, KS, LA, MA, MS, NE, NJ, OR, PA, SD, TN, VA, WA | 7 state– AL, HI, ID, IN, MI, NY, RI | 9 states– CO, IA, KY, MD, ME, NH, UT, WV, WY | 32 |
| Slight gain | 3 states– IL, NV, SC | 1 state– NC | 2 states– MO, OK | 6 |
| Slight decline | 1 state– WI | | 1 states– CT | 2 |
| Moderate-to-large decline | | | | |
| Insufficient years for trend | 6 states– AR, AZ, MT, NM, TX, VT | 2 states – ND, OH | 2 states– AK, MN | 10 |
| Total | 26 | 10 | 14 | 50 |

Table reads: Thirty-two states have demonstrated moderate-to-large gains on state middle school (usually grade 8) math tests since 2002. Of these, 16 states also showed gains on NAEP grade 8 math tests between 2003 and 2005, while 7 states had flat results on NAEP, and 9 states showed declines.

[1] This table compares trends in the percentage of students scoring proficient on state middle school (usually grade 8) math tests since 2002 with trends in the percentage scoring basic on NAEP grade 8 math tests between 2003 and 2005.





the appendix; the years covered between 2002 and 2006 varied by state, depending on which years yielded comparable test data. To calculate the average yearly percentage point gain on NAEP for each state, we subtracted the percentage scoring basic in 2003 from the percentage scoring basic in 2005, then divided by two years.

The correlations are shown in **table 19**. A correlation coefficient of +1 would indicate a perfect relationship between average annual increases on state tests and average annual increases on NAEP; that is, the states with the greatest gains on their own tests would also be the ones with the greatest gains on NAEP. The only statistically significant result—meaning a result not likely to be due to chance—is in grade 4 reading, which has a low-to-moderate correlation coefficient of .364. The correlations in elementary math and in middle school reading and math were positive but very low and not statistically significant. Figures 5 through 6 illustrate the results in tables 15 and 16 in the form of scatterplot graphs. Ideally, if a high correlation exists between the average yearly gains made on both the state and the NAEP tests, the states would cluster to form a rough line extending from the lower left-hand corner to the upper right-hand corner of each figure. For the sake of comparison, **figure 4** depicts an ideal scatterplot where the results from two tests match nearly perfectly, with a correlation coefficient of .997.

## Table 19. Correlations between Average Annual Gains in Percentage Proficient on State Tests (Post-NCLB) and in Percentage Basic on NAEP (2003-2005)

| Comparison | Correlation Coefficient (Statistical Significance)[1] |
|---|---|
| State elementary reading v. NAEP grade 4 reading | .364 *(.019)* |
| State middle school reading v. NAEP grade 8 reading | .165 *(.315)* |
| State elementary math v. NAEP grade 4 math | .012 *(.941)* |
| State middle school math v. NAEP grade 8 math | .125 *(.444)* |

Table reads: A correlation analysis comparing average yearly gains in the state percentage proficient since 2002 with average yearly gains in the NAEP percentage basic between 2003 and 2005 produced a correlation coefficient of 0.364 in elementary reading, which is statistically significant.

[1] A significance level of .05 means that there is a 5% probablility that this result would occur by chance.



## Figure 4. Sample Scatterplot of a Near-Perfect Match Between Two Numerical Variables

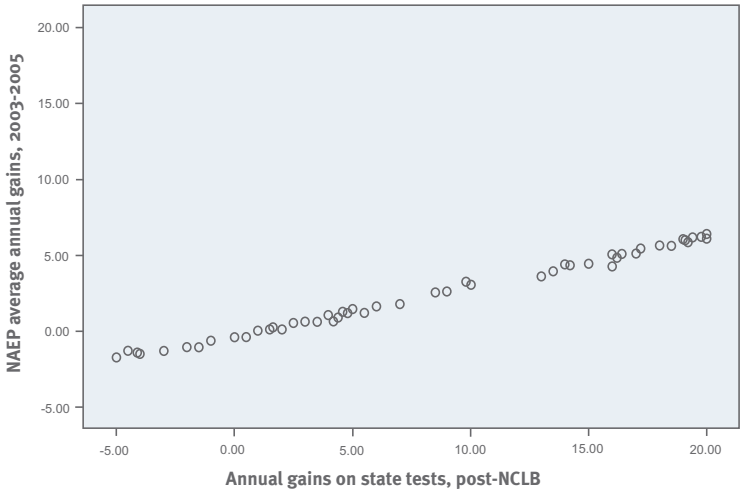


**Figure 5**, which graphs the results for elementary reading (the only statistically significant result), shows a loose cluster, but with many outliers. For example, the position of Massachusetts in the upper left hand corner indicates that it posted large gains on NAEP but a decline in state scores in elementary reading. In contrast, West Virginia, at the top of the figure, showed more consistent results, with relatively large gains on both its state test and NAEP.

**Figure 5. Comparison of Average Annual Gains on State Elementary Reading Tests with Average Annual Gains on NAEP Grade 4 Reading Tests[1]**

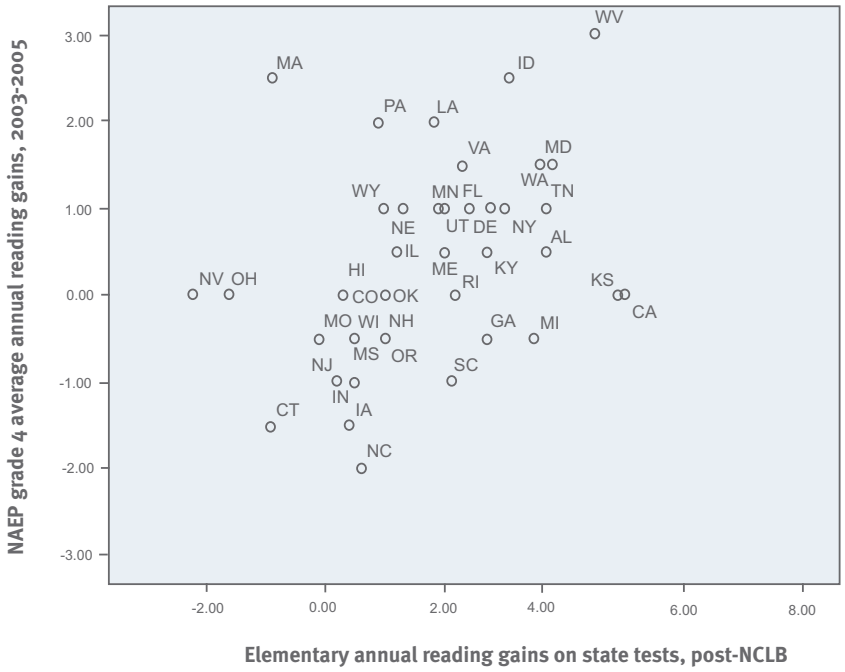



Figure reads: Massachusetts, in the upper left of the scatterplot, demonstrated average yearly declines of less than one percentage point on its state elementary reading test (horizontal axis) and average yearly gains of greater than two percentage points on the NAEP grade 4 reading assessment (vertical axis).

[1] This figure compares the percentage of students scoring proficient on state elementary level (usually grade 4) reading tests since 2002 with the percentage scoring basic on the NAEP grade 4 reading tests between 2003 and 2005.

**Figure 6**, which illustrates the results of the correlation analysis in middle school reading, does not exhibit any discernable trends; the pattern is essentially scattershot. No significant correlations were found between NAEP and state test gains. Though not shown here, the same was true of elementary and middle school math.

**Figure 6. Comparison of Gains on State Middle School Reading Tests with Gains on NAEP Grade 8 Reading Tests[1]**

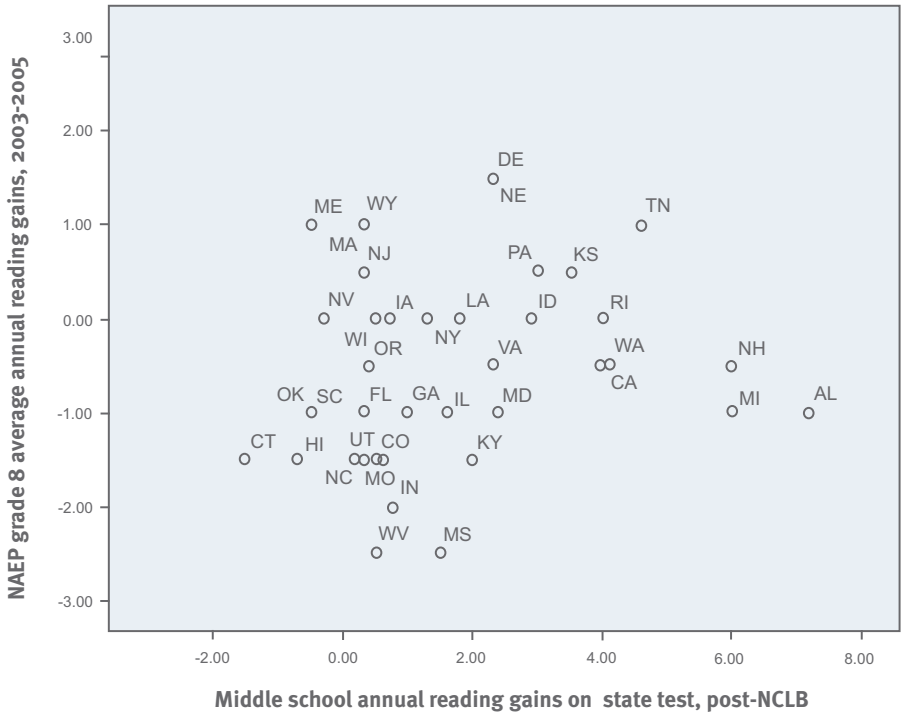



Figure reads: Rhode Island, near the center of the scatterplot, demonstrated average yearly gains of about four percentage points on its state middle school reading test (horizontal axis) and flat results on the NAEP grade 8 reading test (vertical axis).

[1] This table compares average yearly gains on state middle school reading tests (usually grade 8) since 2002 with average yearly gains on the NAEP grade 8 reading test between 2003 and 2005.

## Possible Explanations for Differences in State Test and NAEP Results

This analysis shows that there is not a lot of congruence between state test trends and NAEP trends since 2002, as measured by the percentages of students scoring proficient on state tests and scoring basic on NAEP. We found that many of the gains on state tests are not confirmed by NAEP. Furthermore, there was little correlation between the magnitude of gains and declines shown by the two types of assessments.

This is not to suggest that NAEP is a "gold standard" that should be used to negate or invalidate state test results. As noted in chapter 2, while state tests are far from perfect, they are the best available standardized measures of the curriculum being taught in classrooms. NAEP provides another independent measure of student achievement but also has limitations.

There are several possible reasons why state and NAEP results are not always consistent:

- ***Alignment to standards and curricula.*** State tests used for NCLB must be aligned to a set of state content standards. Ideally, there should be a match between the content of tests and what is taught in most classrooms. However, NAEP is not aligned to any particular state's standards and therefore may not be instructionally sensitive—in other words, it may not reflect what students are actually being taught (Popham, in press). For example, Jacob (2007) found that for Texas 4th graders, the difference in scores between NAEP and the state exam was largely explained by differences in the set of math skills (multiplication, division, fractions, decimals, etc.) covered by the two tests, and the format in which test items were presented (word problems, calculations without words, inclusion of a picture, etc.).

- ***Motivation.*** The state tests used for NCLB have high stakes for educators and somewhat high stakes for students. Federal and state sanctions or incentives for districts and schools are determined largely by the results of state tests. Teachers and administrators often go to great lengths to encourage students to take these tests seriously. In addition, individual scores are reported to students and parents. In some states, the tests used for NCLB have very high stakes for students because these tests are also used to determine whether students will be promoted to the next grade or graduate. NAEP, on the other hand, is a low-stakes test for schools and students because it is not connected to any rewards, sanctions, or future outcomes for administrators, teachers, or students—indeed, students do not even receive their individual NAEP results. For this reason, students may not be motivated to perform their best on NAEP, especially at the higher grade levels. The National Research Council noted the high rate of omitted and incomplete answers on NAEP high school tests, producing "scores of questionable value" (National Research Council, 1999. p. 80).

- ***Inclusion.*** The percentages of students with disabilities and limited-English-proficient students included in both state exams and NAEP have varied widely, raising questions about whether the NAEP sample reflects the same population assessed by state tests. As discussed in chapter 3, policies have changed over the years regarding how the test scores of these two subgroups are included in determinations of percentages proficient and adequate yearly progress. This is also true of NAEP, which started testing students with disabilities and LEP students in 1997 and has been steadily increasing the representation of these two subgroups in the sample of test-takers since that time. However, there is still a great deal of variation over time and between states. For example, in 2005, Delaware excluded 8% of its 4th grade students with disabilities and LEP students from NAEP math testing, while Alabama excluded 1%. Changes have also been made within a state: Hawaii excluded 9% of these students on the 4th grade NAEP math test in 2000 but only 3% in 2005 (NCES, 2005c).

- ***Score inflation on state tests.*** Many researchers believe scores on state tests have become inflated as a result of overly narrow teaching to the state tests (e.g., Koretz, 2005). Because test preparation is so specifically targeted at the content and format of the state tests, the gains do not generalize to NAEP.

- ***Some differences in grades and years analyzed.*** The grades and time spans we compared for state tests and NAEP were not always identical. As stated earlier, for a small number of states we substituted another elementary grade if state test data for grade 4 were not available. In addition, while the NAEP time frame was 2003 through 2005, our state test trends covered varying years between 2002 and 2006, depending on the data available from each state. Due to these data limitations, these results should be considered suggestive rather than definitive.

It is likely that some combination of the above factors explains the inconsistencies in NAEP and state test results, and different factors may be present in different states and at different grade levels. For example, the fact that we found a significant relationship for 4th grade reading but not the other grades and subjects might be attributable to the alignment issue noted above. It could be that state 4th grade reading tests and the NAEP 4th grade reading tests are relatively similar in content and format.

# Chapter 7
# Quality and Limitations of State Test Data



## Key Findings

- Considering the massive amount of test data that states are required by NCLB to report and considering the weight that the law places on these data, one might assume that the information needed to reach definitive conclusions about achievement would be accessible and fairly straightforward. But the house of data on which NCLB is built is at times a rickety structure. Our efforts to conduct a rigorous study of test score trends were hampered by missing or inconsistent data, limited availability of data beyond percentages proficient, and breaks in the comparability in test data. The major reasons why data are limited include overburdened state departments of education; technical or contractual issues with testing contractors; and continual corrections and revisions in test results.

- Data necessary to do rigorous analyses of achievement trends—including mean scale scores and standard deviations—were not available in many states. Only 30 states provided both the means and standard deviations necessary to calculate effect sizes for analyses of overall achievement trends, and only 27 states provided both these types of data by subgroup to enable us to analyze gap trends. The numbers of students tested, which should be available in all states, were available for students in general in 44 states and for subgroups of students in 43 states. Although NCLB does not require states to make these data available, they are basic types of information that all testing programs should have.

- NCLB spurred massive changes in and expansion of state testing programs. As a result, 37 states had breaks in their test data that affected the comparability of test results across the years and limited the types of analyses of test score trends that could be done. These breaks occurred because states switched to a different test, changed their proficiency levels or cut scores, revised their scoring scales, or made other significant changes in their testing programs.

- Although public reporting of test results is a critical tool for accountability under NCLB, our experience suggests it would be difficult for the average citizen in some states to obtain information about trends in test scores. Transparency in state test data could be improved by such steps as posting data in an easy-to-find place on state Web sites and providing clear information and cautions about breaks in trend data. To help researchers reach accurate conclusions about trends in student achievement, states should also make available standard deviations, mean scale scores, and numbers of test-takers.

## Disparity between Ideal and Available Data

In addition to describing our findings about achievement trends, this report is meant to serve an educational purpose by making people aware of data issues that work against the goal of transparency in NCLB data and that hamper efforts to do comprehensive studies of student achievement. This chapter addresses that purpose.

The No Child Left Behind Act requires states to publicly report a massive amount of information on student performance and other issues in the form of state "report cards" that parents can understand. For test data in particular, NCLB requires all states to report the percentages of students scoring proficient in both math and reading, broken down by subgroup. Most states did not do this before 2002, so in that sense, the law has compelled states to report information about student achievement that did not exist before or was not made public. **Table 20** specifies the types of assessment data that must be included in NCLB state-level report cards (in addition to data about AYP and teacher quality).

**Table 20. Assessment Data Requirements for State-Level Report Cards under NCLB**

| Type of Data | All Students | Major Racial & Ethnic Groups | Students With Disabilities | Limited English Proficient | Economically Disadvantaged | Migrant | Male/ Female |
|---|---|---|---|---|---|---|---|
| Percentage of students tested | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Percentage of students achieving at each proficiency level | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Most recent 2-year trend data in student achievement for each subject and grade assessed | ✓ | | | | | | |

Table reads: The state report cards required by NCLB must include data on the percentage of students tested for all students in the state as well as for the major racial and ethnic subgroups, students with disabilities, limited-English-proficient students, economically disadvantaged students, migrant students, and students of both genders.

Note: The subgroups of migrant students and male and female students are required subgroups for reporting purposes but are not required subgroups for AYP determinations.

*Source: U.S. Department of Education, 2003.*

These test data serve other important purposes in addition to informing the public about student progress. NCLB also requires states and school districts to use test data to reach decisions about whether schools and districts have made adequate progress in raising achievement and whether they should be identified for improvement or more stringent sanctions.

In light of these NCLB requirements, we expected that the evidence necessary to do a thorough and rigorous analysis of overall achievement and gap trends would be accessible and fairly straightforward. We soon realized this wasn't always the case. We did find a great deal of data, but we were sometimes surprised and disappointed to find that the house of data on which NCLB is built is a rickety structure. Our efforts to gather all the information we needed were hampered by three notable factors—inconsistent, hard-to-find, or incomplete data on state Web sites; a lack in some states of test data beyond percentages proficient; and breaks in comparable data.

This chapter describes the types of data we requested and the number of states that did and did not provide the requested data during phase I of our study. It also explores the problems created by the three factors mentioned above. Finally, the chapter makes suggestions for improving transparency in NCLB data.

## Availability of Data

Using the methods described in the appendix, we sought to collect the following information for each state, for each year from 1999 and 2006:

- Descriptive information about the state's testing program, including information about changes that might affect the comparability of results
- The percentages of students scoring at each achievement level by grade and subject for students in general in the state and for subgroups
- Mean test scores and standard deviations by grade and subject for students in general and for subgroups, so we could compute effect sizes
- The numbers of test-takers by grade and subject for students in general and for subgroups, so we could determine which subgroups were small or had changed significantly in size

**Table 21** lists the number of states that provided each of the desired types of data and the source from which it was obtained. The first two rows represent data that NCLB requires states to include on state report cards. The last six rows depict data that NCLB does not require states to make public on their report cards but that are integral to all testing systems.

As shown in the first two rows of table 21, all 50 states made percentage proficient results available for all students and for subgroups, as required by NCLB. The other types of data listed in the remaining rows of the table were less widely available, for reasons discussed below.

**Table 21. Number of States Providing Each Type of Test Data and Source of Data**

| Type of Data | Web site | Web site, augmented through state contacts | State verification process | Not available | Not available in time for report |
|---|---|---|---|---|---|
| **Data Required by NCLB** | | | | | |
| State percentage proficient | 25 | 25 | 0 | 0 | 0 |
| Subgroup percentage proficient | 22 | 27 | 1 | 0 | 0 |
| **Data Not Required by NCLB** | | | | | |
| State mean scale scores | 14 | 19 | 8 | 8 | 1 |
| Subgroup mean scale scores | 11 | 18 | 10 | 10 | 1 |
| State standard deviations | 6 | 12 | 16 | 14 | 2 |
| Subgroup standard deviations | 3 | 10 | 16 | 19 | 2 |
| State N-counts | 21 | 19 | 4 | 5 | 1 |
| Subgroup N-counts | 17 | 20 | 6 | 6 | 1 |

Table reads: In 14 states mean scale scores for the state as a whole were obtained from the state Web site. In 19 states these data were obtained initially from the state Web site with additional data supplied through state contacts, and in eight states they were provided through the state verification process. Mean scale scores were not available in time for this report in one state and were not available at all in eight states.

*Source: CEP and HumRRO, 2007.*



## HARD-TO-FIND, INCONSISTENT, OR INCOMPLETE DATA

Although states do post achievement data on their Web sites, state Web sites vary in their quality and accessibility. Some Web sites were easy to navigate and had much of the data we sought, but others were seriously lacking. It was not always obvious where to find key data. In addition, it became apparent early on in our data-gathering process that some of the test data reported on the Web were outdated or inconsistent or were missing key information.

These problems were not wholly unexpected, nor do they mean that states are not making a good faith effort to fulfill their reporting responsibilities under NCLB. As discussed later in this chapter, there are several reasons why achievement data are hard-to-find, limited, or incomplete. Moreover, guidance from the U.S. Department of Education does not explicitly require states to post their report cards on the Web, although non-regulatory guidance encourages states "to disseminate state report cards in multiple ways" and notes that "states might post their report card on the state's Web site and make copies available in local schools, libraries,

parent centers" (U.S. Department of Education, 2003). Additionally, states are not required to report some of the data we sought. Still, we found holes and discrepancies even in some of the mandatory data. As shown in table 21, we could not obtain complete data from state Web sites on percentages proficient for students in general in 25 states and for subgroups of students in 27 states.

To obtain accurate and complete data for our analyses, we had to address these problems. In many cases, we had to contact state officials directly to fill in gaps, answer questions, and make corrections. As an additional check, we asked all states to verify the data we had gathered, using the process mentioned in chapter 2 and explained more fully in the appendix. As the third and fourth columns of table 21 illustrate, we would not have obtained as much data as we did without contacting states and asking them to verify data. In some states, the verification process entailed multiple contacts with state staff or personal appeals from CEP's president. In a few cases, CEP provided states with funding to defray the extra costs of correcting or supplying information, or securing unpublished data from test contractors. Most states made some modifications in their data during verification.

In short, the professionals involved in this study spent many months and considerable effort tracking down information and working with states to verify data. We could not have completed this task without the cooperation of state officials, who by and large were helpful in providing what they could when asked, and we are grateful for their assistance. Our experience suggests how difficult it would be for the average parent or citizen—who does not have the contacts and resources of a national organization—to obtain information about test score trends in some states. Suggestions for improving the transparency of data are outlined at the end of this chapter.

## MEAN SCORES AND STANDARD DEVIATIONS

Mean test scores and standard deviations were more difficult to obtain than percentages proficient, especially for subgroups. As table 22 near the end of this chapter indicates, 30 states provided some data to compute effect sizes for post-NCLB achievement trends, and 27 provided some data to compute effect sizes for achievement gaps. Many states, however, did not provide enough years of comparable mean scores or standard deviations to measure trends in both subjects for all grade spans and subgroups. Therefore, our analyses for different grades, subjects, and subgroups are based on different subsets of states. The total numbers of states with sufficient data to do each type of analysis are indicated in the tables in chapters 4 and 5.

These limitations affected our ability to compute effect sizes for a large proportion of states. On one hand, this is not surprising, since NCLB law does not require states to report mean scale scores and standard deviations. On the other hand, all states would be expected to have these data, or at least be able to access them through their testing contractors, in order to conduct simple analyses of their test results.

### NUMBERS OF TEST-TAKERS

Most, but not all, states were able to provide counts of the numbers of students tested. The numbers of students tested were available for all students in 44 states, and for subgroups of students in 43 states (table 21). It is surprising that some states did not provide these N-counts, because NCLB requires states to report the percentage of students tested, and this figure cannot be determined without knowing the number of students tested. For the purposes of our study, these numbers were used in the achievement gap analyses described in chapter 5 to determine whether some subgroups should be excluded because their numbers of students were too small or had changed rapidly.

### IMPEDIMENTS TO OBTAINING DATA

State education officials were helpful in providing this information when asked. The reasons for these data limitations were largely institutional in nature. The main impediments were as follows:

- ***State capacity.*** As highlighted in a recent CEP study of state capacity to implement NCLB, many state departments of education face serious funding constraints and are understaffed in terms of both numbers of employees and expertise (CEP, 2007d). State assessment offices are particularly overburdened, especially since implementation of NCLB. Although officials in many states made considerable efforts to voluntarily retrieve the data needed, not all states could spare the time or staff.

- ***Contractor issues***. Most states contract with testing companies to develop and score their tests. To obtain the data needed for this study, many state officials turned to their contractors, which sometimes could not supply the data due to technical or contractual issues. In the latter case, providing these data went beyond what was explicit in the state's contract with the testing company.

- ***Changing data.*** The initial reporting of test results in most states is rarely the last word. Test results on state Web sites are continuously modified and updated. Often there are revisions, appeals from schools and districts, corrections, rescoring, or various administrative problems that plague even the best-run state testing systems. These factors make accurate and complete test data somewhat of a moving target.



## Breaks in Trend Lines

As explained in chapter 3, the integrity of our analyses depended on identifying breaks that limited the number of comparable years of data available for analysis. Thirty-seven states had breaks in the data used for this study; these states are marked with asterisks in table 22 later in this chapter. Below are examples of common actions states took in implementing or changing testing systems that affected our ability to compare one year of test data with another. The list is not exhaustive, but it does suggest the difficulties of tracking test score results over time.

### CHANGED OR NEW TESTS

Since the starting point of our analysis (1999), many states have introduced new tests or changed existing tests in ways that have created breaks in comparability of test results. For example, Alaska's state tests were changed in grades 3-8 in 2005, allowing us to compare results only for 2005 and 2006 in elementary and middle school grades. The state test in

grade 10, however, was changed in 2006, so we analyzed performance for that grade based on Alaska's previous grade 10 test for 2002 through 2005.

New Jersey phased in new assessments over time, so the years for which data are comparable vary across grades and subject areas. In 4th grade, reading scores are comparable from 2001 through 2006, and math scores are comparable from 1999 through 2006. In 8th grade, reading and math scores are comparable from 1999 through 2006. In 11th grade, reading and math scores are comparable from 2002 through 2006. Therefore, only limited pre- and post-NCLB comparisons were possible, but in all cases post-NCLB trends are based on four full years of data.

### CHANGES IN CUT SCORES

Many states changed their cut scores defining proficient performance, which also created breaks in year-to-year comparability of test data. For example, Arizona changed its cut scores in 2005; as a result, comparisons were restricted to 2005 and 2006 results. Similar actions were taken in Delaware and Oregon. In Michigan, proficiency levels changed in 2003, and in 2006 content standards were revised, new standards were set, and the assessment window shifted from winter to fall, rendering results not comparable with previous years. Because of these changes, we were only able to compare performance at grades 4 and 7 in reading and grades 4 and 8 in math from 2003 through 2005. We do not have information about whether cut scores were set lower or higher.

### OTHER CHANGES

Other state actions created various types of breaks in test data. For instance, Arkansas introduced new standards and a new vertical scoring scale in 2005, so analyses of grade 3-8 results were limited to 2005 and 2006—too short to be called a trend. In Nevada, an official in the Office of Assessments, Program Accountability, and Curriculum advised that we should compare data only from 2004 onward, due to changes in the pool of test items and their quality.

## Summary of Available Data and Suggestions for Improvement

**Table 22** summarizes the number of states that provided the data necessary to do the different types of analyses in phase I of this study. The states with asterisks had breaks in their data due to changes in their testing systems. Since many factors influence the availability of data, this table should not be seen as a ranking of states or as a judgment on their testing systems.

**Table 22. States Categorized by Types of Analyses Possible**

| States | Percentage Proficient Trends | | Effect Size Trends | | Gap Trends | | Total # of States |
|---|---|---|---|---|---|---|---|
| | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | Pre-NCLB | Post-NCLB | |
| CT*, DE*, KS*, KY, LA, MO*,NC*, PA, SC, WA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 10 |
| OR | ✓ | ✓ | | ✓ | ✓ | ✓ | 1 |
| NJ* | ✓ | ✓ | ✓ | ✓ | ✓ | | 1 |
| AR*, HI, ID, IN*, IA, MI*, MN*, MS, MT*, NV*, NM*, ND*, TN*, UT*, WI* | | ✓ | | ✓ | ✓ | ✓ | 15 |
| GA*, OK*, WV | | ✓ | | ✓ | ✓ | | 3 |
| CO*, MA*, NH*, NY*, VA, WY* | ✓ | ✓ | | | ✓ | | 6 |
| AL*, AK*, AZ*, CA*, FL* IL*, ME, MD*, NE* OH*, RI*, SD*, TX* | | ✓ | | | ✓ | | 13 |
| VT* | | | Very little usable data | | | | 1 |

Table reads: Ten states provided data necessary to analyze pre- and post-NCLB overall trends in terms of both percentages proficient and effect sizes. These same states also provided the data necessary to analyze pre- and post-NCLB trends in achievement gaps in terms of both percentages proficient and effect sizes.

*State has a break in its testing data that limits analyses to some extent. For details on specific years of comparable data and reasons for breaks, see the online state profiles at www.cep-dc.org/pubs/stateassessments.



Our experience with gathering information for this study suggests a need for improvements to ensure that the data necessary to track student achievement are available and accessible. Granted, states are being pulled in countless directions to comply with all the demands of NCLB, and making data more transparent is often not their highest priority. Nevertheless, strengthening accountability is a major goal of NCLB and public reporting of test results is a critical tool for accomplishing this goal, as the U.S. Department of Education has noted in its 2003 guidance on NCLB report cards.

As an initial step, the data that states are required to report could be presented in a more transparent way. Posting state report cards on the state Web site would help greatly to make data accessible, but this alone does not guarantee transparency. The posted data should also be easy to find and presented in a way that is understandable to parents and others.

Addressing the issue of breaks in data is another step. At a minimum, states should report any changes in their testing program that affect comparability of test results. They should

also clearly caution people about which types and years of comparisons are not appropriate. In addition, states should take into account comparability of data when making judgments about whether to change testing programs. Although states cannot be expected to stop changing their testing systems merely for the sake of continuity, they should make changes only for sound educational reasons.

Finally, states should report mean scores, standard deviations, and N-counts—data that should be available for any testing system. These data are vital information for studies of achievement trends. Since these data are primarily of interest to researchers, they do not have to be part of the state report card, but they should be posted in a clear location on state Web sites.

**Box E** lists the type of test-related information that ideally should be available on state Web sites. Many state officials interviewed for phase II of this study pointed out that they are overwhelmed with requests for data. Providing these data in an easily-accessible location, such as a Web site, might eliminate many of those requests.

## Box E. Ideal Test Information on State Web Sites

A fully transparent assessment system would include the state, district, and school report cards in a form appropriate for parents, as well as the elements listed below. While these are recommended for all tests used for NCLB, similar documentation would be appropriate for all statewide tests.

1. A full set of assessment results for students as a whole and disaggregated by demographic subgroups (racial/ethnic categories, low-income students, students with disabilities, and limited English proficient students). This set would include results for reading and math at each grade level (and science, once implemented).

    1.1. Percentage of students at each achievement level
    1.2. Mean scale scores
    1.3. Standard deviations
    1.4. Numbers of students tested and percentage of population tested

2. A description of any substantive changes in the assessment program each year by test and grade level, including the changes in the following aspects:

    2.1. Content standards
    2.2. Cut scores for each achievement level
    2.3. Test scoring scale
    2.4. Test format, such as the addition or removal of open-ended questions, changes in test length, or the introduction of adaptive or norm-referenced components
    2.5. Scoring rules
    2.6. Rules for including students in testing, including changes in policies for students with disabilities and LEP students
    2.7. Rules for categorizing students by subgroups
    2.8. Stakes attached to test results, such as making the test a condition for promotion or graduation decisions
    2.9. Timing of the test, such as a shift from spring to fall testing
    2.10. Equating methodology
    2.11. Testing contractor
    2.12. Method for calculating AYP

3. Rationales for any changes that create breaks in comparability in test data

4. Links to important assessment studies, including the following types of studies:

    4.1. Alignment
    4.2. Linking
    4.3. Standard-setting
    4.4. Cut score values

*Source: Center on Education Policy and HumRRO*

# Appendix
# Study Methods





This report is based on the data gathered and analyzed during phase I of CEP's achievement study. Additional qualitative information will be collected from a subset of approximately 20 states during phase II of the study.

## Collecting and Verifying Phase I Data

The process of collecting and verifying phase I data began in July 2006 and lasted through March 2007. All 50 states provided at least some of the information we sought. As explained below, officials from each participating state were asked to verify the accuracy of the state information.

### TYPES OF DATA COLLECTED

During phase I, CEP and HumRRO sought the following information from every state for each year between 1999 and 2006:

1. Name of test(s) used for NCLB accountability (back to at least 1999)
2. Testing contractor/developer with point-of-contact information
3. Independent quality assurance/verification contractor, if any, with point-of-contact information
4. Key state department of education staff (1-2 people, such as the testing director) with contact information
5. Test type (norm-referenced, criterion-referenced, augmented criterion-referenced, computerized adaptive, other)
6. Test scoring scale
7. Description of and cut scores for achievement levels on the test
8. Longitudinal/vertical scale or separate scale by test level
9. Grades tested in reading and mathematics

10. Grades included in determinations of percentages proficient in reading and math for AYP purposes

11. Subjects tested other than reading and math (just a list with tested grade levels)

12. Timing of test (spring or fall)

13. Frequency of testing (usually annual)

14. Item types (only for reading and math tests used to calculate AYP)

15. Equating methodologies (summarized by type, such as Stocking-Lord, Rasch) and length of time this method has been used

16. Most current state-level mean score, numbers of test-takers (N-counts), and standard deviations for reading and math at all tested grades

17. Mean scores, N-counts, and standard deviations for all subgroups reported by the state

18. Mean scores and standard deviations for the entire state and by subgroup for each comparable year back to 1999

19. Percentages of students scoring at each achievement level reported by the state, for the entire state and by subgroup at all grade levels back to 1999

20. Start date (if longitudinal scale is present)

21. Major changes to the assessments from 2002 to the present—for example, changes in standards setting, cut scores, or in tested grade levels

22. Studies of alignment of test to state content standards



## PILOT TEST AND INITIAL DATA COLLECTION

To prepare for data collection, two experienced HumRRO education researchers conducted a pilot test of the collection process in two states, Maryland and Kentucky. Working from a list of desired information, each researcher collected as much information as possible from the Web sites of the two state departments of education.

The data from the pilot test were used to design templates and instructions for the full data collection. Using these templates and instructions, senior HumRRO staff trained the other staff members who were assigned to collect the data.

As a first step in collecting the desired data, HumRRO staff searched Web sites maintained by state departments of education. Supporting information was also collected from the Web sites of the Council of Chief State Schools Officers, the Education Commission of the States, the National Center for Education Statistics, and the U.S. Department of Education. When Web sites lacked crucial information, HumRRO staff attempted to directly contact state department of education staff by e-mail or phone.

The data from each state were put into one descriptive MS Word table and one MS Excel file with one worksheet per assessment year. The Excel file contained separate quality control worksheets that highlighted extreme changes from year to year, which could indicate possible data entry errors.

HumRRO staff prepared a summary report for CEP that included tables and charts of all the data gathered as of September 2006. The report and the accompanying data were reviewed by the expert panel and CEP staff, and were discussed at a meeting on September 21-22. At this meeting, a decision was made to expand the data collection process to include results from 2005-06 testing and to ask all 50 states to explicitly verify the data collected before any major analyses were done.

HumRRO staff spent several weeks collecting additional data.

## STATE VERIFICATION PROCESS

To get ready for the state verification process, HumRRO staff reformatted the data files to make them more easily understandable to state personnel who would not have the benefit of training or in-person assistance from HumRRO. In addition, HumRRO prepared a CD for each state containing the following information already collected:

- Test characteristics file
- State assessment scale score data file
- State assessment percentages proficient data file
- Data verification checklist, which state officials were asked to sign
- Directions for completing the verification tailored for each state (see attachment A at the end of the appendix)
- A letter from CEP to the chief state school officer (see attachment B at the end of the appendix)

CEP and HumRRO took the following steps to ensure that officials from every state verified the accuracy of their state's data and filled in missing information:

1. ***Initial letter.*** In mid-November 2006, CEP President Jack Jennings sent a letter (attachment B) and a CD to three people in each state: the state superintendent or other chief state school officer, the deputy superintendent, and the assessment director. The letter asked state officials to verify the data on the CD and to make necessary corrections or additions by December 15, 2006. The letter included an offer from CEP to help defray costs.

2. ***First extension.*** Many states asked CEP for more time to verify the data, so CEP extended the deadline to December 22, 2006.

3. ***Second extension.*** Some states needed still more time, so on December 13, 2006, CEP sent an e-mail to the states that had not yet responded, extending the deadline to January 10, 2007.

4. ***Follow-up efforts.*** HumRRO staff and CEP's president communicated with individual states to maximize participation. The last data file was received March 8. The last bit of related documentation (the test characteristics file and verification checklist) was received March 21, 2007.

5. ***Questioning anomalies.*** During the analysis phase, HumRRO staff contacted state officials when anomalies in the data arose, up until April 2, 2007.

During the verification process, some state officials added missing data themselves. Other states hired outside help (sometimes with CEP funding) to add missing data. Still others provided raw data in various forms to HumRRO staff, who added the data to the appropriate tables and sent the revised file back to the state.

Most states returned modified data Excel files and modified files of test characteristics file by e-mail and faxed the signed verification checklist. Throughout the verification period, HumRRO maintained a log of all e-mail and phone communications with states (444 records).

## Analyzing Phase I Data

Analyses of the data collected were conducted from August 2006 through April 2007. Analyses were expanded and revised as data from more states became available, as the scope of the project evolved, and as states made corrections and filled in missing data. Moreover, the process changed as CEP, HumRRO, and the expert panel developed rules to guide the analyses.

### PROCESS FOR ANALYZING STATE DATA

In August and September 2006, HumRRO staff did an initial analysis of the data collected at that point from a limited number of states. Brief findings from these initial analyses were included in the report HumRRO submitted to CEP for the September 2006 panel meeting. As a result of this meeting, the scope of the project was expanded. The analysis process was put on hold until additional data were collected and verified by the states.

HumRRO resumed the analysis process in December 2006. One HumRRO staff member reviewed the test characteristics files and state verification logs to produce a guideline that defined which years and grade levels should be analyzed for each state, explained below. Verified data files were left intact, and separate Excel files were created to extract data from these files and conduct analyses.

For the meeting with CEP and the expert panel on January 24-25, 2007, HumRRO produced a series of tables and figures for the 26 states with verified data at that point. In addition, HumRRO drafted a brief, initial analysis of national trends, based on the 19 states that had provided the data necessary to analyze achievement trends. During that meeting and in the weeks that followed, CEP and HumRRO developed a set of rules for analyzing trends, based on advice from the expert panel. These rules are described in chapter 3.

After each state verified its data, HumRRO staff produced a draft profile, consisting of data tables and figures for that state and a brief description of trends or changes gleaned from HumRRO's analysis of the data. One HumRRO staff person was assigned to conduct a series of quality control checks on each state profile, primarily to confirm that the profile matched the original source data file. HumRRO provided the draft profiles to CEP as they were completed.

A senior CEP consultant with expertise in educational testing carefully reviewed the data and initial trends analysis in each state's profile, and made revisions as necessary in the profile's summary of findings and description of trends. This consultant also made sure that the analysis adhered to the rules that had been developed. Another senior CEP consultant edited the text, developed new templates for tables, and reformatted and edited the tables as needed to help present the information in the state profiles as clearly as possible.

## SELECTION OF GRADES TO REPORT AND ANALYZE

For all of the achievement analyses, CEP and HumRRO looked separately at the elementary, middle, and high school levels. In states that tested multiple grades within each of these grade spans, HumRRO collected data for all the grades tested between 1999 and 2006, then made decisions about which specific grades to report on and analyze. These decisions were based on a fixed set of criteria, applied consistently across all states and developed without regard to whether achievement was high or low at a given grade.

In some states, the analysis of the overall percentages of students scoring proficient covered more grades than the analysis of achievement gaps or effect sizes. As a result, data are available for grade levels and years that are not analyzed in this report.

To determine which grades to cover in the analysis of overall percentages proficient, HumRRO went through the following process at each grade span (elementary, middle, high school):

1. Within each span, the grade with the longest post-2002 trend line was selected as a starting point for the analysis. Because states introduced tests in different grades at different times, the years covered sometimes varied at the elementary, middle, or high school levels.

2. Every grade in place over the same trend period as the longest trend line was included. For example, if a state had tests for grades 3-6 in place from 1999-2005, all of these grades were shown in the appropriate tables and figures and were analyzed. But if another state tested grades 3 and 4 from 1999-2005, then introduced grades 5 and 6 in 2000, only grades 3 and 4 were shown. HumRRO included multiple grades to counter any criticism that the study was "cherry picking" good or bad examples of achievement.

For the analyses of overall achievement in terms of effect size and for both types of achievement gap analyses, HumRRO selected one representative grade at the elementary, middle, and high school levels from among the grades included in the overall percentage proficient analysis. As noted above, the grade was selected without regard to whether performance was high or low at that grade. HumRRO used the following criteria:

1. ***Elementary.*** The first choice was grade 4. If grade 4 was not among the grades selected in steps 1 and 2 above, another grade was chosen in the following order of preference: grade 5, 3, or 6.

2. ***Middle.*** The first choice was grade 8. Other choices in order of preference were grades 7, 9, or 6.

3. ***High.*** The first choice was grade 10. Other choices in order of preference were grade 9 or 11. Some states use end-of-course exams as their high school tests for NCLB; these tests are administered when a student takes a specific course, rather than in a fixed grade.

4. ***No grade repetition.*** HumRRO never used the same grade to satisfy two grade spans (for example, grade 6 could not be used to represent both the elementary and the middle school spans).

## NATIONAL SUMMARY ANALYSIS

After all 50 state profiles were completed, HumRRO generated an Excel file that summarized trends across all 50 states. A senior consultant to CEP analyzed this summary to develop the findings in chapters 4 and 5.

## Attachment A. Sample Directions to State about Verifying State Data Instructions for State Assessment Data Verification — Alaska (AK)

HumRRO staff collected information about your state assessments from your state department of education web site and other sources. Our goal is to gather descriptive information about your assessment as well as assessment results over time. We tried to collect assessment results from 2002 to the present, and from 1999-2002 if your current assessment was in place during those years.

In order to facilitate analysis and reporting, we have organized data for all states into a common format. As you proceed through the following steps, please do not add or remove rows or columns from the worksheets. If a demographic category is not used in your state, or if you do not conduct assessments at all the grade levels listed, please leave those cells blank.

In order to verify your state's assessment data, please complete the following steps:

1. **Review and Edit the Test Characteristics File *(Alaska (AK) Test Characteristics.doc).*** This Word document contains basic characteristics about your state's assessment system that HumRRO was able to glean from your website.

    a. **Check information in the table to make sure it is correct.**

    b. **Fill in any missing information.**

    c. **Pay particular attention to the "Major Changes to the Assessments" field.**

        i. We want to know everything major that has changed in your state accountability system since the implementation of No Child Left Behind (NCLB) in 2002. If your NCLB assessment was already in place in 2002, we are interested in changes going back to 1999.

        ii. We are especially interested in any changes that would render results incomparable from year to year. Examples include changing assessment instruments, conducting new standard setting, or altering the content standards.

2. **Review and Edit the State Assessment Scale Score Data *(AK Scale Scores.xls).*** This Excel file contains Mean Scale Scores, N-Counts, and Standard Deviations for the state and disaggregated into demographic groups by grade and content area. Each assessment year has a separate worksheet.

    a. **Check information in the spreadsheets to make sure it is correct.**

    b. **Add any missing data that you may have.**

3. **Review and Edit the State Assessment Proficiency Data *(AK Proficiency.xls).*** This Excel file contains Proficiency data for the state and disaggregated into demographic groups by grade and content area. Please note that these files have two sections: Above the solid black line is the percentage of students meeting the standard (i.e., proficient or above) from your state. Below the line are raw data for each performance level. Each assessment year has a separate worksheet.

   a. **Check information in the spreadsheets to make sure it is correct.**

   b. **Add any missing data that you may have.**

4. **Complete the Data Verification Checklist *(Verification Checklist.doc).*** When you have reviewed each of the three files above, please complete and sign this checklist to indicate your approval and return it to HumRRO with the rest of your information.

5. **Return All Materials to HumRRO.** When you have finished verifying the data and adding additional information, please return the following files to HumRRO:

   a. **Revised Test Characteristics File**

   b. **Revised State Assessment Scale Score Data File**

   c. **Revised State Assessment Proficiency Data File**

   d. **Completed and signed Data Verification Checklist**

**Please send the above information in whatever format is convenient for you (via e-mail, diskette, or CD) to Sunny Becker at HumRRO by December 15th, 2006:**

Sunny Becker
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314-1591
e-mail: SBecker@HumRRO.org

The signed Data Verification Checklist may be FAXed to Dr. Becker at (703) 548-5574 if other files are provided via e-mail.

**Specific Notes about Alaska Data**

We have culled all of the information that we could from the Alaska Department of Education website, but we were unable to locate some of the data.

- For your Proficiency file, we were able to locate data for most demographic groups from 2002-2005. Please provide any missing information that you have available, particularly for Title I students (if available), which we could not locate for any year.

- For your Scale Score file, we were able to locate only N-Counts from 2002-2006. Please provide mean scale scores and standard deviations across ALL groups for these years.

- Please verify that the disaggregate group American Indians was expanded to include Alaska Natives in 2006.

- We were unable to locate state assessment data for the years 1999-2001. Please verify that your state did not have assessments in place at that time.

**Questions?**

We are available and happy to answer any questions you have during the data verification process.

For questions regarding the files discussed above, please contact:

- Sunny Becker (HumRRO) at SBecker@HumRRO.org or (800) 301-1508, extension 679.

For questions regarding financial assistance or the purpose of the project, please contact:

- Jack Jennings (CEP) at cep-dc@cep-dc.org or (202) 822-8065.

Given the sensitivity and importance of the data with which we are working, we want to ensure that you will have adequate time to review the information. If you feel that you will need more time to do the data verification, please let us know. Otherwise, please recall that a lack of response from you by December 15th, 2006 will indicate your approval of these data. We greatly appreciate your time and assistance in this very important line of research.

## Attachment B. Sample Letter to Chief State School Officer about Verifying State Data

CENTER ON EDUCATION POLICY CEP

1001 Connecticut Ave., NW, Suite 522
Washington, DC 20036

phone 202.822.8065
email cep-dc@cep-dc.org
fax 202.822.6008
web www.cep-dc.org

November 17, 2006

Dear State Superintendent :

The most important questions in determining the impact of the No Child Left Behind Act are 1) whether students are learning more and 2) whether achievement gaps between groups of students are narrowing. To address these questions, the Center on Education Policy is undertaking a major analysis of state assessment data and related information on achievement. I am writing to ask you to verify the accuracy of the information we have collected so far about your state and to let you know more about this project.

The project is funded by the Carnegie Corporation, the Hewlett Foundation, and the Ford Foundation. Our analysis is being overseen by a panel of testing and policy experts that includes Laura Hamilton of RAND; Eric Hanushek of the Hoover Institution; Frederick Hess of the American Enterprise Institute; Robert Linn of the National Center for Research on Evaluation, Standards, and Student Testing; and W. James Popham of the University of California at Los Angeles.

For the first phase of the project, we are collecting vital information on state assessment programs and results from the years before NCLB was enacted in 2002 up through 2006. Early next year, we will publish a report containing profiles of every state, including yours. Our purpose is to make available basic data on all state assessment systems and to determine whether achievement has increased and achievement gaps have narrowed since 2002. Toward this end, our contractor, Human Resources Research Organization (HumRRO), has composed a draft profile for every state based on data from the state department of education Web site, the U.S. Department of Education NCLB Web site, the Council of Chief State School Officers Web site, and direct contacts with state education agency personnel.

Could you please verify the factual accuracy of the enclosed profile of your state? We realize that state assessments are complex, and we want to be even-handed and accurate in presenting a picture of student achievement. If you have additional information that you feel we should consider, or if you wish to correct or ask a question about the data we display, please contact Sunny Becker of HumRRO (703-549-3611 or sbecker@humrro.org).

We recognize that you and your assessment personnel face many demands, but we must complete our project on schedule to provide timely information to policymakers for the

NCLB reauthorization. Therefore, we must receive corrections or additions from you by December 15, 2006; otherwise, we will assume that the enclosed information is correct. We would be willing to help defray costs involved in verifying or completing this information with some of our limited foundation grant funding. For instance, your test contractor might charge you to provide data on means or standard deviations for your assessments. If you need this assistance, please contact me directly at 202-822-8065 before you begin your work.

In the second phase of our project, we will ask a select subset of states to work with us to answer more completely the two key questions about achievement. During this phase, HumRRO researchers would visit a state for one day. Based on the information gathered, the Center will write a report that aims to determine in greater detail on the national level whether student achievement has increased and achievement gaps have narrowed since 2002. We are reviewing the initial data we have collected and may approach you about your state's participation in this second phase. If you would like more information about this phase, I would be glad to call or e-mail you.

During the upcoming reauthorization of NCLB, there will be considerable discussion about student achievement. The objective of our analysis is to concentrate attention on the fundamental issue—whether students know more and whether achievement gaps have narrowed. We hope we can count on your assistance in addressing this issue. Feel free to contact me for further information. For your information, we have sent your deputy and your assessment director a copy of this letter and also of the disc with you state's profile.

Sincerely,
Jack Jennings
President and CEO

cc: Deputy Superintendent
Assessment Director

# Glossary of Technical Terms





**Accumulated annual effect size –** The cumulative gain in effect size tracked over a range of years (see the definition of *effect size* below). For example, to determine the accumulated annual effect size between 2002 and 2004, one would calculate the change in effect size from 2002 to 2003 and from 2003 to 2004, then add these differences together.

**Cut score –** Minimum test score that a student must earn to pass a test or attain a specific achievement level such as "proficient" or "advanced." For purposes of No Child Left Behind, each state sets its own cut score to define proficient performance. Cut scores are typically set by panels of educators based on their expert judgment about what students should know and be able to do at a certain grade level, as well as on actual student performance on the test.

**Effect size –** A statistical tool intended to convey the amount of change or difference between two years of test results (such as 2005 and 2006) or between two subgroups (such as Hispanic and white students) using a common unit of measurement that does not depend on the scoring scale for that particular test. For example, to measure the amount of change between two years of test results, an effect size is computed by subtracting the 2005 mean test score from the 2006 mean score, then dividing the result by the average standard deviation (see the definition of *standard deviation* below). If there has been no change in the average test score, the effect size is 0. An effect size of +1 indicates an increase of 1 standard deviation from the previous year's mean test score. The resulting effect size can be compared with effect sizes from other states that use different tests on different scales.

**Mean scale score –** The arithmetical average of a group of test scores, expressed on a common scale for a particular state's test (see the definition of *scale score* below). The mean is calculated by adding the scores and dividing the sum by the number of scores.

**Percentage proficient –** The proportion of students in a group, such as all students in a particular state, that scores at or above the cut score for "proficient" performance on a test. The percentage proficient is the primary measure of achievement used to determine progress under NCLB.

**Scale score –** A type of test score that converts a student's raw score (the actual number of questions answered correctly) into a score on a common scale for a particular state's test, in order to control for slight variations between different versions of the same test. Scale scores are helpful because each year most testing programs use a different version of their test, which may differ from a previous version in the number or difficulty of questions. Scale scores make it possible to compare performance on different versions of the same test from year to year.

**Standard deviation –** A measure of how spread out or bunched together the numbers are in a particular data set. In testing, a standard deviation is a measure of how much test scores tend to deviate from the mean. If scores are bunched close together (all students' scores are close to the mean), then the standard deviation will be small. Conversely, if scores are spread out (many students score at the high or low ends of the scoring scale), then the standard deviation will be large.

**Z-score –** A statistical tool that allows one to compare changes in the percentage proficient across states with different proficiency cut scores. Briefly, z-scores represent in standard deviation units the difference between the percentage proficient for a specific state (or subgroup or grade level) and 50% proficient (mean performance). The conversion table to make this transformation can be found in standard statistics books. This report uses Z-scores to address the following situation: When the percentage proficient for a state or subgroup falls near the center of the standard distribution of scores, a small increase in the mean score could result in a large increase in the percentage proficient, but when the percentage proficient falls at the higher or lower end of the distribution, the same small increase in the mean score could produce a much smaller increase in the percentage proficient. By controlling for this situation, Z-scores allow one to compare achievement changes more fairly across subgroups and states.

# References


Achieve. (n.d.). *2005 NAEP results: State vs. nation.* Retrieved February 26, 2007, from www.achieve.org/node/482.

Center on Education Policy. (2002). *What tests can and cannot tell us.* Washington, DC: Author.

Center on Education Policy. (2003). *Implementing the No Child Left Behind Act: A first look inside 15 school districts in 2002-03.* Washington, DC: Author.

Center on Education Policy. (2004). *From the capital to the classroom: Year 2 of the No Child Left Behind Act.* Washington, DC: Author.

Center on Education Policy. (2005). *States test limits of federal AYP flexibility.* Washington, DC: Author.

Center on Education Policy. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Author.

Center on Education Policy. (2007a). District survey of NCLB implementation, February 2007. [Unpublished data]

Center on Education Policy. (2007b). *It's different now: How exit exams are affecting teaching and learning in Jackson and Austin.* Washington, DC: Author.

Center on Education Policy. (2007c). *No Child Left Behind at five: A review of changes to state accountability plans.* Washington, DC: Author.

Center on Education Policy. (2007d). *Educational architects: Do state education agencies have the tools necessary to implement NCLB?* Washington, DC: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dillon, S. (2005, November 26). Students ace state tests, but earn D's from U.S. *New York Times.* Retrieved May 2, 2007, from www.nytimes.com/2005/11/26/education/26tests.html?ei=5088&en=fdf05ea7edbf1440&ex=1290661200&partner=rssnyt&emc=rss&pagewanted=print.

Education Week (2006). *Quality counts at 10. A decade of standards-based reform.* Bethesda, MD: Editorial Projects in Education.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement.* Working paper 06-1. Retrieved April 18, 2007, from pace.berkeley.edu/NCLB/WP06-01_Web.pdf.

Hamilton, L. (2003). Assessment as a policy tool. In Floden, R.E., ed., *Review of Research in Education, 27,* 25-68.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*(1), 3-17.

Jacob, B. A. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments.* Cambridge, MA: National Bureau of Economic Research.

Kolen, M. J. (1988, Winter). Traditional equating methodology. *ITEMS Instructional Topics in Educational Measurement.* Madison, WI: National Council on Measurement in Education.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: Harvard Civil Rights Project.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms. *Journal of Educational Measurement: Issues and Practice, 9*, 5-14.

Linn, R. L. (n.d.). Unpublished data.

National Center for Education Statistics. (2005a). *The nation's report card: Reading 2005.* Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2005b). *The nation's report card: Mathematics 2005.* Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2005c). Percentages of students in states and jurisdictions identified, excluded, and assessed in mathematics. Retrieved May 1, 2007, from nces.ed.gov/nationsreportcard/nrc/reading_math_2005/s0094.asp?printver=.

National Research Council. (1997). *Educating one and all: Students with disabilities and standards-based reform.* McDonnell, L. M, & McLaughlin, M. J., eds. Washington, DC: National Academy Press.

National Research Council. (1999). *Grading the nation's report card. Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

Pellegrino, J. (2007, March). Should NAEP performance standards be used for setting standards for state assessments? *Phi Delta Kappan, 88* (7), 539-541.

Peterson, P. E., & Hess, F. M. (2006). Keeping an eye on state standards. *Education Next, 3.* Retrieved April 11, 2007, from www.hoover.org/publications/ednext/3211601.html.

Peterson, P. E., & West, M. R. (2003). The politics and practice of accountability. In Peterson, P. E., & West, M. R., eds., *No Child Left Behind? The politics and practice of school accountability.* Washington, DC: Brookings Institution.

Phillips, G.W. (2007). *Linking NAEP achievement levels to TIMSS.* Washington, DC: American Institutes for Research.

Popham, W. J. (2006, August 29). Let the AYP pigeons fly. *The Pulse.* Retrieved May 10, 2006, from www.districtadministration.com/pulse/commentpost.aspx?news=no&postid=16875.

Popham, W.J. (in press). Instructional insensitivity of tests: Accountability's dire drawback. Forthcoming in *Phi Delta Kappan.*

Stockburger, D.W. (1996). *Introductory statistics: Concepts, models, and applications.* Retrieved March 9, 2007, from www.psychstat.missouristate.edu/introbook/SBK11.htm.

U.S. Department of Education. (2003, September 12). Report cards: Title I, Part A, non-regulatory guidance. Retrieved May 13, 2007, from www.ed.gov/programs/titleiparta/reportcardsguidance.doc.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

## Credits and Acknowledgments

This report was written by Naomi Chudowsky, Victor Chudowsky, and Nancy Kober, CEP consultants. Research and analysis for the report were conducted by the Human Resources Research Organization (HumRRO) under the director of Sunny Becker. Other key HumRRO staff involved in the project were Art Thacker, Lauress Wise, Hilary Campbell, Monica Gribben, Lee Koger, and Emily Dickinson. Jennifer McMurrer, CEP's research associate, developed tables 11 and 14 and generally assisted with the project. Jack Jennings, CEP's president and CEO, and Diane Stark Rentner, CEP's director of national programs, provided advice and assistance.

We would like to thank our panel of expert advisors—Robert Linn, W. James Popham, Laura Hamilton, Eric Hanushek, and Frederick Hess—for their invaluable advice. Additionally, we are grateful to chief state school officers and state assessment personnel for their cooperation in providing information on state testing programs and student achievement data. We also want to thank Gene Wilhoit and Rolf Blank of the Council of Chief State School Officers for their assistance with this project.

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 Ewing Marion Kauffman Foundation, The Carnegie Corporation, The William and Flora Hewlett Foundation, and The Ford Foundation for their support of our work on this study of achievement and the No Child Left Behind Act. The George Gund Foundation, The MacArthur 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.



www.cep-dc.org


**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