The Impact of the No Child Left Behind Act on Student Achievement and Growth: 2005 Edition

APRIL 2005

John Cronin
G. Gage Kingsbury
Martha S. McCall
Branin Bowe

A technical report from the NWEA Growth Research Database

NWEA
Northwest Evaluation Association
Partnering to help all kids learn

Table of Contents

PREFACE: An Introduction to the Research Series

Welcome to the first in a yearly series of studies investigating the impact of the No Child Left Behind act on the achievement of students in the United States. This legislation holds out great promise for education but it also has strong requirements and includes a host of provisions that have never been tried on this scale before. This series will use achievement information from a broad sample of students and schools to provide evidence about the changes that have occurred since the law passed.

We hope that the series will be useful for policy agencies, educational researchers, and others with an interest in improving education. While the series will use statistical procedures to identify trends and levels of impact, it should be accessible to anyone with a basic knowledge of experimental design.

This study uses the Growth Research Database from the Northwest Evaluation Association to provide achievement information about hundreds of thousands of students in school districts across the country. Since the database was founded several years ago, it has provided an archive of individual student growth that was unavailable in the past. The database allows the comparison of student achievement and student growth on a common, stable scale. This simple fact provides a tool that enables researchers to investigate educational change more completely than ever before. The application of this research tool to NCLB is natural and timely.

While the full effect of the legislation will only emerge over time, this series of studies is designed to watch it as it unfolds. Results from the study of any one year will give us a single snapshot of the law as it is implemented, while the series of studies will identify trends as they occur. It is unlikely that anyone can identify the long-term impact of NCLB at this point, but it is important that we investigate to inform mid-course corrections that might be required.

The authors of this series have no political agenda other than the enhancement of education for students. While the U.S. system of public education has been one of the best in the world for the past 200 years, there is always room for improvement. Since international comparisons almost always suffer from language and cultural differences, it is only by inspecting the changes that occur in our own educational system that will lead to its improvement. We hope this series of studies will add to the information needed to foster that improvement.

G. Gage Kingsbury
Director of Research, Northwest Evaluation Association

CHAPTER 1: Introduction

In January of 2002, President Bush signed the No Child Left Behind (NCLB) act into law. This law was the reauthorization of the Elementary and Secondary Education Act, and had within it a broad spectrum of changes to the federal role in public education. It included accountability provisions which required states to test all of their students, and sanctions to schools related to low student performance on those tests. It also required states to provide additional educational opportunities for students in the school under sanction.

While we will discuss NCLB in more detail below, it is useful to understand the assumptions underlying the act, and the expectations that individuals have concerning the law and its associated regulations. While people speaking about the legislation will have slightly different interpretations, the following elements are common in the understanding of the law:

- The law will provide an accountability system to identify which schools are doing a good job with their students.
- The law will enhance the opportunities for students who are in danger of not learning the skills that are needed in reading and mathematics.
- The law will enhance the capacity for all students to become proficient.
- The law will reduce the achievement gaps seen among students in a variety of subgroups.

This study will examine how well the law is beginning to meet its promise in its first years of implementation. It will investigate how much student achievement status has changed since the law was implemented. It will investigate whether and to what extent student achievement growth has changed since the law was implemented. Finally, it will investigate the impact of the law on the achievement status and growth of students by ethnic group.

This is the first year in the series of studies that will investigate the impact of NCLB. Each year, the study will be repeated and expanded to give a broader picture of the manner and the extent to which the law affects student achievement. NCLB is just beginning to apply sanctions and add requirements to education practice in low achieving schools. States are just beginning to increase the percentage of students achieving proficiency in order to be identified as successful. States are moving to expand assessment programs to include all grades from 3 to 8 and high school. Any or all of these factors may have an impact on student education in the years to come.

While this study is quite large in terms of the number of students and schools involved, it should be interpreted with caution, since the law has only been in effect for three years. It is expected that the law will have a cumulative effect on student achievement, since many aspects of the law are being phased in over the next few years. The findings of this study may be indicative of the potential of the law, but additional time will be needed to identify its ultimate effectiveness.

To allow concentration on key elements of the law, the study will investigate the areas of student achievement and student growth. While there are a host of individual aspects of the law that require research, this study will emphasize only achievement. In particular, the study will deal with the following research questions:

- Are students' achievement scores higher than they were when NCLB first went into effect?
- Is student achievement growth higher than it was when NCLB first went into effect?
- Are achievement gaps among ethnic groups shrinking under NCLB?
- Given current rates of change in achievement, are schools likely to meet the requirements of NCLB?

To investigate these questions, the study will use a large sample of student achievement scores, selected from a broad cross section of school districts throughout the United States. These districts were chosen because they have achievement scores measured on a common scale from the 2001-2002 school year until the 2003-2004 school year. This enables the comparison of student achievement prior to specific influence of NCLB and following the implementation of the law. It also allows the comparison of individual student growth patterns prior to and following the implementation of the law. This sample allows a strong investigation of the impact of the law for individual students and for schools.

CHAPTER 2: NCLB Legislation and Educational Improvement Efforts

Educational improvement principles have been enacted in state policy and law since the 1980's. They have been largely effective in raising achievement for a broad range of students. The No Child Left Behind act, enacted in 2002, has set admirable, but very challenging goals which may cause us to overlook the considerable success that public schools have had in the decades previous to the passage of the law. It is useful to consider the impact of NCLB in the context of the standards-based education movement that was underway prior to its passage.

Background

The No Child Left Behind act (NCLB), the current version of the 1965 Elementary and Secondary Education Act (ESEA) passed on January 8, 2002 with broad bipartisan support. ESEA provides funds for schools that serve students from families in poverty to bring opportunity for poor students more in line with that of their peers. The intent of the law has always been to promote equitable academic achievement in public schools, but the method for achieving this goal has evolved over time. Earlier versions of the law focused on using teaching methods, curricula and textbook adoptions that would give needy children better access to skills. At that time legislators believed that the role of federal funds flowing to states was to equalize school funding and ensure equal treatment which would inevitably lead to academic success. ESEA funds were to be spent on specific services to low-performing students in schools serving poor neighborhoods.

Much discussion focused on differences in teaching and learning philosophies that seemed arcane to the community at large. As time went by with little achievement change for schools receiving ESEA funds, and a public perception that American schools were not internationally competitive, policy makers concluded that schools were not expecting enough of students, particularly poor students. By the 1980's education reformers from across the political spectrum sidestepped debates about methods and textbooks by allowing schools to adopt whatever they wished as long as they could prove that students were learning appropriate material.

The National Conference on State Legislatures (2005, page 5) comments on the national change in thinking.

> Standards-based accountability systems moved from a focus on the equality of opportunity to equality of outcomes for all students – especially for minorities and poverty-stricken students. Advocates became no longer concerned just with student access and school equality (i.e., inputs) but with equality of student performance (i.e., outcomes).

The 1994 version of ESEA, the Improving America's Schools Act (IASA) required states to devise standards and assessment systems, piggybacking on existing state education reform laws, but allowing states to devise their own accountability systems for Title I programs. By the time NCLB was signed, all but two states had devised standards-based accountability systems of their own. States had assessment systems in place at selected benchmark grades, and most had published

explicit content standards spelling out what students were expected to learn and had set performance standards specifying the level of achievement required. Many states published school and district results expressed in percentages of students passing state standards. Results were published both overall and disaggregated by ethnic group, students in poverty, disabled and limited English proficient groups. Many states developed growth models as part of their accountability system (Shields, Esch, Lash, Padilla, & Woodworth, 2004). NCLB added the following elements to the ongoing reform movement:

- A single federal accountability system for all states, eliminating growth models.
- A concrete goal of having 100% of students meeting standards by 2014.
- A set of uniform sanctions for schools and districts not meeting goals.
- A requirement that disaggregated (as well as whole group) results carry sanctions.

The introduction to Title I of NCLB clarifies its outcomes-based focus:

> The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments.

To provide opportunity, the law requires each state to apply and report the NCLB accountability measure, adequate yearly progress (AYP), for all public schools, not just those receiving federal NCLB funds. Schools and districts that do get NCLB money are not required to directly tie federal funds to services, but are subject to sanctions if they do not meet achievement goals. These sanctions include allowing students to transfer to schools with better performance and, eventually, closing schools that fail to meet targets.

As the National Council of State Legislatures (2005) points out, NCLB changes the relationship between federal and state powers. States have been responsible for public education constitutionally, since it is not listed as a federal power, and historically, as states have funded and governed public schools. The federal government derives its authority from the spending clause, which allows it to attach conditions to states for the receipt of federal funds. NCLB expands federal powers by extending federal evaluation definitions to districts and schools that do not get these funds and by overriding state definitions of AYP.

> Under NCLB, the federal government's role has become excessively intrusive in the day-to-day operation of public education by trying to incorporate the principles of individual state standards-based reforms and condensing them in one federal statute that imposes a one-size-fits-all accountability system. (NCSL, 2005, p.11)

Has this change in the federal role been worthwhile in terms of raising academic skill? Does NCLB add benefit to states beyond that already provided by existing educational reform policies? These are open questions that this study will help address.

Assumptions of NCLB

Prior to IASA, policymakers assumed that equalizing funding would result in equalized academic results. Under education reform and NCLB a new set of assumptions have to be made.

Assumption: Sanctions Will Cause Higher Academic Performance

Roderick & Engel (2001) showed that, although many low-performing students improved performance in reaction to external goals, a core of low-performers remained unaffected even in the face of considerable sanction. They concluded that policies that put all responsibility for success on the performance of children are bound to fall short of high proficiency for all.

Amrein & Berliner (2002, 2003) found that students in states with high-stakes policies fared no better on SAT, ACT and AP examinations than those in low-stakes states. They found a similar pattern with National Assessment of Educational Progress (NAEP) results and concluded that performance on state tests fails to generalize to other instruments. Although Amrein & Berliner's SAT, ACT and AP analysis stands, other researchers refute their NAEP finding. Depending on the type of analysis and the definition of "high stakes" other researchers (Braun, 2004; Rosenshine 2003; Carnoy & Loeb, 2003) tend to find that high stakes testing increases NAEP cross-sectional performance but may weaken cohort gain results.

Research findings about the role of external goals and sanctions in promoting school success are, at best, mixed. Some of this has to do with who gets sanctioned. States may have high stakes for students (tests determine graduation or promotion) but low stakes for teachers, principals and superintendents. Hanushek & Raymond (2004) report that states attaching consequences to their accountability systems have greater success than states that only report results. They could not determine whether NCLB had an effect because state systems had been well established prior to NCLB's implementation.

Assumption: High Expectations Will Cause Higher Performance

NAEP results are often used in national studies because they are based on the only nationally administered test allowing direct state comparisons. Since NAEP consequences for students are low, scores are believed to be free of coaching or test preparation activities. On the other hand, students are not as motivated to perform on NAEP as they are on their state tests, which vary in rigor. Furthermore, although NAEP's achievement standards are uniform, its exemption policies are not, confounding state-to-state comparisons. (Using NWEA data allows a cross-state analysis of both cross-sectional and cohort achievement on a uniform scale with standards that approximate those of the state.) Nevertheless, researchers report common finding with regard to NAEP results (Linn, 2004; McCombs, et. al., 2004; Gissmer, et. al., 2000). McCombs, et. al. (2004)

note that in every state more than 50% of students score below NAEP proficiency levels. The average proficiency rates were 30% on the fourth grade test and 32% on the eighth grade test.

There has been steady significant improvement in NAEP mathematics scores and slight improvement in reading in the past decade. Mathematics scores have increased significantly in both fourth and eighth grades from 1990 to the present (National Center for Educational Statistics, 2005; Campbell, Hombo, & Mazzeo, 2000; Gissmer, et. al., 2000). The percent of fourth and eighth graders reaching NAEP proficiency in reading rose between 1992 and 2003, although the percent meeting basic and advanced levels did not and mean scores remained the same (NCES, 2005).

The mathematics improvement, while modest in any particular year, is remarkable taken in the aggregate over time. One of the criticisms of state tests is that they will encourage narrow teaching that will not generalize beyond state measures. Yet NAEP, established to provide comparability between states and nationally over time, shows steady mathematics improvement in a large and diverse nation. It is worth noting here that these increases began well before the passage of NCLB, when standards-based systems were prevalent in states using their own accountability formulas and that scores increased during a period when states were cutting back resources.

The rate of improvement is far lower than that needed to meet NCLB goals. Linn (2004) notes the discrepancy between mandated and observed growth.

> In mathematics, for example, the percentage of students at the proficient level or above on NAEP would have to have an annual rate of improvement between 2003 and 2014 that is 2.3 times as fast at grade 4 as the rate actually realized between 2000 and 2003. At grade 8, the rate of improvement in the percentage of students at the proficient level or above in mathematics would need to be 6.5 times as rapid between 2003 and 2014 as it was between 2000 and 2003. Such rapid acceleration of achievement trends is unrealistic. In reading, the rate of increase in percentage proficient or above is an even more unrealistic jump. For grade 4 the annual rate of improvement would have to be 15.7 times as fast for the next 11 years as it actually was between 1998 and 2003. At grade 8, annual rate of improvement would need to be 10.2 times as great as it was between 1998 and 2003.

Since this kind of growth has never been known to occur, there has been a call for a more realistic evaluation of policy goals (NCSL, 2004; Packer, 2004; Linn, 2003). However, some policymakers continue to believe that schools could make these hitherto unseen rates of growth if they were to radically change instruction (Chubb, Linn, Haycock & Wiener, 2005).

State standards and assessment results. State tests differ in the rigor of performance targets and thus cannot be compared to one another, but can be compared with themselves across time. Results are generally consistent with those of NAEP. The Education Trust (2004) examined elementary grade results for 2002, 2003, and 2004 and found that most states have improved in that time according to their own standards, although state rates of improvement are sometimes more dramatic than NAEP gains. The authors also find that even when improvement is greater on state tests than on NCES, rates of change are too low to meet NCLB's goal of 100% proficiency in 2014.

The Center on Education Policy (2005) did a comprehensive study of how states have been affected by NCLB. Of the 49 states surveyed, 72% (36 states) reported that achievement had improved, 16% (8) said it had remained the same, 4 were unsure and only 1 state reported a decline. Districts receiving NCLB funds reported similar results (72% improving, 22% remaining the same, 6% declining) with more large and medium sized districts reporting themselves as improving (95% and 80% respectively). Nevertheless, state and district officials realize that increases are not sufficient to reach NCLB's goal of 100% proficiency in 2014.

The authors conclude that, although achievement gains are real, they cannot be attributed to NCLB both because it is too early to see the law's effects and because the AYP model is an insufficient statistic for drawing conclusions. Chapter 3 of the CEP study profiles states and districts that had more sophisticated accountability models in place prior to NCLB. For them, the imposition of the NCLB AYP model has hampered reform efforts and hindered communication. However, for states and districts without systems in place NCLB provided a framework for proceeding with reform.

Assumption: Holding High Standards Will Close the Achievement Gap

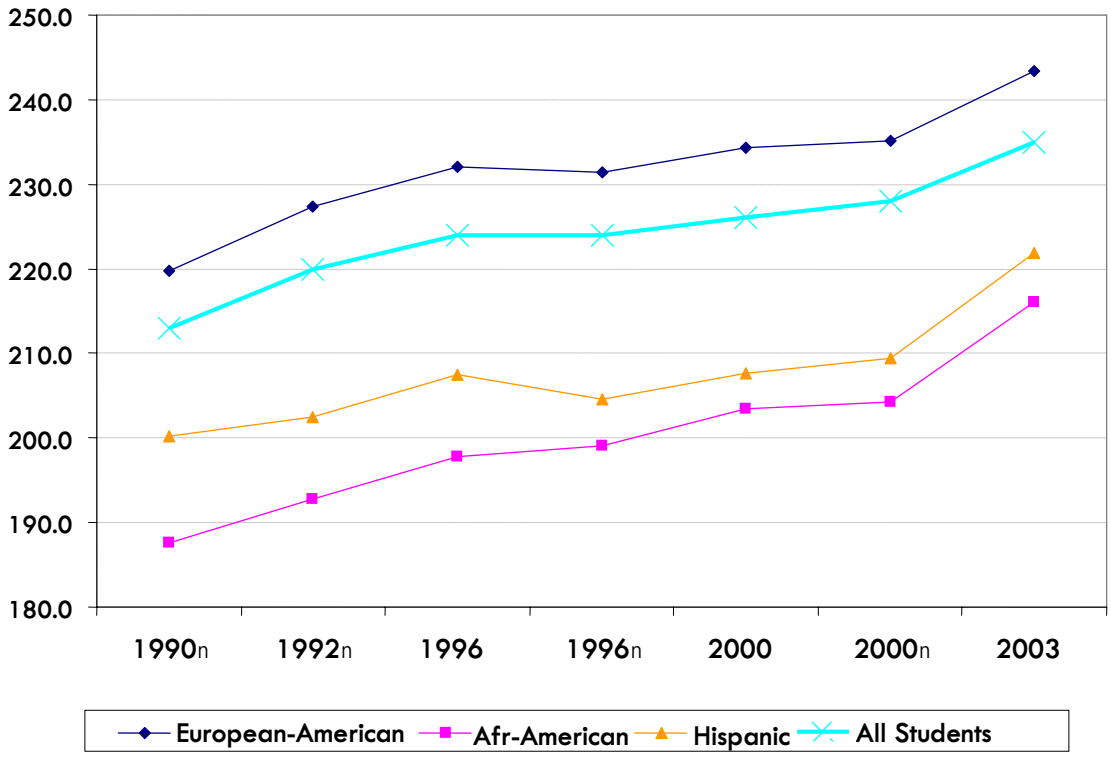
The Education Trust (2004) and Center for Educational Policy (2005) concur that more states report a narrowing of the achievement gap than a widening of the gap since the passage of NCLB. This is true for each ethnic group (African-American, Asian, Hispanic and Native American) as well as for Limited English Proficient students and students with disabilities. At the same time, however, long term trends on NAEP show the African-American/European-American (i.e. white, Caucasian) achievement gap remaining constant throughout the 1970s, narrowing in the 1980s and widening in the 1990s. The Hispanic/European-American gap narrowed throughout the 1970s and 80s, increasing in the early 1990s and narrowing again in the late 1990s. These trends held for reading, mathematics and science. During the period of educational reform, minority performance in mathematics increased as did the performance of European-American students (whites), but minority gains were not sufficient to narrow the achievement gap. The chart on the next page illustrates this pattern for fourth grade NAEP progress in mathematics for 1990-2003.

Hanushek & Raymond (2004) find that while attaching state consequences to results raises NAEP achievement overall and for subgroups, gains for African-American and Hispanic students are less than those of European-American students, resulting in a larger achievement gap.

National Assessment of Educational Progress

National/Mathematics Composite/Grade 4/2003, 2000, 1996, 1992 and 1990

Student race/ethnicity based on school records (supplemented in some cases by student self-reported data) [SDRACE]



n Accommodations were not permitted for this assessment

NOTE: The NAEP Mathematics scale ranges from 0 to 500. Observed differences are not necessarily statistically significant.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003, 2000, 1996, 1992 and 1990 Mathematics Assessments.

Researchers who study growth measures, which are not currently included under NCLB's AYP model, have found that when the initial condition is taken into account, growth rates are fairly constant across ethnic groups (Goldschmidt, 2004). This is notable in that it indicates similar cognitive processes come into play in the learning continuum. However, because the proportion of minority students at lower levels is typically greater than that of European-American students, aggregate performance at any point in time favors European-American students. Clearly, growth rates of minority students need to increase in order to close the achievement gap.

The reform movement has been instrumental in calling attention to the achievement gap and making it a top priority. It has also insisted that all students be held to the same academic goals and that these goals be made explicit and public. This has been effective in raising achievement for all groups, but not in narrowing the gap. In addition, reporting proportions of students meeting performance criteria can mask continuation of the gap. If all students were to meet proficiency, but the patterns of group difference in mean scale score performance were preserved, access to top jobs and schools would still be denied to minority groups.

Conclusions

The research evidence to date addresses the potential for NCLB to succeed in all of its aspects. Among the most important implications of the research to date are the following:

- The standards-based education movement, which gained popularity prior to NCLB, was shown to increase student achievement when used with state accountability models.
- The presence of content and performance standards in a system of accountability for schools and students has been shown to increase academic performance.
- The addition of a set of federal sanctions has not been shown to further increase performance.
- The addition of high-stakes testing has not been shown to reduce the achievement gaps among students of different ethnicity.
- The AYP model currently in use in NCLB may not identify schools that are doing a good job of helping low performing students grow and could mask achievement gaps.

CHAPTER 3: Method of Study

In order to determine what effects, if any, the No Child Left Behind act (NCLB) has had on student achievement, a dataset was assembled that included data from the 2001-2002 and the 2003-2004 academic years. These years represent time before NCLB took effect (2001-2002) and after NCLB took effect (2003-2004).

Student Sample

The dataset included reading assessment data from over 320,000 third through eighth grade students in more than 200 school districts located in 23 states. Mathematics data came from over 334,000 third through eighth grade students in more than 200 school districts located in 22 states. All assessment information came from NWEA tests, which put all scores onto a single, stable measurement scale (the RIT scale). This enables meaningful comparisons across time, and allows the calculation of growth scores for individual students.

Table 1 – Number of students, schools and school districts included in the reading study by state and academic year

State	Number of Unique Districts	Number of Unique Schools	2002 Student Count	2004 Student Count
AZ	1	3	1093	1110
CA	7	104	32293	32979
CO	17	116	20143	20343
FL	1	1	56	52
IA	30	72	10640	10900
ID	24	59	5562	5755
IL	2	9	1741	1740
IN	77	247	50772	52269
KS	1	2	294	311
KY	1	9	936	946
MI	5	9	1117	1112
MN	5	12	2144	2122
MT	7	21	3528	3645
NE	3	3	333	333
NM	6	29	3560	3653
NV	2	7	504	510
OH	2	5	574	585
OR	3	14	1181	1181
PA	2	6	1677	1688
WA	12	83	16824	17166
WI	5	7	971	935
WY	3	26	2773	2822
Total	**216**	**844**	**158716**	**162157**

Table 2 – Number of students, schools and school districts included in the mathematics study by state and academic year

State	Number of Unique Districts	Number of Unique Schools	2002 Student Count	2004 Student Count
AZ	1	3	1120	1111
CA	7	104	29709	29761
CO	19	158	27285	27672
FL	1	1	56	59
IA	30	74	10895	11178
ID	24	63	5922	6056
IL	3	17	4161	4217
IN	76	250	50687	51898
KS	1	3	721	701
KY	1	9	894	899
MI	5	12	1734	1664
MN	6	13	2320	2206
MT	7	21	3694	3767
NE	4	4	376	345
NM	6	32	4258	4358
NV	2	10	737	778
OH	2	4	546	555
OR	2	14	1385	1403
PA	2	6	1680	1693
WA	11	83	15152	15419
WI	9	10	1239	1215
WY	3	22	1686	1763
Total	222	913	166257	168718

In order to be included in the dataset a student must have had both a valid fall and spring NWEA assessment in either the 2001-2002 or 2003-2004 academic years. All of the assessment records were extracted from NWEA's Growth Research Database (GRD)[1] matching on a unique student identifier key. All invalid assessments were excluded from the study. Within a subject, a student could have one test for each term.

In addition to these constraints, we included only schools that tested about the same amount of students. The number of tested students in a school could not differ by more than 20% between the 2002 and 2004 academic years. We employed this constraint in order to ensure that

[1] The GRD contains longitudinal student assessment information from Measures of Academic Progress (MAP) and Achievement Level Tests (ALT) from over 1400 school districts and over 8900 schools in 45 states dating back to 1996. All records in the GRD are uniquely identified to ensure that students and their assessments can be accurately tracked across time.

comparisons between grades were made with schools testing the same proportion of students. Students in schools who exceeded this threshold were not included in the dataset.

Another element included in the dataset was whether or not a state test was administered in a grade level. Tables 3 and 4 indicate the grades in which state assessments were given in both the 2002 and 2004 academic years. Each NWEA test record had an indicator of whether or not it was administered in a grade and subject in which a state test was also administered. This allowed comparison of student growth in grades where a state test was administered to growth in grades without a state test.

It should be noted that Indiana actually administers their assessment in the fall for grades 3, 6 and 8. For the purposes of this study, we chose to assign the state assessment flag to the prior grade since the majority of the instruction for the assessment took place in the prior grade.

Table 3 – Grades in which a state reading test was administered in 2002 and 2004 (grey shading indicates an assessment)

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
AZ	■		■			■
CA	■	■	■	■	■	■
CO	■	■	■	■	■	■
FL	■	■	■	■	■	■
IA		■				■
ID	■	■	■	■	■	■
IL	■		■			■
IN			■		■	
KS			■			■
KY		■			■	
MI		■			■	
MN	■		■			■
MT		■				■
NE		■				■
NM	■		■			■
NV	■		■			■
OH		■		■		
OR	■		■			■
PA			■			■
WA		■			■	
WI		■				■
WY		■				■

Table 4 – Grades in which a state mathematics test was administered in 2002 and 2004 (grey shading indicates an assessment)

State	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
AZ	■		■			■
CA	■	■	■	■	■	■
CO			■	■	■	■
FL	■	■	■	■	■	■
IA		■				■
ID	■	■	■	■	■	■
IL	■		■			■
IN			■		■	
KS		■			■	
KY		■			■	
MI		■			■	
MN	■		■			■
MT		■				■
NE		■				■
NM	■		■			■
NV	■		■			■
OH		■		■		
OR	■		■			■
PA			■			■
WA		■			■	
WI		■				■
WY		■				■

This study also incorporated an analysis of growth by student ethnicity. In the GRD, the student ethnic code is standardized to one of seven values: American/Alaskan Native, Asian/Pacific Islander, African-American, Hispanic, European-American, Unknown and MultiRacial. Tables 5 and 6 show the count and percentage of the sample in each grade that comes from each ethnic group.

Table 5 – Number of students taking reading assessments by ethnicity and grade

Grade	Ethnicity	American/ Alaskan/ Native	Asian/ Pacific Islander	African-American	Hispanic	European-American	Unknown	Multiracial	Total
3	Count	1145	1636	1758	7404	34613	6020	295	52871
	Percent	2.17	3.09	3.33	14.00	65.47	11.39	0.56	100
4	Count	1209	1302	1884	7232	34134	3962	307	50030
	Percent	2.42	2.60	3.77	14.46	68.23	7.92	0.61	100
5	Count	1166	1377	1903	7777	41122	4128	268	57741
	Percent	2.02	2.38	3.30	13.47	71.22	7.15	0.46	100
6	Count	1311	1243	1567	7955	40223	3068	163	55530
	Percent	2.36	2.24	2.82	14.33	72.43	5.52	0.29	100
7	Count	1181	1421	1931	8026	38282	4486	216	55543
	Percent	2.13	2.56	3.48	14.45	68.92	8.08	0.39	100
8	Count	1170	1256	922	7048	35273	3349	140	49158
	Percent	2.38	2.56	1.88	14.34	71.75	6.81	0.28	100

Table 6 – Number of students taking mathematics assessments by ethnicity and grade

Grade	Ethnicity	American/ Alaskan/ Native	Asian/ Pacific Islander	African-American	Hispanic	European-American	Unknown	Multiracial	Total
3	Count	1293	2404	2130	8342	38429	6791	324	59713
	Percent	2.17	4.03	3.57	13.97	64.36	11.37	0.54	100
4	Count	1475	1824	2015	7953	36633	4207	311	54418
	Percent	2.71	3.35	3.70	14.61	67.32	7.73	0.57	100
5	Count	1273	2183	2115	9386	44865	4206	272	64300
	Percent	1.98	3.40	3.29	14.60	69.77	6.54	0.42	100
6	Count	1345	1665	1604	8377	41028	4060	153	58232
	Percent	2.31	2.86	2.75	14.39	70.46	6.97	0.26	100
7	Count	1308	1394	1752	8257	36295	5823	219	55048
	Percent	2.38	2.53	3.18	15.00	65.93	10.58	0.40	100
8	Count	1294	988	817	6118	30279	3630	138	43264
	Percent	2.99	2.28	1.89	14.14	69.99	8.39	0.32	100

Characteristics of NWEA Assessments

All scores for the NWEA assessment in a subject area reference a single, cross-grade, equal-interval scale developed using Item Response Theory methodology. These scales are referred to as RIT scales (Ingebo, 1997). The RIT scale is designed to measure student growth and performance across time as well as to take advantage of strong measurement theory and experimental design, and have been demonstrated to be extremely stable over twenty years of development and use (Kingsbury, 2003). This stability holds for each subject area measurement scale (reading, mathematics and language usage) and across grades levels from 3 to 8 within subjects (Northwest Evaluation Association, 2002).

The dataset included NWEA assessments delivered by both the computerized adaptive Measures of Academic Progress (MAP) and the paper-and-pencil based Achievement Level Tests (ALT). Although these assessments are delivered in two mediums, our studies have shown that mode of test administration does not affect the student's achievement level estimate (Kingsbury, 2002).

Measures of Academic Progress (MAP) assessments are administered via computer and item difficulties adapt in difficulty depending on the student's performance. Once an item is answered, the student achievement level is estimated and another appropriate item is shown to the student. If the student answers a question correctly, a more difficult item is displayed. Conversely, if a student answers a question incorrectly, a less difficult item is displayed. As the items are selected within the test, the estimate of achievement becomes more precise. This iterative item selection process is repeated until the test is completed. The advantage of this type of assessment is that each child is given a custom test better suited to the student and much more accurate than a traditional test (Northwest Evaluation Association, 2003.)

Achievement Level Tests (ALT) are paper-and-pencil delivered assessments designed around the difficulty of the content rather than the age of the student. ALT assessments are built by taking a broad range of content-specific material and breaking it down into relatively small, targeted ranges of item difficulty. A grade-specific test will use only one form to measure student achievement within a class, while an ALT assessment has between 7 and 9 levels to choose from based on student ability. This means that each student taking an ALT test will be challenged with items appropriate for their achievement level. Grade-level assessments will be challenging only to students who are at or around the mean achievement level for that grade.

NWEA's assessments are designed to align directly with each state's content standards. NWEA accomplishes this by cross-referencing the state's content standards with the index that organizes the NWEA item bank. NWEA's MAP and ALT assessments have a combined item bank of more than 12,000 multiple choice test items. NWEA also has conducted state alignment studies for **17 states** that relate state proficiency scores to the RIT scale (Kingsbury, et. al.; 2003).

Student Growth

One measure of whether No Child Left Behind has had an effect on student achievement is how much students in each grade level grew before and after the law was implemented. One way to calculate a growth statistic is to simply subtract the beginning assessment score from the ending assessment score (Raw Growth).

Another way to compare student growth is through indicators of unexpected growth. One such measure is the Growth Index, which is simply the student's Raw Growth minus the expected growth given the initial score. To identify the expected growth for a student, RIT Block Growth Norms are used (Northwest Evaluation 2002). The RIT Block Norms were created by selecting a large sample of students, and then dividing them into 10 point blocks based on their initial test score. The average growth for all of the students in a particular block is the expected growth for students in that block.

Given this information, the Growth Index can easily be calculated. As an example, if an eighth grade student scored 217 on a fall mathematics assessment, they would be expected to grow approximately eight RIT points based on their RIT Block Growth Norm. If the student scored 227 on their spring assessment (10 RIT points of growth), their Growth Index score would be +2. If the same student grew only 6 points, the Growth Index score would be -2.

Analysis

In this study, we distinguish between **achievement level** (the score that a student has at one point in time) and **achievement growth** (the difference in scores for a single student from one point in time to another). These two ways of looking at achievement are useful because we want to ask cross-sectional questions (How does the achievement level of this year's fourth grade class compare to last year's fourth grade class?) and we also want to ask longitudinal questions (How much achievement growth have this year's fourth grade students made since they were in third grade last year?). This distinction differentiates many of the analyses in the study.

The analysis in the study takes the following two primary forms:

> First, we compare achievement level and achievement growth prior to the implementation of NCLB (the 2001-2002 school year) and following the implementation of NCLB (the 2003-2004 school year). This cross-sectional panel comparison allows us to examine the state of student performance, and provides some indication of the impact of the law. Conclusions from these analyses are limited because the passage of time contains many more elements than just the implementation of a federal law. Many school districts were involved in funding cutbacks during the same time period. Events in world politics may have changed educational focus to social studies during the time in question. A host of other events tangentially related to education may have influenced student achievement. The results from these analyses are meaningful, but we need to avoid causative statements concerning the results.
>
> Second, we compare achievement growth across time for set of students. This type of analysis allows us to investigate changes in growth and patterns of growth seen before and during implementation of NCLB. This repeated measures comparison provides strong evidence of change for a specific group of students during a specific time period. Limitations of this type of analysis center around the sample characteristics of students who have growth scores. These students tend to be slightly more stable than their peers for whom growth scores can't be calculated.

Within the types of analyses, the statistical approach is fairly straightforward. Univariate and multivariate statistics are used to draw a picture of the change accompanying the implementation. Effect sizes are calculated where appropriate.

CHAPTER 4: Results: Performance and Growth

The No Child Left Behind act (NCLB) was signed into law in January of 2002. This study looks at changes in student performance and growth that occurred between the school year prior to implementation of the law (2001-2002) and the most recently completed school year (2003-2004).

Implementation of NCLB is still in a relatively early stage, so dramatic effects on student achievement or growth would not be anticipated. Several elements of the act are not yet fully implemented. For example, the requirement that schools test all students in reading and mathematics in grades 3 through 8 is not fully implemented until 2005-2006. In addition, the new accountability requirements of the act are only beginning to affect many schools.

Furthermore, it is difficult to say whether to say whether changes in performance can be attributed to NCLB. NCLB encodes certain principles of assessment in law, primarily the principle that students should be tested in every grade between grades 3 and 8. However, many states had embraced this principle in their own testing programs. States such as South Carolina, for example, tested all students in grades 3 through 8 prior to implementation of NCLB. The fact that many states had implemented high stakes testing prior to NCLB passage in benchmark grades allows us to make robust comparisons of performance and growth between students in grades that are and are not tested.

If NCLB and its associated regulations are having an impact, that impact should be manifested in some of these ways:

- Students entering a grade in 2004 should achieve higher test scores than students entering the same grade in 2002.
- Students entering a grade in 2004 that participated in state testing the prior year should enter that grade with higher test scores than students in the same grade who were not tested during the prior year.
- Students enrolled in a grade that participates in state testing should show greater growth during the school year than those enrolled in a grade that does not participate in state testing.

The results section is organized around these hypotheses. In addition to examining these hypotheses in relation to the whole group, we also disaggregated the analysis by ethnic group in an effort to determine whether some groups experience different learning effects from the implementation of high stakes testing and the law. This disaggregated information follows in the next chapter.

Performance Comparisons

Comparing fall 2001 and fall 2003 achievement. Table 7 shows a comparison between fall 2001 and fall 2003 achievement in mathematics. Readers will note that we use three different types of statistics to express differences in performance and growth. These are the ***average weighted difference, the weighted cumulative difference,*** and the ***effect size.***

Table 7 – Fall mathematics scores for students in the study sample

	Fall 2001			Fall 2003			Change	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
3	29762	190.16	12.07	29951	191.4	12.17	1.25	0.10
4	27175	201.43	11.99	27243	202.39	12.17	0.96	0.08
5	31958	209.65	13.1	32342	210.47	13.1	0.82	0.06
6	29088	216.84	14.26	29144	217.65	14.74	0.81	0.06
7	27151	223.09	16.01	27897	223.39	16.01	0.30	0.02
8	21123	229.34	17.02	22141	229.6	16.79	0.26	0.02
Cumulative Difference							4.18	
Average Weighted Difference							0.76	0.05

The ***average weighted difference*** represents the average difference in performance across all grades weighted for the number of students in the sample at each grade (as represented by the initial fall score). The use of weighting assures that grade levels with lower counts do not receive greater weight than deserved in the sample. Thus in mathematics, while the improvement in fall scores between 2001 and 2003 ranged from .26 to 1.25 RIT points in magnitude, the average weighted difference would be .76 RIT points across all grades when the differences in count are taken into consideration. In this case, fifth grade results are weighted more heavily because over 31,000 students were included at this grade, while the performance difference among 21,123 fall 2001 eighth graders receives less weight.

The ***cumulative difference*** is the sum of the measured differences across all grades sampled. This is an estimate of difference that is used to represent what kind of gain (or loss) might occur for a group of students if differences in grade level performance were sustained over time.

Because the data represent a snapshot of performance in two seasons, we would caution readers that it is hazardous to assume that students would sustain and accumulate these improvements in performance across time and ask them to take this caution into consideration when judging cumulative differences. Nevertheless, we are also convinced that reporting differences at only a single grade would tend to greatly understate the effects that small learning improvements might generate over time. Indeed, reports of historical NAEP results show that most documented improvements in learning over the past two decades have been the result of small improvements in performance that have been sustained. It seems only reasonable, therefore, to attempt to represent the effect that small improvements (or declines) might have over a number of years.

In this particular case, while the gains in mathematics starting achievement attributable to any one grade level are relatively small, these differences would lead to substantial improvement in student learning if they were sustained over time. Assume that the weighted cumulative difference cited in Table 7 is sustained for a group of 2003 third graders through their eighth grade year. This would raise the average scale score for the entire population improved by about 4 points over that length of time. Such gains would bring about a meaningful improvement in proficiency rates.

Here is an example that illustrates the point. The state of Oregon uses the RIT scale to establish their proficiency standards. Oregon's cut score for mathematics proficiency is set at a point equivalent to 235 on the RIT scale. According to the most recent norms (Northwest Evaluation Association, 2002), about 50% of the norm sample performs at or above this standard.

If the NWEA norm population improved its overall performance by an average of 4 points, this improvement would result in 59% of the NWEA norm population achieving the Oregon proficiency standard, a gain of 9%. While this kind of gain is not sufficient to ensure that goals of NCLB are met, it would represent a marked improvement in student achievement.

Finally, ***effect size*** in this study is expressed as the ratio of the difference in results to the pooled standard deviation of scores for all students in a grade and subject. This allows readers to judge the size of a change by comparing it to the variation of performance that is present in a group. Framing differences from three possible reference points gives readers the best opportunity to reach their own conclusions about the results of this study.

In mathematics, the results indicate that students entering a grade in 2003 had higher beginning (fall) test scores than students entering the same grade in 2001 (see Table 7). These differences were greater in grades 3 through 6 than they were in grades 7 and 8. Stated in terms of their effect size, the differences ranged from .10 at grade 3 to about .02 at grades 7 and 8. The average weighted difference across all grades was 0.72 RIT, which translated to an average weighted effect size of .05. In other words, student scores in mathematics were modestly higher in 2003, by about 5% of a standard deviation.

The differences in reading were considerably smaller than those in mathematics, with only students in grades 3 achieving a gain large enough to equate to an effect size improvement of .05 or better. The weighted average difference across all grades was only 0.19 RIT point, which translated to a weighted effect size of .01. The cumulative difference across all grades was slightly under 1 point. While this difference is statistically significant, it is not large enough to project an improvement in proficiency rates with any confidence.

Overall, although gains were made in both reading and mathematics, only the mathematics gains were large enough to project to substantive improvements in overall performance and proficiency rates over time.

Table 8 – Fall reading scores for students in the study sample								
	Fall 2001			Fall 2003			Change	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
3	26384	189.59	15.54	26487	190.3	15.22	0.72	0.05
4	24903	199.54	14.96	25127	199.81	14.86	0.28	0.02
5	28658	206.67	14.63	29083	206.84	14.38	0.18	0.01
6	27676	211.66	14.46	27854	211.78	14.86	0.12	0.01
7	27160	216.14	14.46	28383	215.98	14.75	-0.16	-0.01
8	23935	220.39	14.64	25223	220.39	14.51	0	0.00
Cumulative Difference							0.83	
Average Weighted Difference							0.19	0.01

The impact of an existing state testing program. Knowing that there was some improvement in starting RIT scores between the fall 2001 and fall 2003 testing periods, we attempted to tackle the second question. Did students who participated in state testing the prior year have significantly greater differences in their fall RIT scores than students who did not?

Table 9 shows that fall 2003 mathematics RIT scores were generally higher than fall 2001 scores for both groups, but that the difference between the two years was greater for students who had participated in their state testing program during the prior year. The average weighted gain of students who participated in state testing was .60 RIT points. The cumulative benefit to students who had participated in state testing the prior year was 4.30 RIT points over those who had not.

Although there was also some overall improvement in fall reading scores, only in grades 7 and 8 did students who participated in state testing the prior year enjoy higher starting scores than students who did not. The overall gains were not large enough to conclude that participating in a state testing program had a meaningful effect on reading scores.

These results show that schools were more effective at sustaining gains in mathematics than they were in reading. We are aware that there is a considerable body of literature that suggests reading and language development are affected by many factors beyond the classroom. The availability of books in the home, the willingness of parents to read to their children, and the language development of parents themselves, all have an effect on reading development. In mathematics, student learning and improvement may depend less on factors outside the classroom and may respond more directly to improvements in instruction. If this were the case, it may prove easier to sustain improvements in mathematics performance than it will be in reading or writing.

Table 9 – Fall mathematics scores for students in the study sample, disaggregated by whether students participated in state testing the prior year

No State Test

	Fall 2001			Fall 2003			Differences		
Grade	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Difference	Effect Size	Instructional Time
3	15408	189.42	12.29	15580	190.70	12.37	1.28	0.10	0.09
4	19746	201.43	11.44	19862	202.07	11.74	0.64	0.06	0.06
5	16239	209.88	12.99	16306	209.78	13.09	-0.09	-0.01	0.11
6	5896	216.85	13.59	5912	217.46	14.18	0.61	0.05	0.09
7	18037	224.97	15.44	18710	225.06	15.64	0.09	0.01	-0.13
8	4434	229.90	17.10	4671	229.70	16.94	-0.20	-0.01	0.09
						Cumulative Difference	2.33		
						Average Weighted Difference	0.44	0.03	

State Test Administered

	Fall 2001			Fall 2003			Differences		
Grade	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Difference	Effect Size	Instructional Time
3	14366	190.91	11.78	14385	192.12	11.93	1.20	0.10	0.15
4	7387	201.52	13.29	7350	203.31	13.19	1.79	0.13	0.17
5	15671	209.45	13.20	15995	211.19	13.04	1.74	0.13	0.07
6	23157	216.84	14.42	23220	217.70	14.88	0.85	0.06	0.11
7	9069	219.41	16.42	9137	220.07	16.18	0.65	0.04	0.10
8	16666	229.20	17.00	17454	229.59	16.74	0.39	0.02	0.00
						Cumulative Difference	6.63		
						Average Weighted Difference	1.04	0.07	

Difference in Fall 2003 Improvement

	State test not administered	State test administered	Advantage for states administering test
Grade 3	1.28	1.20	-0.08
Grade 4	0.64	1.79	1.15
Grade 5	-0.09	1.74	1.83
Grade 6	0.61	0.85	0.24
Grade 7	0.09	0.65	0.57
Grade 8	-0.20	0.39	0.59
	Cumulative Difference		4.30
	Weighted Difference		0.60

Table 10 – Fall reading scores for students in the study sample, disaggregated by whether students participated in state testing the prior year

No State Test

	Fall 2001			Fall 2003			Difference		
Grade	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Difference	Effect Size	Instructional Time
3	12785	189.33	15.61	12964	189.77	15.46	0.45	0.03	0.08
4	14845	199.88	14.31	15092	200.58	14.25	0.71	0.04	0.08
5	11098	206.86	14.35	11189	206.44	14.21	-0.43	0.03	0.08
6	6456	213.03	13.52	6406	213.14	13.91	0.10	0.00	0.01
7	18283	217.10	13.74	19167	216.44	14.13	-0.66	-0.08	-0.3
8	4441	219.80	14.75	4628	219.38	14.48	-0.42	-0.02	-0.08
						Cumulative Difference	-0.24		
						Average Weighted Difference	-0.03	0.00	

State Test Administered

	Fall 2001			Fall 2003			Difference		
Grade	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Difference	Effect Size	Instructional Time
3	13642	189.78	15.51	13569	190.75	15.02	0.96	0.06	0.06
4	10058	199.03	15.84	10035	198.66	15.66	-0.38	-0.02	0.01
5	17560	206.54	14.80	17894	207.10	14.48	0.56	0.04	-0.02
6	21220	211.24	14.71	21448	211.37	15.11	0.13	0.01	-0.02
7	8878	214.18	15.66	9216	215.02	15.91	0.84	0.05	0.01
8	19494	220.52	14.61	20595	220.61	14.51	0.09	0.01	0.06
						Cumulative Difference	2.21		
						Average Weighted Difference	0.34	0.02	

Difference between gain of students in states that administered test in the prior year

	State test not administered	State test administered	Difference
Grade 3	0.45	0.96	0.52
Grade 4	0.71	-0.38	-1.08
Grade 5	-0.43	0.56	0.98
Grade 6	0.10	0.13	0.03
Grade 7	-0.66	0.84	1.50
Grade 8	-0.42	0.09	0.51
	Cumulative Difference		0.28
	Cumulative Difference		2.5
	Weighted Difference		0.37

Growth Comparisons

The analysis of student achievement status, as measured by the fall 2001 and fall 2003 assessments showed improvement, with greater gains in mathematics than reading. Introducing fall to spring growth data allows us to triangulate these findings, by determining whether students were sustaining gains during the school year that might translate to additional long term achievement gains.

For purposes of this study we included only records in which the student posted both a fall and spring test score within a subject for either the 2001-2002 or the 2003-2004 school year. We calculated raw growth by subtracting each student's fall school from his or her respective spring score.

Based on prior studies (Northwest Evaluation Association, 2002) we have found that reading and mathematics growth in our norming sample differs significantly based on the student's grade and starting position on the RIT scale. In particular, these studies found that higher performing students tend to grow less on average than those who start at a lower level of performance.

To control for these differences we calculated a ***growth index*** statistic, which is estimated by subtracting each student's growth from the average growth of students in the 2002 norming study who started in the same grade and same ten point RIT range. A positive growth index statistic would indicate that the student grew more than a peer group who started in the same grade with about the same RIT score, while a negative number would indicate that the student grew less.

Changes in growth index scores between 2001-2002 and 2003 -2004. Tables 3 and 4 compare mathematics and reading growth for students in the study sample for the 2001-2002 and 2003-2004 school years. Growth, as reflected in the growth index numbers, declined slightly in both reading and mathematics between these two school years. In mathematics, differences ranged from -.07 (grade 5) to -.65 (grade 8) RIT, with effect sizes ranging from -.01 to -.09. In reading the differences in growth index scores ranged from -.04 (grade 4) to -.39 (grade 3), with effect sizes ranging from -.01 to 0.04.

While fall status test scores seem to have improved somewhat overall, the rate of growth clearly slipped. At least during the early stages of NCLB implementation, schools in this sample did not achieve the gains in either achievement status or academic growth that would be needed to meet the ambitious goals set by the law.

Table 11 – Mathematics growth for students in the study sample 2001-2002 v. 2003-2004							
	2001-2002		2003-2004		Pooled Standard Deviation	Difference	
	Count	Mean	Count	Mean		Change in growth	Effect Size
3	29762	-0.10	29951	-0.23	7.20	-0.14	-0.02
4	27175	0.37	27243	0.25	6.97	-0.12	-0.02
5	31958	0.15	32342	0.08	6.95	-0.07	-0.01
6	29088	-0.78	29144	-1.10	6.90	-0.32	-0.05
7	27151	-0.56	27897	-1.03	7.00	-0.47	-0.07
8	21123	-1.59	22141	-2.23	7.16	-0.65	-0.09
	Average Weighted Difference					-0.27	-0.04

Table 12 – Reading growth for students in the study sample 2001-2002 v. 2003-2004							
	2001-2002		2003-2004		Pooled Standard Deviation	Difference	
	Count	Mean	Count	Mean		Change in growth	Effect Size
3	26384	0.06	26487	-0.33	7.78	-0.39	-0.05
4	24903	-0.37	25127	-0.41	7.17	-0.04	-0.01
5	28658	-0.67	29083	-0.74	6.76	-0.07	-0.01
6	27676	-0.56	27854	-0.72	6.87	-0.16	-0.02
7	27160	-0.93	28383	-1.01	6.96	-0.08	-0.01
8	23935	-0.46	25223	-0.75	6.78	-0.29	-0.04
	Average Weighted Difference					-0.17	-0.02

Effect of State Testing Programs on Growth Index Scores

Students enrolled in a grade that participated in their state's respective testing program consistently showed greater growth in mathematics than those students enrolled in a grade in which a state test was not administered for both testing periods, although the average weighted difference for the 2003-2004 testing period was slightly smaller than that for the 2001-2002 testing period (see Table 13). For the 2001-2002 testing period, differences ranged between +.75 to 1.30 RIT growth index points with corresponding effect sizes that ranged from .00 to .19. For the 2003-2004 testing periods, the differences ranged from +.24 to 1.26 RIT growth index points, with corresponding effect sizes ranging between .01 and .18.

Once again, although the differences seem modest, they would have a substantive effect on academic achievement and student proficiency rates if sustained over time. Let's use Oregon once again as an example. Oregon currently tests students at grades 3, 5, and 8. They are required by NCLB to add testing at grades 4, 6, and 7 by the 2005-2006 school year. If the addition of the test in these years resulted in the average bump in growth index scores that was demonstrated in 2003-2004, students would experience an average increase of approximately 2 RIT points in their math scores (.75 growth index gain times 3 grades) between grades 3 and 8 by that decision alone. This

would translate to an increase from about 50% to 54% of the number of proficient students, relative to our national norm sample. By our estimate, the reading gain would translate to approximately 1 RIT point and an increase of about 2% in the number of proficient students.

In reading, students participating in a state test once again showed greater gains than those who did not, with students in grade 8 showing the greatest difference (.85 RIT in 2001-2002 and .95 RIT in 2003-2004 with effect sizes of .12 and .14). Interestingly, the smallest changes occurred in grade 3 (+.08 RIT in 2001-2002 and -.12 RIT in 2003-2004), suggesting that testing in reading may have brought more focus at the upper grades, while the addition of a state mathematics test seemed to have more effect in the lower grades. Reading has traditionally received more time and energy from teachers in the lower grades than mathematics, so it would make sense that adding emphasis to the subject by introducing a state test in mathematics at that grade might improve growth for those students. Similarly, reading has received less emphasis in grades 6 through 8, and adding a state reading test in those grades might spur more growth than it would in the lower grades where reading has always received great emphasis.

Table 13 – Mathematics growth disaggregated by school year and participation in state testing

State Test Administered
School Year 2001-2002

	No			Yes			Difference		
Grade	Count	Mean Growth Index	Std Deviation	Count	Mean Growth Index	Std Deviation	Pooled Standard Deviation	Difference	Effect Size
3	21036	-0.46	7.27	8726	0.77	7.09	7.2	1.23	0.17
4	16235	0.15	6.96	10940	0.71	6.86	6.97	0.56	0.08
5	6387	-0.86	6.69	21983	0.45	6.81	6.95	1.30	0.19
6	16743	-1.19	6.84	10011	-0.44	6.42	6.9	0.75	0.11
7	5838	-0.57	6.8	21313	-0.56	6.93	7	0.02	0.00
8	10496	-2.01	7.18	10627	-1.17	7.01	7.16	0.84	0.12
						Average Weighted Difference	**7.04**	**0.84**	**0.12**

State Test Administered
School Year 2003-2004

	No			Yes			Difference		
Grade	Count	Mean Growth Index	Std Deviation	Count	Mean Growth Index	Std Deviation	Pooled Standard Deviation	Difference	Effect Size
3	21130	-0.61	7.11	8821	0.66	7.21	7.2	1.26	0.18
4	16417	-0.1	7.07	10826	0.78	6.89	6.97	0.88	0.12
5	6494	-0.34	6.77	22178	-0.1	7.05	6.95	0.24	0.04
6	16619	-1.52	7.22	10139	-1.11	6.42	6.9	0.4	0.06
7	5924	-1.07	7.03	21973	-1.02	7.1	7	0.05	0.01
8	11055	-2.63	7.16	11086	-1.84	7.2	7.16	0.79	0.11
						Average Weighted Difference	**7.05**	**0.75**	**0.11**

Table 14 – Reading growth disaggregated by school year and participation in state testing									
State Test Administered School Year 2001-2002									
	No			Yes			Difference		
Grade	Count	Mean Growth Index	Std Deviation	Count	Mean Growth Index	Std Deviation	Pooled Standard Deviation	Difference	Effect Size
3	14707	0.02	7.93	11677	0.1	7.86	7.78	0.08	0.01
4	10926	-0.5	7.35	13977	-0.27	6.8	7.17	0.22	0.03
5	5812	-0.99	6.27	19615	-0.63	6.66	6.76	0.36	0.06
6	16452	-0.76	6.95	8893	-0.43	6.33	6.87	0.33	0.05
7	5166	-1.26	6.74	21994	-0.86	6.76	6.96	0.4	0.06
8	10757	-0.92	6.79	13178	-0.08	6.47	6.78	0.85	0.12
						Average Weighted Difference	7.11	0.35	0.05
State Test Administered School Year 2003-2004									
	No			Yes			Difference		
Grade	Count	Mean Growth Index	Std Deviation	Count	Mean Growth Index	Std Deviation	Pooled Standard Deviation	Difference	Effect Size
3	14840	-0.27	7.61	11647	-0.41	7.71	7.78	-0.14	-0.02
4	11125	-0.48	7.57	14002	-0.35	7.04	7.17	0.13	0.02
5	5906	-1	6.51	19835	-0.79	7.01	6.76	0.2	0.03
6	16576	-0.96	7.07	8865	-0.4	6.63	6.87	0.56	0.08
7	5271	-1.5	7.56	23112	-0.9	7.06	6.96	0.6	0.08
8	11374	-1.27	6.99	13849	-0.32	6.81	6.78	0.95	0.14
						Average Weighted Difference	7.11	0.37	0.05

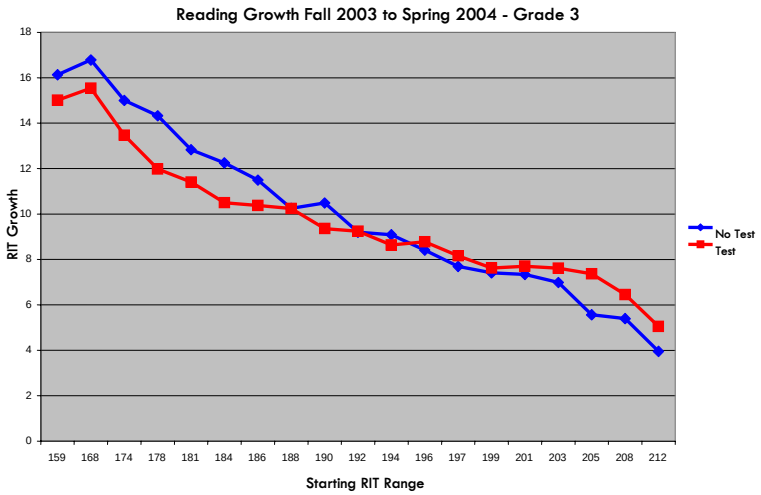
Differences in growth by starting achievement. NCLB focuses on improving the proportion of students who achieve proficiency on their state test. As a result, there is some possibility that educators may invest more of their efforts on students who are near the proficiency bar than students who are too far above or below the proficiency bar to cross it during a given school year. Thus we were also interested in whether any improvements in growth were evenly distributed. That is, if more growth occurred, did all students improve or was the improvement focused on selective groups?

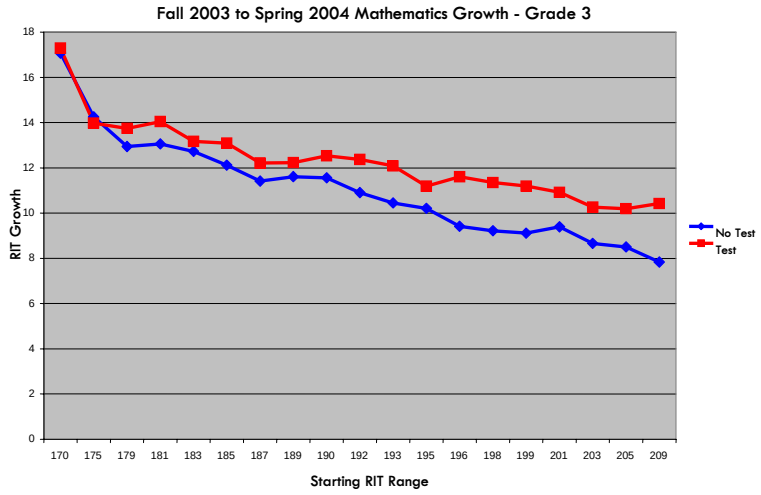
The figures on the next several pages show the fall 2003 to spring 2004 growth achieved by students based on their starting RIT score in fall of 2003. In mathematics, high performing students in grades administering their state test generally showed greater growth than high performing students in grades that did not administer a state test. This would seem to suggest that the presence of a high-stakes assessment may do more to stimulate math growth in high

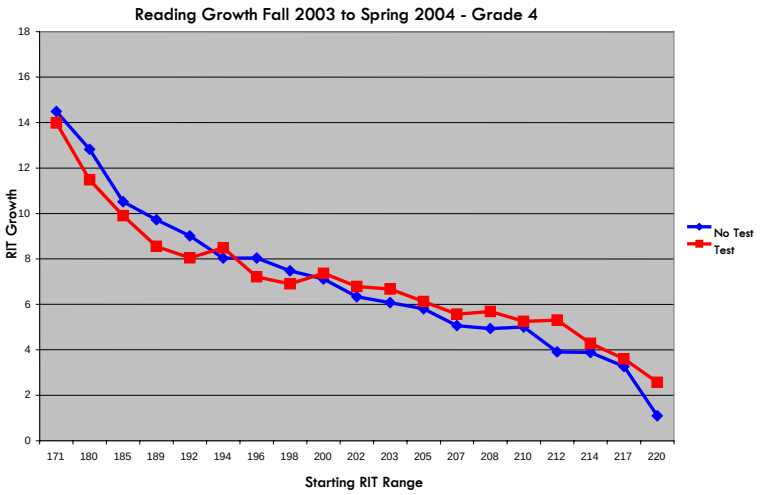
performing students than others. A similar, but slightly less pronounced pattern was visible in reading. High performing students generally showed higher growth in reading if a state test was administered in their grade.

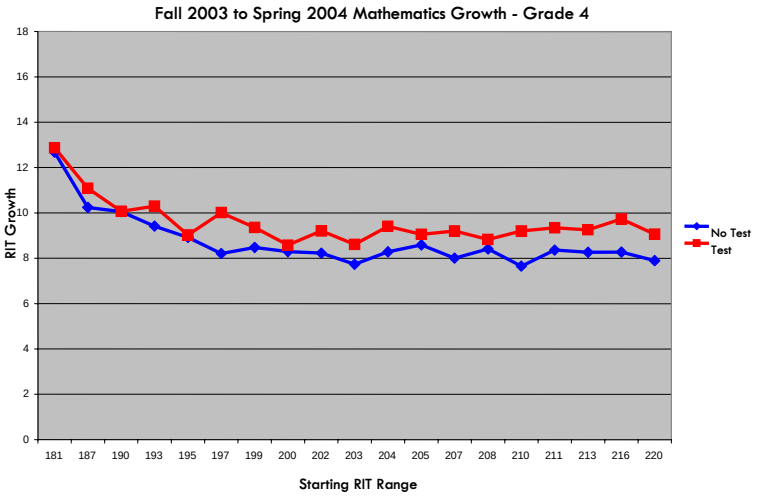
The data show that tested students at the low end of the achievement scale tend to grow less than similar students who were not tested. On the other hand, tested students at the high end of the achievement scale generally showed substantively greater growth than students who were not tested.

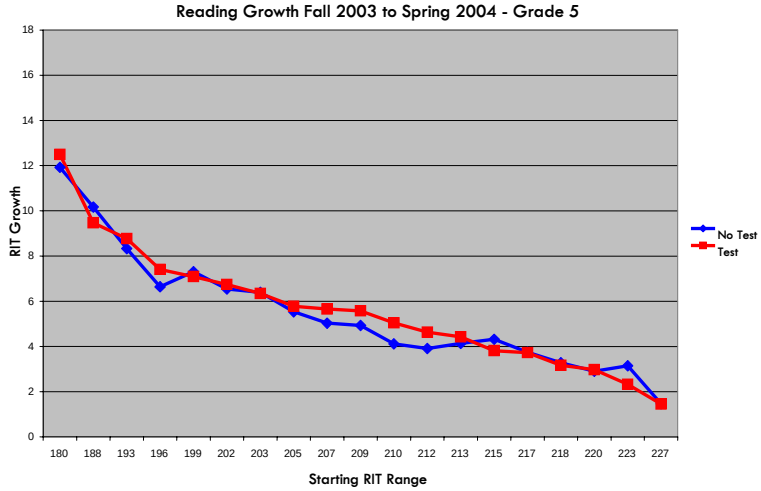
One of the goals of NCLB was to stimulate higher performance among disadvantaged students. One of the fears that critics of NCLB have voiced is that the law may lead educators to focus instruction on teaching a narrow range of basic skills and test taking techniques to assure the success of low performing students, neglecting high performing students who are very likely to reach standards. While our conclusion is tentative, this fear seems somewhat unfounded. Indeed, the introduction of testing may have actually stimulated greater improvement among high achieving students. Unfortunately, it is also not definitively established that testing has yet stimulated substantively greater growth from low performing students.

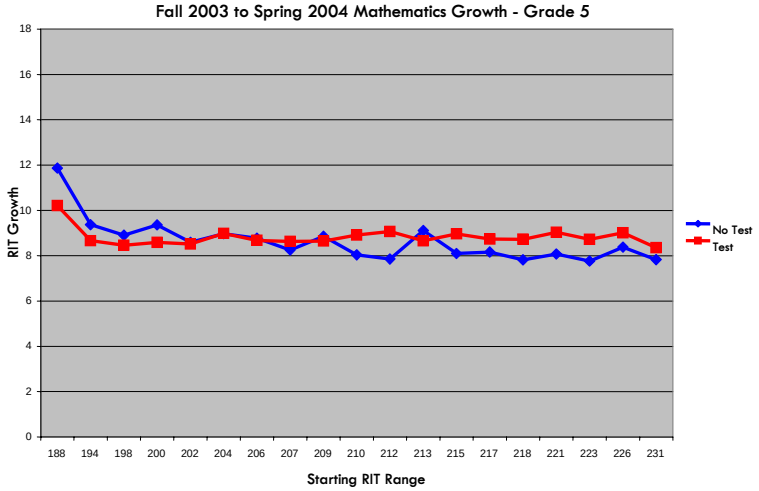


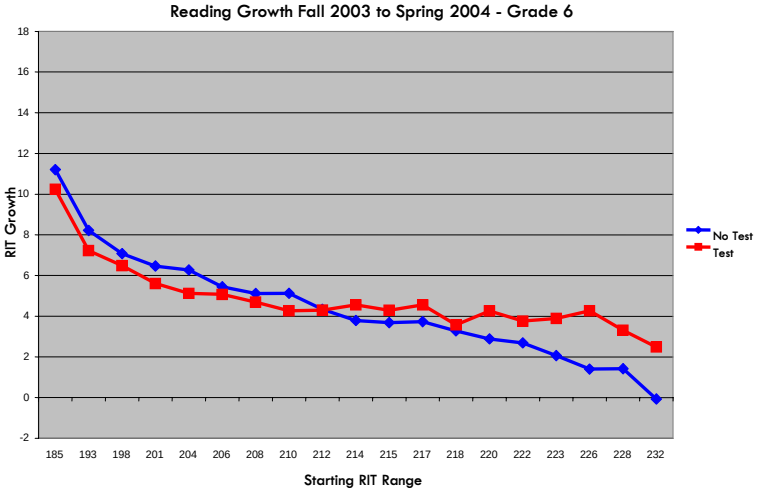


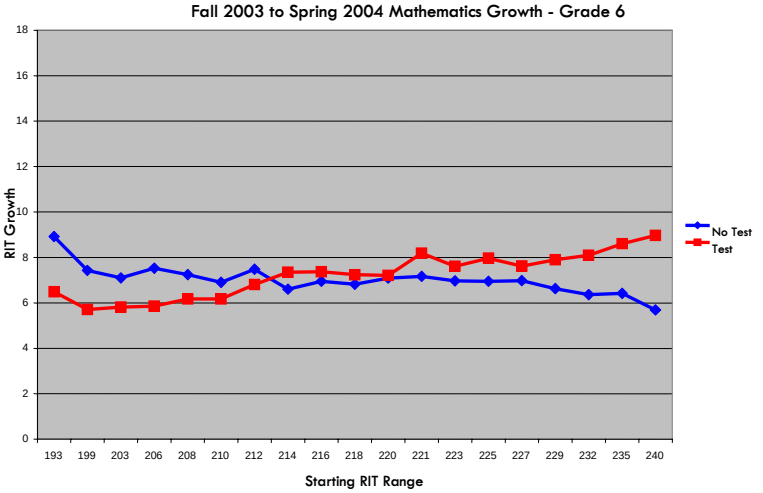

Reading Growth Fall 2003 to Spring 2004 - Grade 4
RIT Growth
Starting RIT Range
No Test
Test

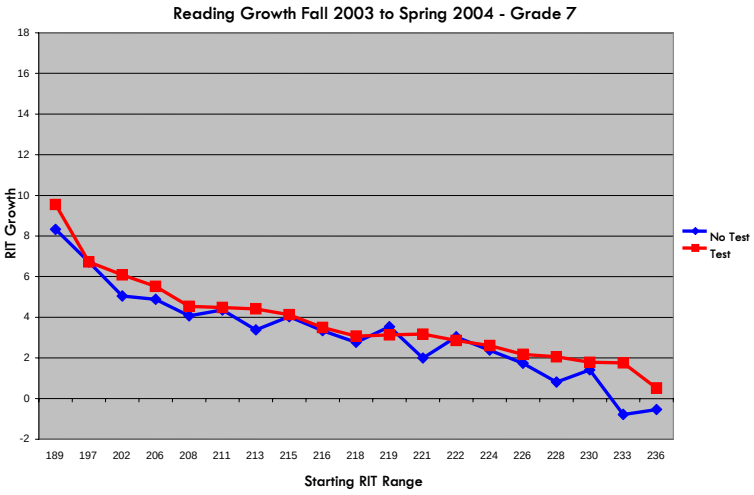

Fall 2003 to Spring 2004 Mathematics Growth - Grade 4
RIT Growth
Starting RIT Range
No Test
Test

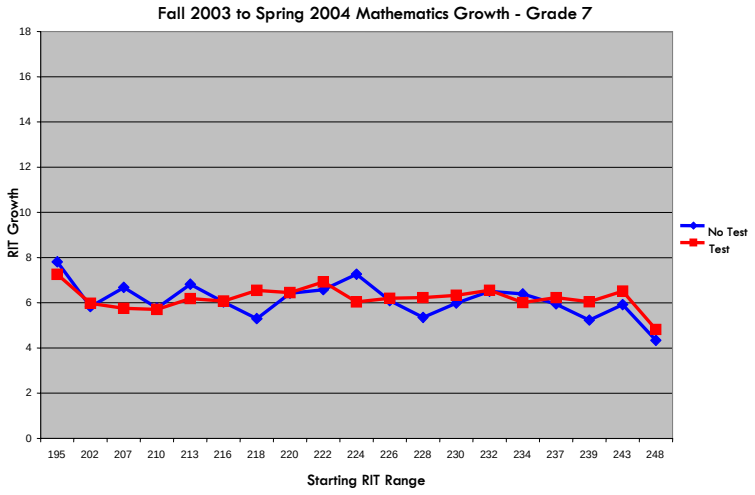

Reading Growth Fall 2003 to Spring 2004 - Grade 5
RIT Growth
Starting RIT Range
No Test
Test

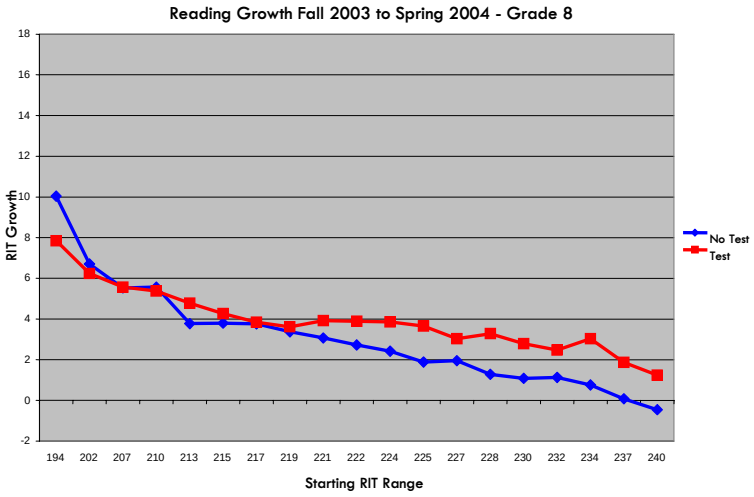

Fall 2003 to Spring 2004 Mathematics Growth - Grade 5
RIT Growth
Starting RIT Range
No Test
Test

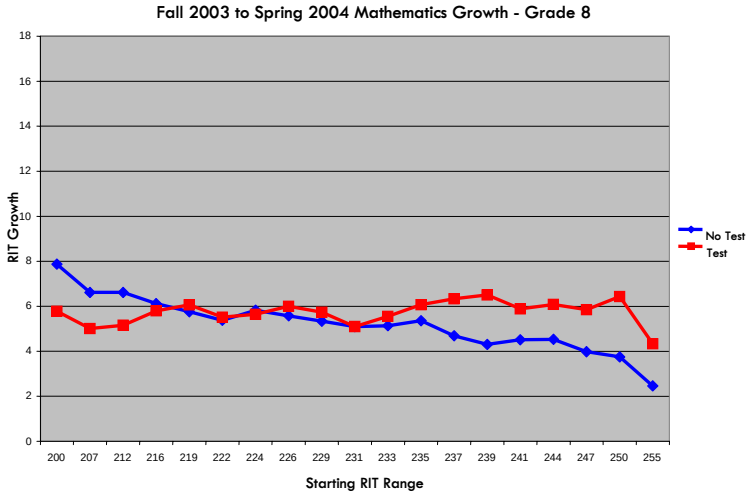

Reading Growth Fall 2003 to Spring 2004 - Grade 6
RIT Growth
Starting RIT Range
No Test
Test


Fall 2003 to Spring 2004 Mathematics Growth - Grade 6
RIT Growth
Starting RIT Range
No Test
Test


Reading Growth Fall 2003 to Spring 2004 - Grade 7
RIT Growth
18
16
14
12
10
8
6
4
2
0
-2
189 197 202 206 208 211 213 215 216 218 219 221 222 224 226 228 230 233 236
Starting RIT Range
No Test
Test


Fall 2003 to Spring 2004 Mathematics Growth - Grade 7
RIT Growth
18
16
14
12
10
8
6
4
2
0
195 202 207 210 213 216 218 220 222 224 226 228 230 232 234 237 239 243 248
Starting RIT Range
No Test
Test


Reading Growth Fall 2003 to Spring 2004 - Grade 8
RIT Growth
18
16
14
12
10
8
6
4
2
0
-2
194 202 207 210 213 215 217 219 221 222 224 225 227 228 230 232 234 237 240
Starting RIT Range
No Test
Test


Fall 2003 to Spring 2004 Mathematics Growth - Grade 8
RIT Growth
18
16
14
12
10
8
6
4
2
0
200 207 212 216 219 222 224 226 229 231 233 235 237 239 241 244 247 250 255
Starting RIT Range
No Test
Test

Impact of Both State Tests and NCLB

This analysis examined the influence of two major variables on student growth. One is the implementation of NCLB. We hypothesized that if the act has stimulated improved student performance, then a significant proportion of the variance in student's unexpected growth (growth index scores) would be explained by the school year in which students were tested. In other words, students tested after NCLB implementation (2004) should show higher growth than students tested prior to implementation (2002). The second was whether a state test was administered in the child's grade, the hypothesis being that a significant proportion of the variance in student growth index scores would be explained by whether students participated in the state testing program.

An analysis of variance was done to test the influence of these two variables on growth index scores (see Tables 7 and 8). Primarily because of the large sample size, we expected the effect of both variables to be statistically significant in both mathematics and reading. This proved to be the case. The F values indicated that the administration of a state test during a grade had a much greater effect on growth index scores than the effect associated with the start of NCLB implementation. In mathematics, for example, the F value associated with administration of a state test during the grade was roughly 9 times as great as that associated with differences stemming from the start of NCLB implementation (214.37 v. 23.29). In reading, the F value associated with administration of a state test during the grade was nearly 6 times as great as that associated with differences stemming from the start of NCLB implementation (794.28 v. 131.58).

The ANOVA results confirm observations made from the analysis of univariate statistics. More growth occurs in grades in which a state test is administered, both before and after the implementation of NCLB. The implementation of NCLB adds some additional growth. Greater gains in growth are observed in mathematics than in reading.

Beyond the findings from the univariate analysis, the ANOVA results indicate that the presence of state testing during a grade explained far more of the variance in growth index scores than whether the school year followed or preceded NCLB implementation. This would seem to suggest that the existence of a state test (whether prior to or after NCLB) has done more to stimulate additional student growth than factors that might be associated with the actual implementation of NCLB.

It should be noted that the strength of the explanatory power of the model is low. This indicates that unexpected student growth is not being controlled by the existence of state tests or by the existence of NCLB. This is to be expected, since individual student differences in interest, motivation, and time spent on learning are likely to be more powerful indicators of student growth than any external mandate or measure.

Table 15 – Results of ANOVA on growth index scores for mathematics

Source	Df	Mean Square	F	Sig.
Between subjects				
2002 v 2004 school year	1	6544.89	131.58	0.000
State tested grade	1	39506.68	794.27	0.000
2002 v 2004 school year * State tested grade	1	71.55	1.44	0.230
Subjects within groups				
Error	334971	49.74	309.65	

R Squared = .003 (Adjusted R Squared = .003)

Table 16 – Results of ANOVA on growth index scores for reading

Source	Df	Mean Square	F	Sig.
Between subjects				
2002 v 2004 school year	1	2417.49	48.48	0.000
State tested grade	1	4227.48	84.77	0.000
2002 v 2004 school year * State tested grade	1	40.97	0.82	0.365
Subjects within groups				
Error	320869	49.87		

R Squared = .000 (Adjusted R Squared = .000)

CHAPTER 5: Results: Performance and Growth by Ethnicity

Problems Surrounding the Concept of the Achievement Gap

One of the goals of NCLB is to reduce achievement gaps among a variety of subgroups in the student population. For this study, differences among groups by ethnicity are examined, since these are the groups of students in which substantive and enduring achievement gaps have been identified in the past. The term "achievement gap" has only recently come into widespread use. Until the passage of NCLB, writers would typically refer to the achievement gap as differences in performance on an academic measure between two or more demographic groups, generally ethnic groups. Thus if Anglos achieved a median score of 212 on a measure and African-Americans achieved a median score of 205, writers spoke of the achievement gap as the 7 point difference in median performance between these two groups.

The passage of NCLB has created a new operational definition for the term achievement gap. We increasingly see references in the press and literature that define the achievement gap as the difference between the percentage of students identified as proficient in two groups of interest.

This study examines whether there is evidence that these gaps started to narrow since the beginning of NCLB implementation by looking at differences in student scores prior to implementation of the act in fall 2001 and starting achievement in the fall of 2003. To facilitate the analysis, records in which the ethnic status of the students was unknown or identified as "other" were removed from different demographic groups who achieve proficiency on a measure. Thus if 75% of European-American students are proficient on their state test and 58% of Hispanic students achieve proficiency, we have a 17 percentage point achievement gap between the two groups.

Defining an achievement gap in this fashion is problematic because the size of the gap depends not only on the performance of the students, but also the position in which the bar was placed.

Consider this example. Let's assume we are measuring the athleticism of two male populations of 100 members each. One sample exercises regularly and maintains normal weight. We'll call them the "fit" sample. The other sample does not exercise at all and is an average of 30 pounds overweight. We'll call them the "fat" sample. To statistically compare the fitness of the two populations we will have them jump over a high bar and determine what percentage successfully completes the task.

For our initial experiment we decide to set the high bar at 6 feet 0 inches. Two members of the fit sample successfully jumped over the bar (2%) and no members of the fat sample made the leap. The initial experiment shows a 2 point achievement gap between the two groups. We are surprised that the gap is this narrow, and decide that it might be because we set the bar too high to be meaningful. So we decide to reconfigure the experiment.

For our second experiment we decide to set the high bar at a more reasonable level, 2 feet 0 inches. 95% of the fit sample was able to cross the bar at this level and 91% of the fat sample was also successful. So the second experiment found a slightly larger achievement gap (4 percentage points) between the two groups. Again we are surprised that the gap is this narrow and decide that we may have set the bar too low to be meaningful. In Goldilocks fashion, therefore, we search for the bar height that is just right.

Thus a third experiment. This time we set the bar for 4 feet 0 inches. We found that about 50% of the fit population was able to successfully leap the 4 foot bar. However only 20% of the fat population successfully completed the jump. This time we found an achievement gap of 30 points, which led us to conclude that fit people have an easier time completing fitness exercises than fat people.

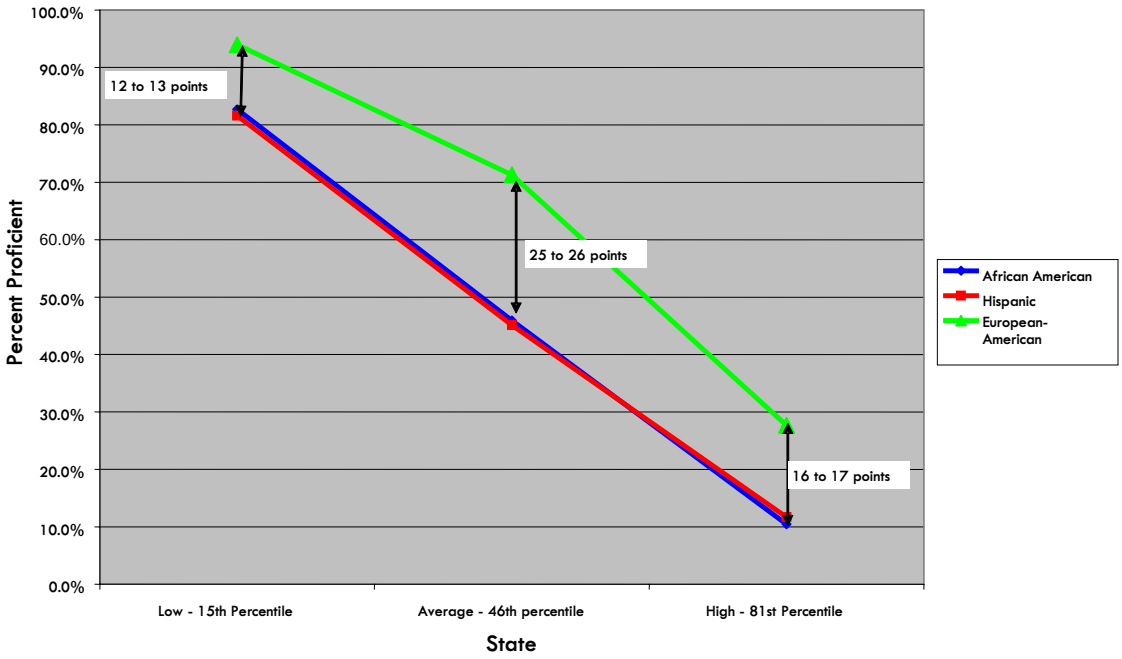
In our example, the size of the achievement gap was not simply a reflection of the difference between the two groups. It was also a function of where we chose to place the bar. NCLB requires that states establish a proficiency standard for their students, but leaves decisions about where to set that standard to the states. In a prior study it was found that states have used this autonomy to set standards that vary greatly in their difficulty (Kingsbury, et. al.; 2003). This study also summarized the results of over 16 prior studies conducted by our organization to determine the scores on the RIT scale that aligned with proficient performance on a variety of state tests.

The large variance in achievement standards has an effect on the size of achievement gaps. Figures ***112 and 113*** show an example of the problem in grade 5 mathematics and reading. For purposes of the illustration we selected the state from our original study with the lowest NCLB proficiency cut score, the state near the median for all states studied, and the state with the highest cut score. We calculated the proportion of the entire spring 2004 fifth grade sample population that would be proficient based on application of each standard and compared the three largest ethnic groups from the population.

The examples show that when we use the lowest bar, we not only report a higher proficiency rate (that's a given) but we also report a substantially lower achievement gap. The effect is similar when we use the highest standard. Applying a more difficult standard again lowers the reported achievement gap. The largest gap is reported when we apply a standard that is closer to the middle of the achievement distribution.

This problem makes it difficult to talk meaningfully about how schools in two different states perform relative to achieving true equity across ethnic groups, because the size of the achievement gap is partly a function of the academic height of the bar. When achievement gaps are reported as the difference in percent of students who are proficient, two states with equal gaps in mean performance between two ethnic groups could easily report differences in the percent of students proficient that are greater than 10 points, if one maintains a standard near the 50th percentile and the other maintains a standard that is closer to the 20th or 80th percentile.

Percent of the Grade 5 Sample Population that is Proficient Relative to Three States' Standards - Spring 2004 Mathematics



For that reason, when we refer to achievement gaps in this study, we will refer differences in mean scores and not differences in the percent of students who achieve proficiency. This is the simplest and most accurate way to report such differences.

We first attempted to determine whether there is evidence that these gaps started to narrow since the beginning of NCLB implementation by looking at differences in student scores prior to implementation of the act in fall 2001 and starting achievement in the fall of 2003. To facilitate the analysis, we removed records in which the ethnic status of the students was unknown or identified as "other."

Performance Comparisons

As in virtually all other studies, European-American and Asian students performed better than Hispanic, African-American, and Native American students on both the mathematics and reading measures (see Tables 13 and 14).

In general, we found that average mathematics scores of the fall 2003 group had improved over the average scores of the fall 2001 group. The weighted differences between groups ranged from improvements of about .7 (Asian) to 2.0 RIT points (African-American), with the effect sizes ranging from .05 to .14. African-American, Hispanic, and Native American students posted the

largest gains. In reading, the fall 2003 group also made gains, although the gains were smaller than those found in mathematics. The weighted differences ranged from about .4 (European-American) to 1.2 RIT points (African-American), with effect size differences ranging from about .02 to .08. African-American, Hispanic, and Asian students posted the largest gains.

Table 17 – Fall mathematics scores disaggregated by ethnicity

	Fall 2001			Fall 2003			Change	
	Count	Mean	Std Dev	Count	Mean	Std Dev	Difference	Effect Size
European-American	19215	192.01	11.51	19214	193.23	11.67	1.22	0.11
Hispanic	3995	184.00	11.84	4347	186.24	11.91	2.24	0.19
African-American	1052	186.26	10.86	1078	188.40	11.23	2.15	0.20
Asian	1073	191.09	13.31	1331	192.82	13.77	1.73	0.13
Native American	660	185.96	11.90	633	186.81	12.40	0.85	0.07
Grade 4								
	Count	Mean	Std Deviation	Count	Mean	Std Dev	Difference	Effect Size
European-American	18277	203.41	10.99	18356	204.36	11.44	0.95	0.09
Hispanic	3714	194.34	12.19	4239	196.53	12.38	2.19	0.18
African-American	968	195.51	11.34	1047	197.00	12.15	1.49	0.13
Asian	864	202.13	12.74	960	203.12	13.82	0.99	0.08
Native American	675	192.61	12.86	800	195.43	11.55	2.82	0.22
Grade 5								
	Count	Mean	Std Deviation	Count	Mean	Std Dev	Difference	Effect Size
European-American	22595	211.39	12.40	22270	212.15	12.39	0.76	0.06
Hispanic	4378	201.88	12.62	5008	204.10	13.22	2.22	0.18
African-American	1016	203.52	12.30	1099	205.74	12.04	2.22	0.18
Asian	994	211.72	13.56	1189	211.43	14.73	-0.29	-0.02
Native American	617	201.23	13.25	656	203.81	13.02	2.58	0.19
Grade 6								
	Count	Mean	Std Deviation	Count	Mean	Std Dev	Difference	Effect Size
European-American	21000	218.92	13.34	20028	219.91	13.76	0.99	0.07
Hispanic	3971	208.07	13.48	4406	208.61	14.34	0.53	0.04
African-American	779	210.64	13.51	825	212.20	13.39	1.55	0.11
Asian	781	219.52	15.60	884	221.33	17.19	1.81	0.12
Native American	624	207.50	14.16	721	205.90	14.11	-1.60	-0.11
Grade 7								
	Count	Mean	Std Deviation	Count	Mean	Std Dev	Difference	Effect Size
European-American	17911	225.95	14.64	18384	226.14	14.90	0.18	0.01
Hispanic	3701	212.17	15.13	4556	213.73	15.72	1.55	0.10
African-American	831	214.46	15.63	921	217.21	16.25	2.75	0.18
Asian	641	225.59	17.03	753	224.17	16.74	-1.42	-0.08
Native American	628	208.76	16.39	680	211.57	16.02	2.80	0.17

Grade 8								
	Count	Mean	Std Deviation	Count	Mean	Std Dev	Difference	Effect Size
European-American	14250	231.83	15.77	16029	232.65	15.59	0.83	0.05
Hispanic	2747	217.33	16.34	3371	219.00	16.23	1.68	0.10
African-American	360	221.63	18.44	457	220.19	18.36	-1.44	-0.08
Asian	412	232.44	17.04	576	228.24	17.89	-4.20	-0.25
Native American	746	219.87	16.93	548	219.32	16.09	-0.54	-0.03
Weighted Differences								
	Average Weighted Difference	Pooled Standard Deviation	Weighted Effect Size Difference					
European-American	0.85	13.93	0.06					
Hispanic	1.66	13.91	0.12					
African-American	1.96	13.64	0.14					
Asian	0.70	13.61	0.05					
Native American	1.41	14.13	0.10					

Table 18 – Fall reading scores disaggregated by ethnicity

	Fall 2001			Fall 2003			Change	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	17277	192.11	14.77	17336	192.87	14.40	0.76	0.05
Hispanic	3534	181.14	15.60	3870	182.51	15.33	1.36	0.09
African-American	860	184.58	14.56	898	186.36	14.37	1.79	0.12
Asian	757	192.24	14.30	879	192.72	14.90	0.47	0.03
Native American	589	185.11	15.93	556	185.01	16.06	-0.10	-0.01
Grade 4								
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	16994	202.01	13.80	17140	202.30	13.87	0.28	0.02
Hispanic	3360	190.18	15.68	3872	191.53	15.09	1.35	0.09
African-American	912	193.05	15.50	972	194.24	14.99	1.19	0.08
Asian	636	201.65	13.31	666	202.29	14.02	0.64	0.05
Native American	554	188.11	16.29	655	191.11	15.85	3.00	0.18
Grade 5								
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	20691	208.73	13.61	20431	208.99	13.36	0.26	0.02
Hispanic	3589	197.14	15.56	4188	198.54	15.06	1.40	0.09
African-American	921	201.29	14.31	982	202.25	13.24	0.97	0.07
Asian	644	209.83	12.58	733	210.02	13.88	0.18	0.01
Native American	562	195.81	15.89	604	196.84	16.53	1.03	0.06

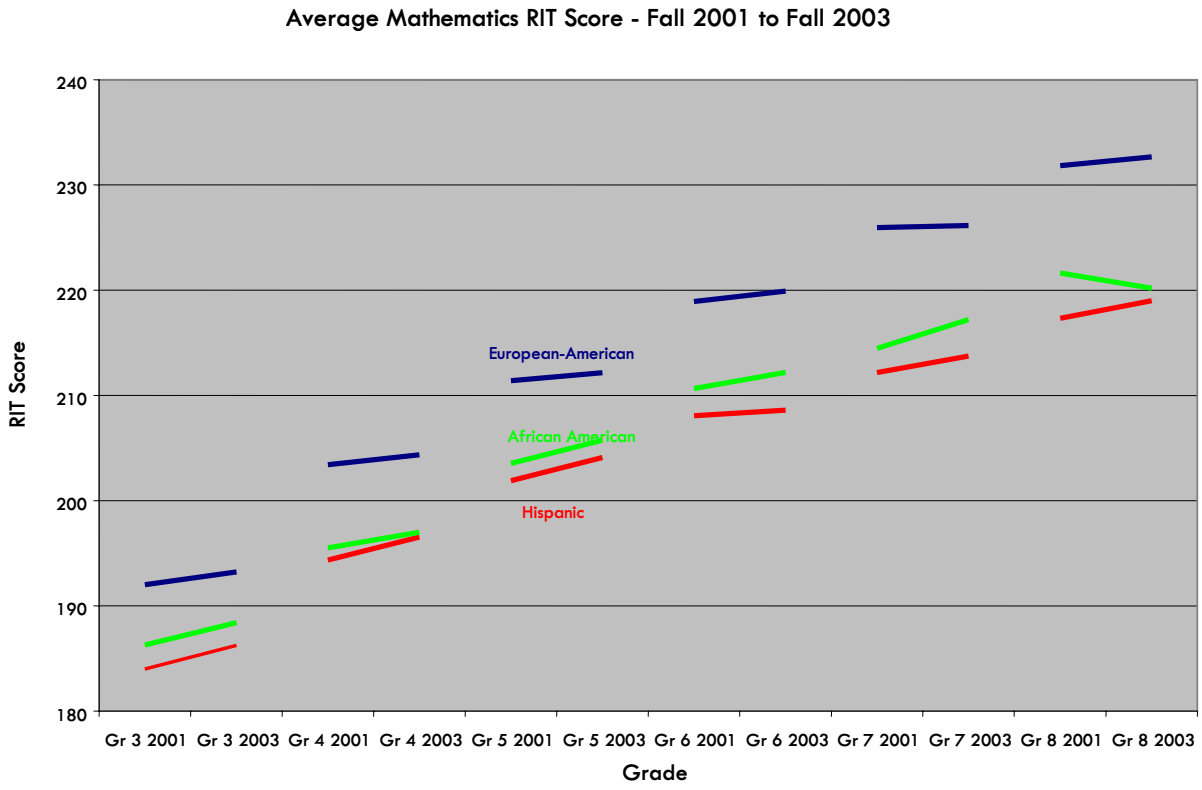
Grade 6								
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	20604	214.00	13.27	19619	214.38	13.52	0.38	0.03
Hispanic	3823	202.41	15.28	4132	201.81	16.09	-0.61	-0.04
African-American	771	206.19	13.43	796	206.81	14.14	0.62	0.05
Asian	586	215.06	13.91	657	216.47	13.13	1.40	0.10
Native American	590	201.76	15.00	721	200.26	15.25	-1.51	-0.10
Grade 7								
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	18532	218.97	12.82	19750	218.55	13.30	-0.42	-0.03
Hispanic	3615	205.97	15.58	4411	206.93	16.05	0.97	0.06
African-American	931	209.12	14.04	1000	210.41	14.84	1.29	0.09
Asian	655	215.77	14.84	766	217.40	15.40	1.62	0.11
Native American	568	204.36	15.84	613	204.60	15.59	0.24	0.02
Grade 8								
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Difference	Effect Size
European-American	16712	222.86	13.03	18561	222.89	12.97	0.03	0.00
Hispanic	3145	209.74	16.27	3903	210.99	16.25	1.26	0.08
African-American	414	214.21	16.27	508	214.00	14.90	-0.20	-0.01
Asian	580	221.52	13.99	676	221.60	15.41	0.08	0.01
Native American	708	211.14	16.24	462	210.88	14.78	-0.27	-0.02
Weighted Differences								
	Average Weighted Difference	Pooled Standard Deviation	Weighted Effect Size Difference					
European-American	0.36	14.76	0.02					
Hispanic	0.89	14.77	0.06					
African-American	1.17	14.79	0.08					
Asian	0.87	14.80	0.06					
Native American	0.60	14.77	0.04					

Figure 1 shows the change in the achievement gap for the subject which showed the greatest change, mathematics.

While these differences may not seem large, they do represent some progress in narrowing the achievement gap in a relatively short period of time. Figure 1 depicts the difference between the fall 2001 and fall 2003 mathematics scores for three ethnic groups, Anglos, African-Americans and

Hispanics. It shows that African-Americans narrowed the achievement gap at every grade but grade 8, while Hispanics narrowed the achievement gap at every grade but grade 6.

Figure 1 -



The next analysis looked for any differences between situations in which state tests had been administered and those in which a state test had not been administered. In mathematics, we found that fall 2001 to fall 2003 improvement was greater for all ethnic groups among students enrolled in grades that administered their respective state test (see Table 19). The weighted improvement between fall 2001 and fall 2003 ranged from about .7 (European-American, Asian, and Native American students) to about 1.6 RIT (Hispanic) in mathematics, with effect sizes ranging between .05 and .12. In reading the differences were again smaller. The weighted average difference between the fall 2001 and fall 2003 means ranged from about -.2 (Native American) to +.8 RIT (African-American), with effect sizes ranging between -.01 and +.05 (see Table 20).

Table 19 – Average weighted differences in fall 2001-fall 2003 mathematics results disaggregated by ethnicity and whether a state test was administered in the grade (complete results available in Appendix A)						
State Test	**No**			**Yes**		
Ethnic Group	**Average Weighted Difference**	**Pooled Standard Deviation**	**Weighted Effect Size**	**Average Weighted Difference**	**Pooled Standard Deviation**	**Weighted Effect Size**
European-American	0.72	13.63	0.05	0.92	13.63	0.07
Hispanic	1.57	13.15	0.12	1.85	13.15	0.14
African-American	0.91	13.24	0.07	2.48	13.24	0.19
Asian	0.67	12.93	0.05	0.06	12.93	0.00
Native American	0.71	14.01	0.05	2.68	14.01	0.19

Table 20 – Average weighted differences in fall 2001-fall 2003 reading results disaggregated by ethnicity and whether a state test was administered in the grade (complete results available in Appendix A)						
State Test	**No**			**Yes**		
Ethnic Group	**Average Weighted Difference**	**Pooled Standard Deviation**	**Weighted Effect Size**	**Average Weighted Difference**	**Pooled Standard Deviation**	**Weighted Effect Size**
European-American	0.27	14.84	0.02	0.20	14.84	0.01
Hispanic	0.66	14.82	0.04	0.98	14.82	0.07
African-American	0.79	14.89	0.05	1.19	14.89	0.08
Asian	0.17	14.92	0.01	0.93	14.92	0.06
Native American	-0.20	14.80	-0.01	2.13	14.80	0.14

On the whole, evidence indicated that small but substantive gains in achievement were made by African-Americans, Hispanics, and Native Americans that would serve to reduce achievement gaps between these groups and European-American and Asian students. For traditionally disadvantaged groups, the gains were greater when those students were enrolled in grades that participated in their respective state testing programs.

Growth Comparisons

The next set of analyses investigated whether and to what extent ethnic groups differed in the fall to spring growth that was unexpected, using the growth index score. Because lower performing students generally grow more, controlling for starting position on the scale with the growth index score better assures that comparisons of progress are reasonable.

Figures 2 and 3 depict the growth index scores for the 2003-2004 school year in mathematics and reading. European-American and Asian students achieved greater growth than their Hispanic, Native American, and African-American counterparts for both the 2001-2002 and 2003-2004 school years. The depiction shows in grade 4, for example, that European-American students showed 2 points greater RIT growth than their African-American counterparts in mathematics, about 1.5 points more growth than Native American students, and about 1 point of growth more than Hispanic students. Differences of approximately this magnitude hold through all grades tested in mathematics and similar differences in growth can be seen in reading.

Figure 2

Fall 2003 to Spring 2004 Growth Index Scores Disaggregated by Ethnic Group - Mathematics

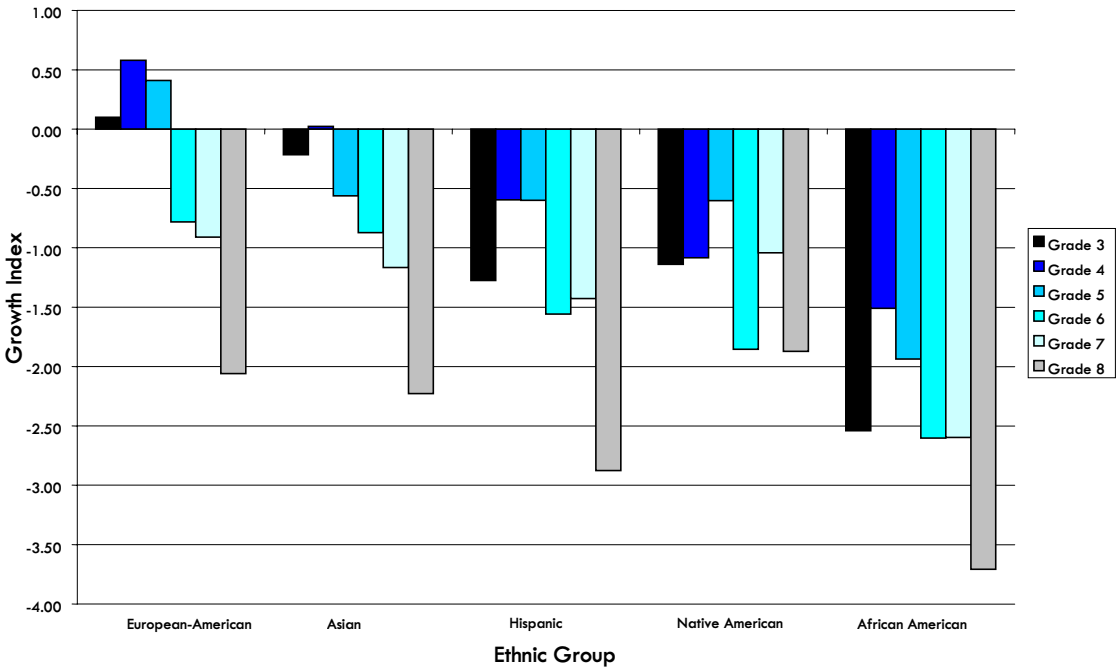
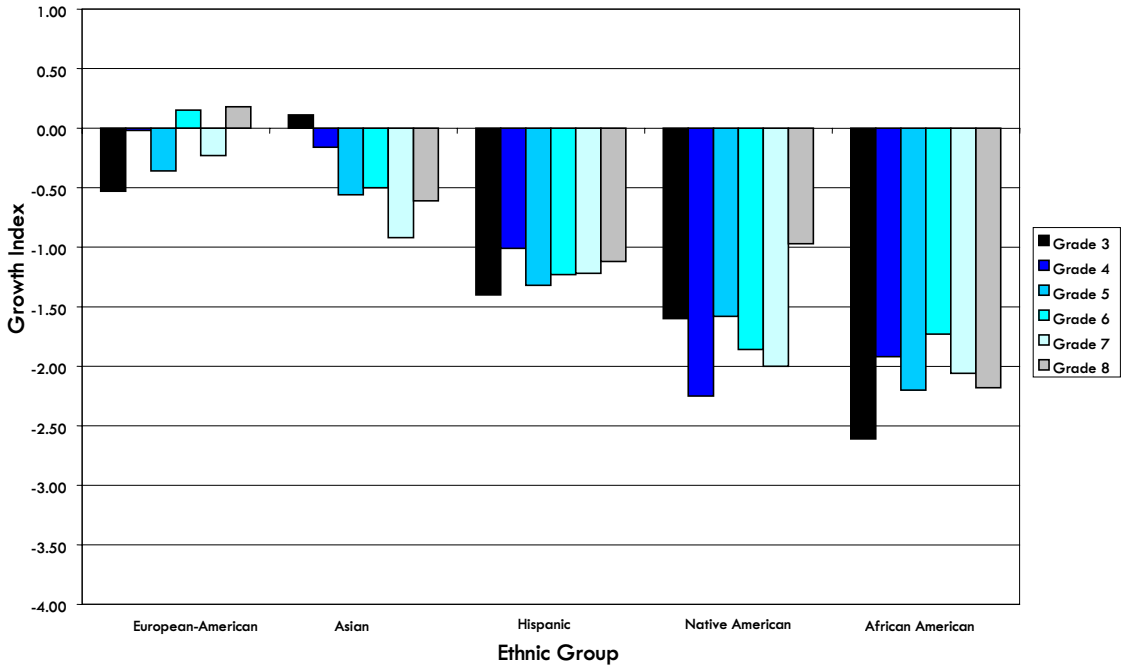


Figure 3

Fall 2003 to Spring 2004 Growth Index Score Disaggregated by Ethnic Group - Reading



Thus, while the data indicate the status achievement gap might be narrowing, we found a substantive growth gap. It is useful to provide an example of the manner in which this can occur.

> Assume that Robert is a hypothetical example that illustrates how. Assume we have Robert, an European-American fifth grader who achieved a fall reading score of 216, which is equivalent to the 75th percentile for all fifth graders. Marissa is an European-American fifth grader who achieved a fall score of 197, which is equivalent to the 25th percentile. Terrance is an African-American student who achieved the same fall score as Marissa, 197.
>
> The European-American students, Robert and Marissa achieve the growth that is typical for students who start in this grade at their position on the scale. In Robert's case the typical growth would be 5.4 points according to NWEA RIT scale norms (NWEA, 2002), which would raise his score from 216 to 221. In Marissa's case, the typical growth is considerably higher, 10.7 points, because lower performing students generally show greater growth. Thus her score would improve from 197 to 208. Next, let's assume that Terrance's growth reflects this average minus the difference between European-American and African-American students in this grade, which is about 2 points. That means that his growth would be 9 points (10.7 rounded

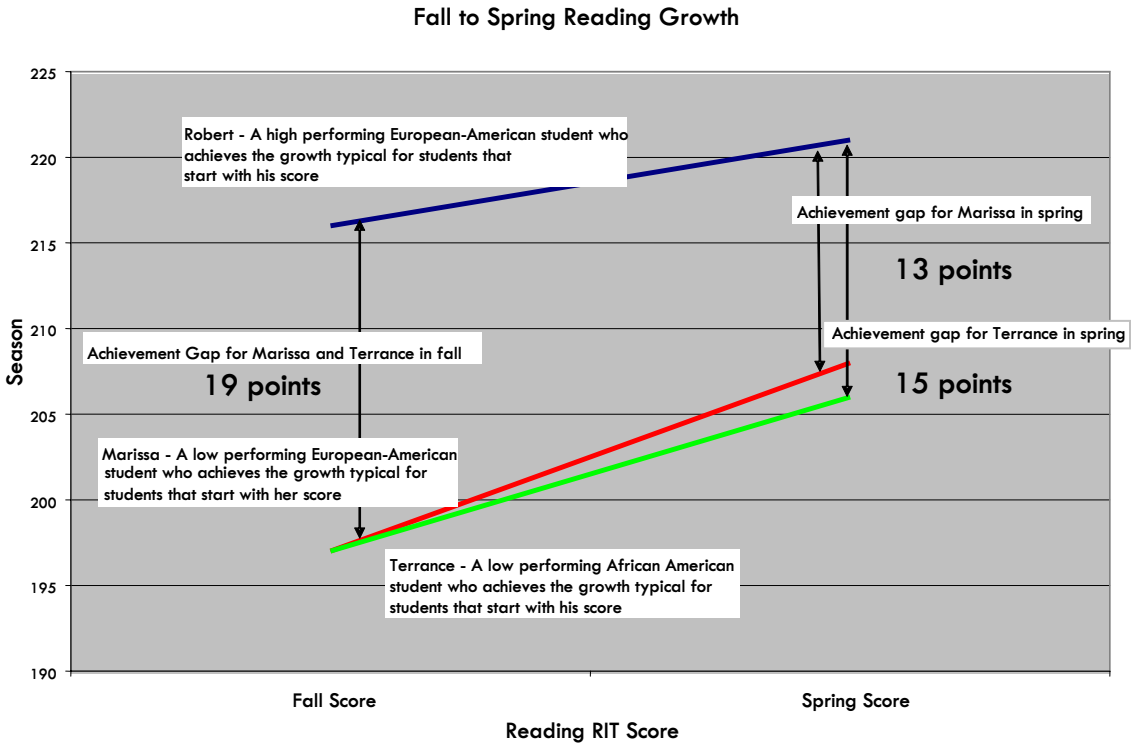
minus the 2 point difference between European-American and African-American students), raising his scale score to 206.

The results, depicted in Figure 4, show that Terrance, who started 19 points behind Robert, now trails him by only 15 points. He has indeed closed some of his achievement gap. But Marissa, the European-American student, closed her gap with Robert to only 13 points because, once you control for starting position on the scale, European-American students achieve about 2 points greater growth than African-American students. In other words, while both students narrowed their achievement gap, the European-American student narrowed her's more.

Extending this example to a population, African-American students generally might narrow their achievement gap relative to European-American students, primarily because more African-American students start with lower scores. But a population of European-American students who started with the same scores would have narrowed it more. Thus narrowing gaps in achievement is not a complete solution to achieving educational equity.

Figure 4

An example illustrating how an achievement gap might be reduced in an environment in which minority students grow less.



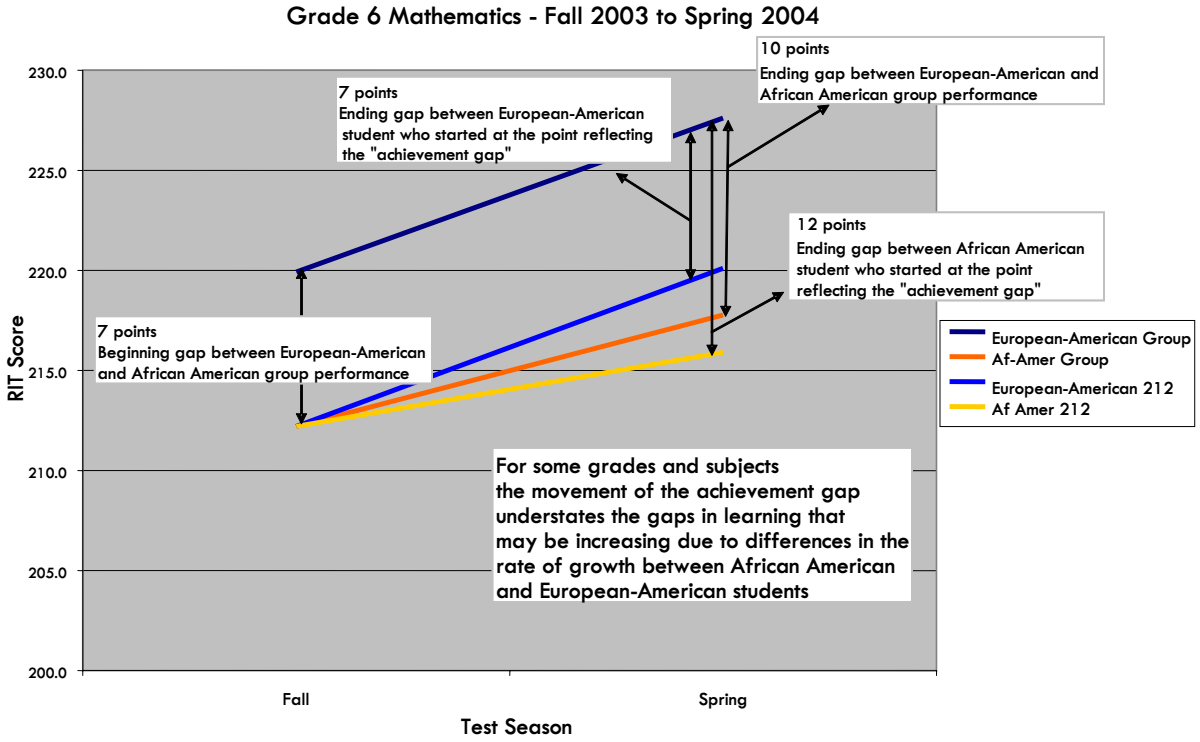
The example was corroborated by other evidence in this study that showed gaps in achievement do not fully reflect gaps in growth that may thwart efforts to achieve equity. Figure 5 shows an example from the sixth grade mathematics data that illustrates the issue.

In sixth grade mathematics, the actual achievement gap between the European-American and African-American students in our sample increased from a gap of 7 points to 10 points between the fall of 2003 and spring of 2004. But the gap in group scores understates the actual differences that are observed with groups of European-American and African-American students who started at the same point on the RIT scale.

The next analysis looked at growth of African-American and European-American students who started at the RIT score that would represent the low point of the achievement gap, that is the average African-American student's performance in Fall 2003 (212 RIT). By spring of 2004, European-American students who started with this score maintained the original achievement gap, 7 points, relative to the average of the European-American group. African-Americans who started with this score, however, fell even further behind than the average of their group would have suggested. While the average African-American student in sixth grade mathematics had fallen 10 points behind their European-American peers by Spring of 2004, African-American students who started with a score of 212, the point representing the original gap, fell 12 points behind. In other words, while the African-American group lost 2 points relative to their European-American peers, the average African-American student who started at 212, lost a full 5 points relative to other European-American students who started with that score.

This example is not an anomaly. At every grade in mathematics, Hispanic and African-American students lost ground relative to European-American peers when each group started with the score representing the low end of the reported achievement gap. In several grades, the change in achievement gaps reported between the European-American group and the Hispanic and African-American groups substantively understated the gap in performance that emerged when we evaluated progress made by students who started at the RIT score representing the low end of the gap.

Figure 5



The next analysis examines changes in growth index scores between the year prior to implementation of NCLB and the 2003-2004 school year (see Tables 17 and 18). In reading, all ethnic groups showed declines in growth index scores that ranged from -.13 (Hispanic students) to -.75 (Asian) students. These effect size changes were generally small, with the exception of the Asian student sample. In mathematics, only African-American students showed slight improvement in growth index scores (+.08) over this time period with other groups showing declines that ranged from -.10 to -.39 RIT. None of the effect size changes in reading would be considered large.

Table 21 – Summary of changes in mathematics growth index scores by ethnic group (complete results available in Appendix B)							
	Fall 2001-Spring 2002		Fall 2003-Spring 2004		2001-2002 to 2003-2004 change in growth		
Ethnic Group	Count	Mean Growth Index	Count	Mean Growth Index	Difference	Pooled Standard Deviation	Effect Size
European-American	113248	-0.10	114281	-0.38	-0.28	7.02	-0.04
Hispanic	22506	-1.18	25927	-1.32	-0.13	7.02	-0.02
African-American	5006	-1.97	5427	-2.34	-0.36	7.02	-0.05
Asian	4765	0.07	5693	-0.68	-0.75	7.03	-0.11
Native American	3950	-1.08	4038	-1.25	-0.17	7.03	-0.02

Table 22 – Summary of changes in reading growth index scores by ethnic group (complete results available in Appendix B)							
	Fall 2001-Spring 2002		Fall 2003-Spring 2004		2001-2002 to 2003-2004 change in growth		
Ethnic Group	Count	Mean Growth Index	Count	Mean Growth Index	Difference	Pooled Standard Deviation	Effect Size
European-American	110810	-0.27	112837	-0.46	-0.19	7.04	-0.03
Hispanic	21066	-1.12	24376	-1.22	-0.10	7.05	-0.01
African-American	4809	-2.20	5156	-2.12	0.08	7.08	0.01
Asian	3858	0.24	4377	-0.16	-0.39	7.08	-0.06
Native American	3571	-1.52	3611	-1.75	-0.23	7.05	-0.03

Finally, results were disaggregated to investigate whether students in the various ethnic groups responded differently when their grade participated in state testing (see Tables 19 and 20). In mathematics, we generally found no substantive differences with the exception of African-Americans. African-American students who were enrolled in grades that participated in the state test showed about .5 less RIT growth than students of their ethnic group who were enrolled in grades that did not participate. The effect size of the difference was -.07. In reading, we found that Asian and Native American students enrolled in grades that participated in state testing achieved .76 and .64 greater RIT growth than their peer group that did not participate. The effect sizes of the difference were .09 and .10 respectively.

Table 23 – Summary of changes in mathematics growth index scores disaggregated by ethnic group and enrollment in a grade in which the state test is administered (complete results available in Appendix C)

	Fall 2001- Spring 2002				Fall 2003 – Spring 2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Change in Growth	Effect Size	Change in Growth	Effect Size	Difference in Growth	Effect Size
European-American	55719	-0.56	57529	0.34	56926	-0.81	57355	0.04	7.04	-0.25	-0.04	-0.30	-0.04	-0.05	-0.01
Hispanic	5919	-1.57	16587	-1.05	6786	-1.51	19141	-1.25	7.05	0.05	0.01	-0.20	-0.03	-0.25	-0.04
African-American	2125	-2.32	2881	-1.72	2422	-2.38	3005	-2.30	7.05	-0.07	-0.01	-0.58	-0.08	-0.51	-0.07
Asian	1677	-0.47	3088	0.36	1986	-1.27	3707	-0.36	7.07	-0.80	-0.11	-0.72	-0.10	0.08	0.01
Native American	2282	-1.27	1668	-0.82	2226	-1.51	1812	-0.94	7.02	-0.24	-0.03	-0.12	-0.02	0.12	0.02

Table 24 – Summary of changes in reading growth index scores disaggregated by ethnic group and enrollment in a grade in which the state test is administered (complete results available in Appendix C)

	Fall 2001- Spring 2002				Fall 2003 – Spring 2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Change in Growth	Effect Size	Change in Growth	Effect Size	Difference in Growth	Effect Size
European-American	47988	-0.50	49601	-0.73	62822	-0.10	63236	-0.24	7.11	0.40	0.06	0.49	0.07	0.09	0.01
Hispanic	3750	-0.81	4459	-0.84	17316	-1.18	19917	-1.30	7.09	-0.37	-0.05	-0.46	-0.07	-0.10	-0.01
African-American	1931	-1.83	2149	-1.95	2878	-2.45	3007	-2.24	7.17	-0.62	-0.09	-0.29	-0.04	0.33	0.05
Asian	1020	0.41	1168	-0.54	2838	0.17	3209	-0.02	7.22	-0.24	-0.03	0.52	0.07	0.76	0.10
Native American	1977	-1.62	1902	-2.17	1594	-1.40	1709	-1.30	7.08	0.23	0.03	0.87	0.12	0.64	0.09

These analyses of univariate statistics show, therefore, that African-American, Hispanic, and Native American students demonstrate lower growth then their European-American and Asian counterparts. Factors such as the implementation of NCLB and the administration of the state test do not seem to have had a differential effect.

Impact of Both State Tests and NCLB

To further examine combined effects of NCLB implementation and state testing on the growth of students in different ethnic groups, another analysis of variance was conducted. This ANOVA introduced ethnicity to the two factors considered in our original analysis, the year in which testing occurred and whether the state test was administered in that particular grade.

The mathematics model (see Table 25) proved a better predictor of growth index scores than the reading model, although both models passed tests of significance (mathematics R squared=.009, reading R squared = .005). Both models indicated that ethnic status of the students was by far the variable with the largest influence on growth (mathematics $F(4,304,821)=370.75$, reading $F(4,294,451)=230.06$). The mathematics model found that the administration of state testing was the variable with second greatest influence ($F(1,304,821)=125.22$) and the school year in which the test was administered the third ($F(1,304,821)=36.29$). The reading model found the interaction between ethnicity and the presence of a state test in the grade to be the second greatest influence ($F(4,294,451)=30.79$) and the school year in which the test was administered the third ($F(1,294,451)=11.51$).

These ANOVA results seem to bear out our tentative conclusions from the univariate analysis. They show that ethnicity has the largest influence on growth index figures. Factors such as growth before and after NCLB and whether a state test was administered to the student had a statistically significant effect on the model, but F values associated with these are far smaller than those associated with the student's ethnic group.

Table 25 – Results of ANOVA on growth index scores for mathematics with ethnicity

	European-American	Hispanic	African-American	Asian	Native American
Ethnic Group	227529	48433	10433	10458	7988
Source	Df	Mean Square	F	Sig.	
Between Subjects					
Ethnicity	4	18285.92	370.75	0.00	
School Year 2001-02 v 2003-04	1	1789.81	36.29	0.00	
State Test Administered	1	6176.23	125.22	0.00	
Ethnic * School Year	4	239.39	4.85	0.00	
Ethnic * State Test	4	653.98	13.26	0.00	
School Year * State Test	1	64.46	1.31	0.25	
Ethnic * School Year * State Test	4	60.38	1.22	0.30	
Error	**304821**	**49.32**			

R Squared = .009 (Adjusted R Squared = .009)

Table 26 – Results of ANOVA on growth index scores for reading with ethnicity

	European-American	Hispanic	African-American	Asian	Native American
Ethnic Group	223647	45442	9965	8235	7182
Source	Df	Mean Square	F	Sig.	
Between Subjects					
Ethnicity	4	11362.80	230.06	0.00	
School Year 2001-02 v 2003-04	1	568.59	11.51	0.00	
State Test Administered	1	38.67	0.78	0.38	
Ethnic * School Year	4	113.12	2.29	0.06	
Ethnic * State Test	4	1520.85	30.79	0.00	
School Year * State Test	1	414.74	8.40	0.00	
Ethnic * School Year * State Test	4	99.97	2.02	0.09	
Error	**294451**	**49.39**			

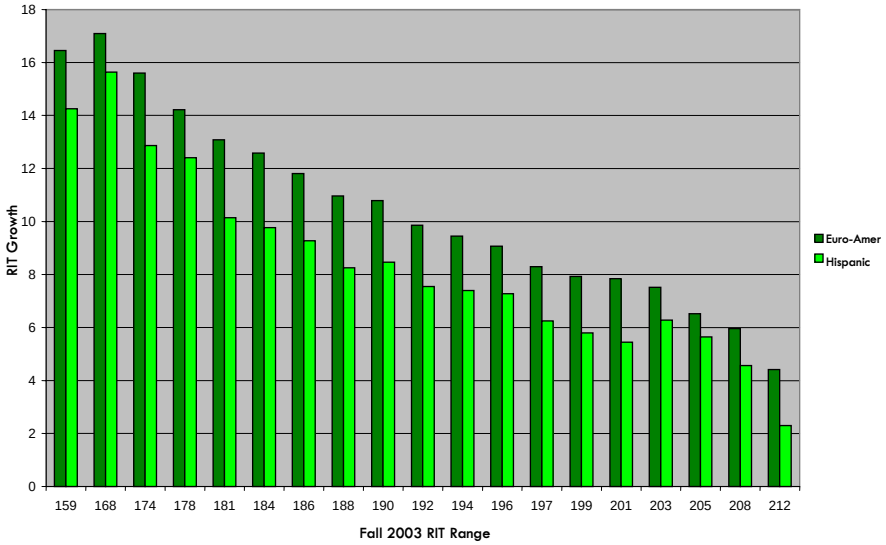
R Squared = .005 (Adjusted R Squared = .005)

Growth by Ethnic Group and Initial Score

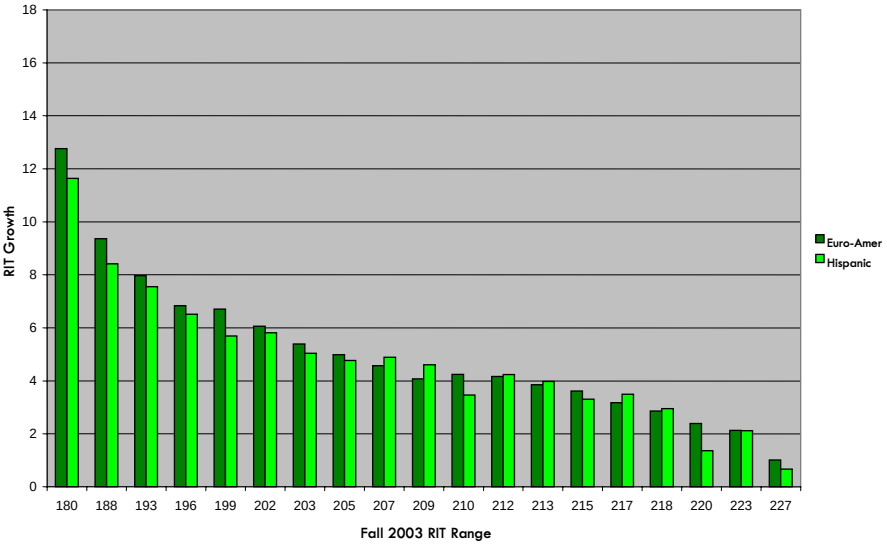
The next analysis investigates whether growth differences among ethnic groups would be even across the RIT scale. For this analysis we limited our comparisons to European-American and Hispanic students, since these two groups had large numbers of students in each grade and starting RIT range. This analysis identifies students who differ in ethnicity but have the same RIT score in the fall and asks whether they are expected to grow the same amount.

The figures on the next two pages compare the average growth of European-American and Hispanic students grouped by fall RIT score. For 50 of 57 comparisons, Hispanic students with the same initial score as their European-American peers grew less from fall to spring. This is exactly the case identified in the hypothetical example above. While the groups may be getting slightly closer together, Hispanic students tend to grow less than European-American students who have the same starting point. In general, the Hispanic students in the middle of the achievement distribution seemed to grow most like their European-American peers. For the lower-performing Hispanic students, every comparison indicated that European-American students with the same initial achievement level estimate were likely to grow more than comparable Hispanic students.

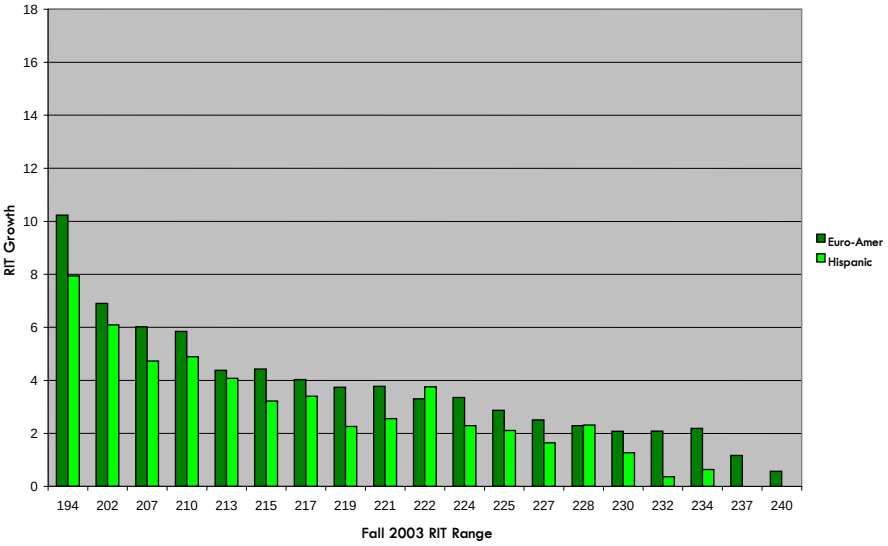
Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 3 Reading

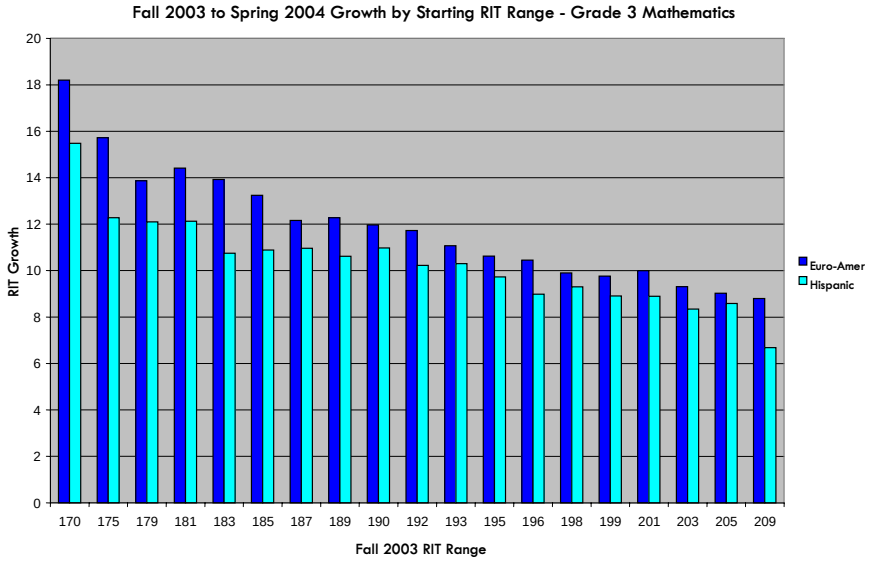


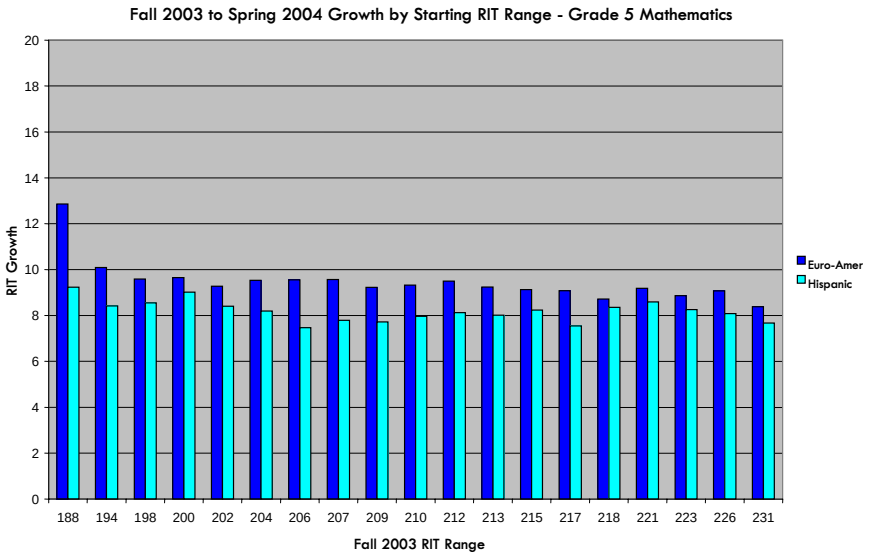
Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 5 Reading

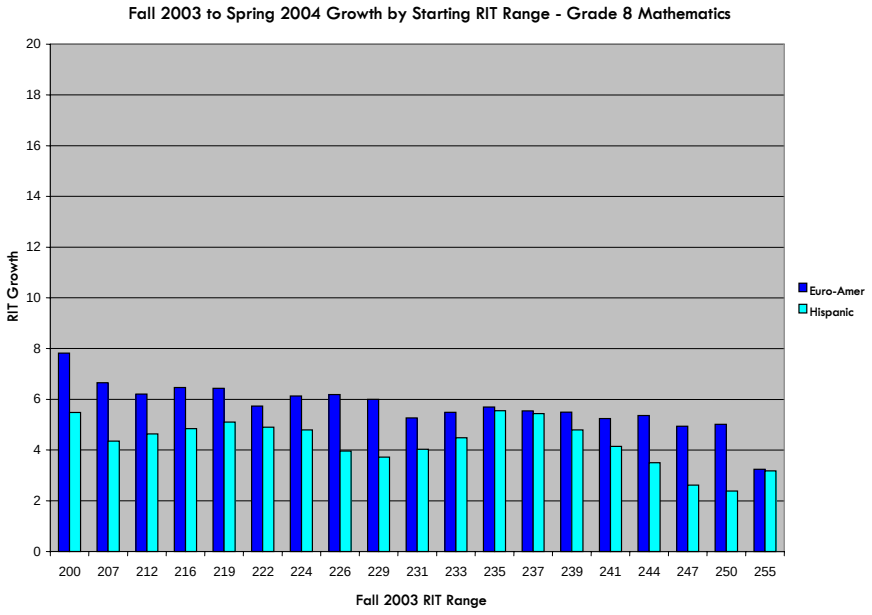


Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 8 Reading




Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 3 Mathematics
RIT Growth
0
2
4
6
8
10
12
14
16
18
20
170 175 179 181 183 185 187 189 190 192 193 195 196 198 199 201 203 205 209
Fall 2003 RIT Range
Euro-Amer
Hispanic


Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 5 Mathematics
RIT Growth
0
2
4
6
8
10
12
14
16
18
20
188 194 198 200 202 204 206 207 209 210 212 213 215 217 218 221 223 226 231
Fall 2003 RIT Range
Euro-Amer
Hispanic


Fall 2003 to Spring 2004 Growth by Starting RIT Range - Grade 8 Mathematics
RIT Growth
0
2
4
6
8
10
12
14
16
18
20
200 207 212 216 219 222 224 226 229 231 233 235 237 239 241 244 247 250 255
Fall 2003 RIT Range
Euro-Amer
Hispanic

CHAPTER 6: Conclusions and Discussion

The primary findings of the study include the following:

- Mathematics and reading scores have improved over the past two years under NCLB.
- Student growth scores have decreased since NCLB was implemented.
- Students in grades with state tests have higher achievement and growth than students who are not.
- Changes in performance in mathematics are greater than those in reading since NCLB was implemented.
- Studies in this area that use lower-stakes assessments to measure improvements in learning may have a greater percentage of unmotivated students.
- Student growth in every ethnic group has decreased slightly since NCLB was implemented.
- Growth of Hispanic students in every grade and subject area tends to be lower than the growth of European-American students with exactly the same initial score.

Mathematics and Reading Scores have Improved Under NCLB

The primary public policy goal of NCLB is to assure that all students in grades 3 through 8 (and one high school grade) are proficient performers in reading and mathematics by 2014. States have reported varying rates of progress toward this goal on their own tests. One of our objectives was to evaluate the rate of progress toward this goal using a common assessment across a multi-state population.

The differences in performance scores observed in this study would indicate that school systems have achieved some improvement in performance since the beginning of NCLB. In mathematics, the improvement in performance was substantial. Fall 2003 results were better than fall 2001 results in every grade with a weighted average improvement of .72 RIT points.

The cumulative difference in performance across grades was over 4 RIT points. One way to interpret the magnitude of this difference is to attempt to frame it in terms of the improvement it might create in proficiency rates, which has become the most common statistic used to represent student performance.

Using a hypothetical example, we can see how the amount of achievement change seen in the study might affect the percentage of students being identified as proficient. If we use the 50th percentile in grade 8 to represent a proficiency standard, the mathematics proficiency level would be a RIT score of 235 (approximately equivalent to the proficiency standard in place in Oregon for

grade 8, based on Kingsbury et. al, 2003). The reading proficiency level would be a RIT score of 225 (approximately equivalent to the proficiency standard used in Arizona for grade 8).

If the NWEA norm group were to sustain an improvement in their mathematics RIT score of 4 points between third and eighth grade, students currently scoring between 231 (the 41st percentile) and 234 would show enough improvement to cross the proficiency bar. This would translate to about a 9 percentage point improvement in proficiency. This would represent a good start toward attaining the ambitious NCLB goals.

The improvement in reading was very slight. The average weighted difference in scores between the 2001 and 2003 was only .14 RIT points and the cumulative difference was less than one point. These differences would not result in substantial improvement in the number of students being identified as proficient.

Student Growth Scores have Decreased Under NCLB

While performance seems to have improved, the rate of student growth actually declined between the school year prior to NCLB implementation (fall 2001 – spring 2002) and the fall 2003 – spring 2004 school year. In mathematics, the average weighted difference was -.25 RIT points and in reading the average difference was -.17 RIT points. While these differences are slight, they indicate that teachers are not yet achieving the gains in learning during the school year that will be needed to sustain large improvements in performance for the future.

This finding is very similar to the results of other studies that have attempted to use results from third party assessments, such as NAEP, to monitor and evaluate improvement in student performance since the implementation of NCLB. These studies consistently report that the rate of improvement in student performance on third party assessments is less than the rate of improvement reported on state assessments. Other researchers have also found that the rate of improvement on both third party assessments and state tests is not sufficient to create any reasonable prospect that all students will achieve proficiency by 2014.

That said, there is evidence of improvement, particularly in mathematics, that would suggest educators have continued to make steady progress toward improving student scores. The gains in performance that we've cited, if sustained over time, would result in significant increases in the number of students reaching the proficiency bar.

Students in Grades with State Tests have Higher Achievement and Growth

Nearly all states had statewide assessment programs in place prior to the implementation of the No Child Left Behind act. A few states tested all students in grades 3 through 8, but most implemented testing for students in two or three selected grades. One of the core policy objectives of NCLB was to expand the testing of students in grades through 8 to all states. The assumption behind this objective is that expanded accountability would serve as an impetus to improved achievement in all grades.

We found that students enrolled in grades that were included in state testing programs showed larger improvements in performance than students who did not. These differences were more substantive in mathematics than they were in reading. In mathematics, the cumulative gain for students who participated in state testing was 2.50 RIT points beyond those achieved by students who did not participate. This gain, if sustained over time, would translate to a 4 to 6 percentage point improvement in proficiency based on the standards employed in the hypothetical example above.

Once again the improvements in reading that were attributable to participation to state testing were smaller, with an average weighted difference of only .10 RIT points and a cumulative difference of only .28 RIT points. In grades 3 through 6, the overall performance of students participating in their state test declined slightly in performance relative to the non-tested proportion of the population. Students in grades 7 and 8 who were tested, however, performed substantively better, with a difference of 1.15 RIT in grade 7 and .51 RIT in grade 8.

Students participating in state testing programs also fared considerably better on the growth measurement than those who did not. In mathematics, these gains would amount to a 3 to 4 point improvement in average RIT performance between third and eighth grade, assuming they were sustained cumulatively. The difference in growth was substantive for reading as well, with students participating in testing programs gaining an average of more than 2 RIT points over students who did not.

It was anticipated that most of the beneficial effects of state testing would fall on either low performing students or on those students who perform near the proficiency bar. The reasoning was that these non-proficient or nearly-proficient students would receive the most attention from educators because the consequences of No Child Left Behind are centered on improving the achievement of this group. The results actually showed that the introduction of state tests may have most benefited students at the higher end of the performance continuum. Depending on the grade and subject tested, the added growth enjoyed by high performing students ranged between about 0 and 2 RIT points.

This data strongly supports the concept that the testing of students in the elementary grades leads to meaningful gains in performance. Most of these gains seem to be attributed to the presence of a state test. It is not clear from our data whether the particular consequences associated with the implementation of testing in an NCLB framework provides any additional benefit, nor did we attempt to determine whether the introduction of testing had greater (or less) benefit in states which imposed higher-stakes consequences on students and teachers. What we can say with some confidence is that the presence of a test that measures, monitors, and reports student achievement provides some impetus to improved learning.

Changes in Mathematics are Greater than Those in Reading Under NCLB

In general, effect size differences in both performance and growth were larger in mathematics than they were in reading. For example, the overall effect size improvement in fall 2001 to fall 2003 mathematics performance was .05, while the improvement in reading performance was only

.01. The average weighted decline in mathematics growth (effect size = .04) was also larger than that in reading (effect size =.02). Virtually every analysis in this study resulted in larger effect sixes for mathematics than for reading.

It seems possible then, that mathematics achievement may be more responsive to changes in curriculum and instruction than reading. NCLB holds educators accountable for assuring that all students are proficient in both reading and mathematics. It is in the national interest to cultivate a citizenry that is well educated in these two disciplines. As a practical matter, however, there are reasons to expect that it might be easier to achieve large improvements in mathematics performance than reading because of the way the disciplines are structured and delivered.

One aspect of this structure is that students develop most of their skill in mathematics through a curriculum pursued in the classroom. This curriculum is fairly well defined and sequenced and while parents may occasionally help their kids through a difficult algebra problem or drill their children on the times tables, this supplements and does not replace the work done in the classroom.

This gives schools two obvious points of leverage that can be used to improve mathematics learning. First, if what needs to be known in mathematics is defined and sequenced, schools have an easier time focusing improvement efforts because the path to improvement is knowable. Second, since most of what is to be learned can be accomplished in the classroom or through well-designed homework, success in mathematics may be less dependent on what goes on inside the home.

The same may not be true in reading. The skills required to develop as a reader are not as clearly defined and sequenced as those in mathematics. Reading improvement is also more dependent on qualities and skills that may require development outside the classroom. Vocabulary development, for example, is critical to developing one's reading power, and most vocabulary development comes from reading that is pursued outside the classroom setting. Students who come from homes with few reading materials and have non-English speaking parents or parents who do not read and use language well themselves, are missing ingredients that may be critical to the support of rapid reading development.

If improvement in mathematics comes more rapidly than improvement in reading, it is doubtful that an implementation of sanctions will assure that all students will eventually get what is needed to reach proficiency. If success is more difficult to achieve in some domains than others, we need to ask whether educators and policy makers are prepared to recognize that sanctions alone may not solve the problem. Assuming that improved reading achievement needs more than just classroom instruction, the solutions rest far beyond the bounds of this study, in the realm of environmental change to enhance reading improvement.

Student Growth in Every Ethnic Group has Decreased Under NCLB

Evidence here indicates substantive gaps in performance between European-American and Asian students and their Hispanic, African-American, and Native American counterparts. But we also

found evidence of a decline in the achievement gap. All ethnic groups showed improvement in performance between fall 2001 and fall 2003. For African-American, Native American, and Hispanic students these improvements were relatively large. These gains were large enough to cause modest declines in the achievement gap between these students and their European-American and Asian counterparts.

Unfortunately, while the achievement gap may have narrowed, the fall to spring growth of Hispanic, African-American, and Native American students in our sample fell far short of the growth achieved by other students. These differences in growth were quite large. For example, the difference in mathematics and reading growth between European-American and African-American students averaged more than 2 RIT points in each grade.

How could Hispanic, African-American, and Native American students achieve reductions in the achievement gap while showing lower growth than European-American students? The answer is that low achieving children generally show greater growth than high achieving children, which would generally cause gaps between low and high achievers to narrow over time. Thus low achieving minority children were closing the achievement gap because their growth was greater than that of higher achieving children.

It is important that public policy doesn't define equity as merely closing the gap between the low performers and the middle. That's why public policy should place more emphasis on closing gaps in growth among students than gaps in achievement. While closing the gap between low performers and the middle would be a step in the direction of greater equity, the product of true equity would not be equal student outcomes. It is one step in the direction of justice when a minority student, perhaps the son or daughter of a single parent trying to raise a family on a very limited income, achieves a level of proficiency that allows him or her to earn a living wage and pursue a happier life. It may be a bigger step toward justice when a particularly talented minority student, the son or daughter of a single parent trying to raise a family on the same limited income, is afforded the kinds of opportunities that assure their talent will truly blossom to its fullest potential.

Growth of Hispanic Students is Lower than Growth of Comparable European-American Students

Despite modestly narrowing the achievement gap, low achieving Hispanic children grew less than European-American children who started with the same score. The gap in growth between low achieving Hispanic and European-American children, for example, is close to 2 RIT points in third grade mathematics and between 1.5 and 2 RIT points for most students in eighth grade. Thus if one goal is to help low achieving students close their gap relative to high achievers, low achieving European-American children are closing that gap more effectively than Hispanic, African-American, and Native American children.

The disadvantage in growth experienced by Hispanic, African-American and Native American children is large enough that it will eventually thwart efforts to close the achievement gap. But closing an achievement gap should not be our only concern. We should also be concerned when

high performing minority students, students who have closed this gap and already perform at a level equal to or beyond that of their European-American peers, achieve less growth then European-American or Asian students. The gaps in growth that we've found are pervasive, they are not limited to low performing students, and they contribute to conditions that make it difficult for many minority students to achieve the level of success that reflects their true capabilities.

Conclusion to the 2005 Study

It is very early to identify the extent to which NCLB will influence educational change in the future.

Two of the positive trends at this point include the following:

- State-level tests tend to improve observed achievement, and therefore increasing the number of grades in which they are given may improve achievement more.
- There is evidence that NCLB has improved student achievement since its adoption (although this effect is much smaller than the testing effect).

Two of the worrisome elements at this point are that:

- If change in achievement of the magnitude seen so far continues, it won't bring schools close to the requirement of 100% proficiency by 2014.
- Students in ethnic groups that have shown achievement gaps in the past grow less under NCLB, and may grow less than comparable European-American students.

During next year's study in this series, the authors will watch these positive and negative trends, and add additional evidence concerning the effectiveness of NCLB. Given the number of aspects involved in the NCLB legislation, the number of methods that states have used to implement the federal requirements, and the potential for the federal government to change regulations or even the law itself, it should be an interesting few years.

REFERENCES

Amrein, A. & Berliner, D. (2002a). High-stakes testing, uncertainty and student learning Educational Policy Analysis Archives, 10(18). Retrieved March 8, 2005 from http://epaa.asu.edu/epaa/v10n18/

Amrein, A. & Berliner, D. (2002b). The impact of high-stakes tests on student academic performance. Education Policy Research Unit. Education Policy Studies Laboratory. Retrieved March 15, 2005 from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211-126-EPRU.pdf

Braun, H. (2004). Reconsidering the impact of high-stakes testing. Educational Policy Analysis Archives, 12(1). Retrieved March 8, 2005 from http://epaa.asu.edu/epaa/c12n1/

Campbell, J. R., Hombo C.M., & Mazzeo, J. (2000). *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance.* Washington, D.C.: National Center for Education Statistics. Retrieved March 12, 2005 from http://nces.ed.gov/nationsreportcard/pdf/main1999/2000469.pdf

Center for Ecucation Policy (2005). *From the capital to the classroom: Year 3 of the No Child Left Behind Act.* Washington, D.C.: Author. Retrieved March 28, 2005 from http://www.cep-dc.org/pubs/nclby3/press/cep-nclby3_21Mar2005.pdf

Chubb, J., Linn, R., Haycock, K. & Wiener, R. (2005). Do we need to repair the monument? Education Next, Spring, 2005. http://www.educationnext.org/20052/8.html

Education Trust. (2004). Measured progress: Achievement rises and gaps narrow, but too slowly. Washington, D.C.: Author. http://www2.edtrust.org/edtrust/images/MeasuredProgress.doc.pdf

Goldschmidt, P. (2004). *Models for school accountability and program evaluation.* Presentation at the Reidy Interactive Lecture Series: Incorporating measures of student growth into state accountability systems, October, 2004, Nashua, NH. Retrieved from http://www.nciea.org/publications/RILS_PG04.pdf

Grissmer, D., Flanagan, A., Kawata, J. & Williamson, S. (2002). Improving student achievement: What state NAEP test scores tell us. Santa Monica, CA: RAND Corporation. http://www.rand.org/publications/MR/MR924/

Hanushek, E., & Raymond, M. (2004). *School Accountability and the Black-White Test Score Gap.* Paper prepared for "50 Years After Brown, What Have We Achieved and What Remains to Be Done?", Harvard University, April 23-24, 2004.

Ingebo, G. (1997). Probability in the Measure of Achievement. Chicago, IL: MESA Press.

Kingsbury, G. G. (2002, April). Comparison of ALT and MAP Scores. Presented to the American Educational Research Association annual meeting New Orleans, LA.

Kingsbury, G. G. (2003, April). A long-term study of the stability of item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Kingsbury, G. G., Cronin, J. C., Hauser, C., and Houser, R. L. (2003, December). The State of State Standards: Research Investigating Proficiency Levels in Fourteen States. Portland, OR; NWEA.

Linn, R. L. (2003). *Accountability: Responsibility and reasonable expectations.* Educational Researcher, 32(7), 3-13.

Linn, R. (2004). *Rethinking the No Child Left Behind Accountability System.* Paper prepared for a forum on No Child Left Behind sponsored by the Center for Education Policy, Washington, DC, July 28, 2004.

McCombs, J.S., Kirby, S. N., Barney, H. D. & Magee, S.J. (2004) *Achieving State and National Literacy Goals, a Long Uphill Road: A Report to Carnegie Corporation of New York.* Santa Monica, CA: RAND Corporation. Retrieved March 12, 2005 from http://www.rand.org/pubs/technical_reports/2004/RAND_TR180.pdf

National Center for Education Statistics. (2005). *The Nation's Report Card.* Retrieved March 16 from http://nces.ed.gov/nationsreportcard/

National Conference of State Legislatures. (2005). *Task force on No Child Left Behind Final Report.* Denver; Author.

Northwest Evaluation Association (2002). RIT Scale Norm. Portland, OR; Author.

Northwest Evaluation Association (2003). Technical Manual. Portland, OR; Author.

Packer, J. (2004). *No Child Left Behind and adequate yearly progress fundamental flaws: A forecast for failure.* Paper presented at the Center on Education Policy Forum on Ideas to Improve the Accountability Provisions Under the No Child Left Behind Act, July 28, 2004

Roderick, M. & Engel, M. (2001). The grasshopper and the Ant: Motivational Responses of Low-achieving students to high-stakes testing. Educational Evaluation and Policy Analysis, 23(3). 197-222.

Shields, P.. Esch, C., Lash, A., Padilla, C. & Woodworth, K. (2004). *Evaluation of Title I accountability systems and school improvement efforts (TASSIE): First year findings.* A report prepared for the U.S. Department of Education by SRI International. Retrieved March 8, 2005 from http://www.ed.gov/rschstat/eval/disadv/tassie1/

Appendix A

Table A1 - Fall mathematics scores of students who participated in a state test the prior year, disaggregated by ethnicity

	Fall 2001						Fall 2003									
Grade 3																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	392	187.88	12.00	268	183.16	11.18	333	189.82	11.94	300	183.48	12.06	1.94	0.16	0.32	0.03
Asian	715	188.84	13.36	358	195.59	12.01	879	190.01	13.29	452	198.28	13.04	1.17	0.09	2.69	0.22
African-American	695	186.30	11.15	357	186.17	10.29	727	188.13	11.31	351	188.96	11.04	1.83	0.16	2.79	0.27
Hispanic	1984	183.27	12.23	2011	184.72	11.40	2175	185.39	11.83	2172	187.09	11.94	2.12	0.17	2.36	0.21
European-American	15356	191.66	11.49	3859	193.42	11.51	15216	192.65	11.59	3998	195.46	11.73	0.99	0.09	2.05	0.18
Grade 4																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	483	191.13	13.02	192	196.32	11.71	561	193.93	11.01	239	198.93	12.03	2.80	0.22	2.61	0.22
Asian	449	198.29	12.11	415	206.28	12.09	521	199.26	13.93	439	207.69	12.21	0.97	0.08	1.41	0.12
African-American	572	195.90	11.20	396	194.95	11.54	657	196.87	11.70	390	197.24	12.88	0.97	0.09	2.28	0.20
Hispanic	1594	194.03	12.20	2120	194.58	12.18	1903	194.93	12.33	2336	197.84	12.27	0.90	0.07	3.25	0.27
European-American	11227	202.79	10.77	7050	204.39	11.26	11561	203.67	11.45	6795	205.53	11.32	0.87	0.08	1.14	0.10
Grade 5																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	169	204.50	12.98	448	200.00	13.15	163	205.17	12.42	493	203.37	13.20	0.67	0.05	3.37	0.26
Asian	108	213.70	11.10	886	211.48	13.81	129	213.67	10.03	1060	211.16	15.19	-0.03	0.00	-0.32	-0.02
African-American	132	203.55	12.14	884	203.52	12.33	157	204.73	12.38	942	205.91	11.98	1.18	0.10	2.40	0.19
Hispanic	554	202.48	11.98	3824	201.79	12.71	625	204.61	11.85	4383	204.03	13.41	2.13	0.18	2.23	0.18
European-American	4828	211.50	12.17	17767	211.36	12.46	4824	212.37	12.16	17446	212.08	12.45	0.88	0.07	0.72	0.06
Grade 6																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	534	206.96	13.91	90	210.72	15.23	633	205.37	14.26	88	209.74	12.41	-1.59	-0.11	-0.98	-0.06
Asian	235	219.64	15.87	546	219.47	15.49	229	222.10	15.74	655	221.06	17.67	2.47	0.16	1.59	0.10
African-American	450	211.63	13.65	329	209.30	13.21	493	212.83	13.71	332	211.26	12.86	1.20	0.09	1.96	0.15
Hispanic	1033	208.77	13.72	2938	207.83	13.38	976	210.51	14.75	3430	208.07	14.18	1.74	0.13	0.24	0.02
European-American	12895	219.29	13.32	8105	218.33	13.35	12088	219.89	13.57	7940	219.95	14.04	0.60	0.05	1.62	0.12

Grade 7																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	450	205.76	15.58	178	216.37	15.97	479	208.51	15.15	201	218.86	15.74	2.76	0.18	2.48	0.16
Asian	78	228.56	15.42	563	225.18	17.21	90	227.57	16.93	663	223.71	16.68	-1.00	-0.06	-1.47	-0.09
African-American	58	216.05	16.08	773	214.34	15.60	125	215.22	13.20	796	217.52	16.67	-0.84	-0.05	3.18	0.20
Hispanic	288	213.73	14.11	3413	212.04	15.20	525	212.96	15.42	4031	213.83	15.75	-0.76	-0.05	1.78	0.12
European-American	3609	226.45	14.37	14302	225.83	14.70	3938	225.26	14.37	14446	226.37	15.03	-1.18	-0.08	0.54	0.04

Grade 8																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	254	230.56	14.56	492	214.35	15.37	57	226.67	16.09	491	218.47	15.88	-3.89	-0.27	4.12	0.27
Asian	92	238.96	14.98	320	230.57	17.16	138	231.89	17.37	438	227.09	17.92	-7.07	-0.47	-3.48	-0.20
African-American	218	222.44	19.47	142	220.40	16.72	263	219.97	20.08	194	220.49	15.79	-2.47	-0.13	0.09	0.01
Hispanic	466	221.28	16.05	2281	216.52	16.29	582	223.15	14.86	2789	218.14	16.37	1.88	0.12	1.62	0.10
European-American	7804	232.52	15.91	6446	230.99	15.55	9299	233.50	15.46	6730	231.49	15.71	0.98	0.06	0.50	0.03

Table A2 - Fall reading scores of students who participated in a state test the prior year, disaggregated by ethnicity

	Fall 2001						Fall 2003									
Grade 3																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	356	188.63	15.40	233	179.73	15.24	299	188.54	16.13	257	180.92	14.98	-0.10	-0.01	1.19	0.08
Asian	359	191.04	13.87	398	193.32	14.61	402	191.91	14.60	477	193.39	15.13	0.87	0.06	0.07	0.00
African-American	493	184.98	15.25	367	184.04	13.59	513	187.08	14.18	385	185.41	14.60	2.10	0.14	1.37	0.10
Hispanic	801	183.51	15.17	2733	180.45	15.66	991	183.90	14.68	2879	182.03	15.52	0.39	0.03	1.58	0.10
European-American	11239	191.79	14.88	6038	192.71	14.54	11247	192.77	14.35	6089	193.05	14.47	0.98	0.07	0.34	0.02
Grade 4																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	335	184.67	16.06	219	193.36	15.25	393	187.16	14.80	262	197.03	15.55	2.49	0.15	3.67	0.24
Asian	140	199.39	12.08	496	202.29	13.59	151	200.93	14.90	515	202.69	13.75	1.54	0.13	0.40	0.03
African-American	510	192.86	15.52	402	193.29	15.50	558	194.07	14.20	414	194.48	16.00	1.21	0.08	1.19	0.08
Hispanic	553	192.44	14.82	2807	189.73	15.81	638	192.86	14.82	3234	191.26	15.14	0.42	0.03	1.53	0.10
European-American	7743	200.82	13.95	9251	203.01	13.60	8170	201.65	13.94	8970	202.88	13.77	0.83	0.06	-0.13	-0.01
Grade 5																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	128	201.88	15.02	434	194.02	15.72	118	201.44	14.83	486	195.72	16.74	-0.43	-0.03	1.70	0.11
Asian	102	208.32	11.22	542	210.12	12.81	135	208.07	11.72	598	210.46	14.29	-0.26	-0.02	0.34	0.03
African-American	131	202.32	14.01	790	201.12	14.36	155	203.60	12.82	827	202.00	13.31	1.28	0.09	0.89	0.06
Hispanic	504	197.41	15.09	3085	197.10	15.64	581	200.42	13.34	3607	198.24	15.29	3.00	0.20	1.14	0.07
European-American	4389	209.23	13.11	16302	208.60	13.74	4357	209.23	13.42	16074	208.93	13.35	0.00	0.00	0.33	0.02
Grade 6																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	510	201.02	14.85	80	206.50	15.19	628	199.40	15.28	93	206.04	13.82	-1.62	-0.11	-0.46	-0.03
Asian	229	213.72	14.19	357	215.92	13.68	233	213.23	13.33	424	218.25	12.68	-0.49	-0.03	2.32	0.17
African-American	466	206.99	13.18	305	204.95	13.75	512	207.49	14.32	284	205.58	13.76	0.49	0.04	0.63	0.05
Hispanic	1137	203.13	14.88	2686	202.11	15.43	1066	202.86	16.03	3066	201.44	16.10	-0.27	-0.02	-0.67	-0.04
European-American	12958	213.81	13.21	7646	214.33	13.37	12270	213.86	13.46	7349	215.26	13.57	0.05	0.00	0.93	0.07

Grade 7																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	396	200.86	15.14	172	212.40	14.46	414	201.50	15.55	199	211.04	13.59	0.64	0.04	-1.37	-0.09
Asian	66	216.68	16.49	589	215.67	14.65	77	217.36	16.01	689	217.40	15.34	0.68	0.04	1.73	0.12
African-American	71	211.00	14.27	860	208.96	14.02	124	209.80	15.14	876	210.50	14.80	-1.20	-0.08	1.53	0.11
Hispanic	265	208.12	14.58	3350	205.79	15.65	516	206.86	15.67	3895	206.94	16.10	-1.26	-0.09	1.15	0.07
European-American	3664	219.24	12.54	14868	218.91	12.88	4083	217.82	13.36	15667	218.74	13.27	-1.42	-0.11	-0.16	-0.01
Grade 8																
	No State Test			State Test Administered			No State Test			State Test Administered			No State Test		State Test Administered	
	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Count	Mean	Std Deviation	Mean Diff	Effect Size	Mean Diff	Effect Size
Native American	252	220.62	12.48	456	205.91	15.70	50	218.34	14.01	412	209.97	14.63	-2.28	-0.18	4.06	0.26
Asian	124	221.79	13.14	456	221.45	14.22	170	219.66	14.03	506	222.26	15.80	-2.13	-0.16	0.81	0.06
African-American	260	213.74	16.91	154	214.99	15.14	287	212.07	15.98	221	216.51	12.98	-1.67	-0.10	1.52	0.10
Hispanic	490	211.25	14.54	2655	209.46	16.56	667	213.39	14.77	3236	210.50	16.50	2.14	0.15	1.04	0.06
European-American	7995	222.22	12.99	8717	223.45	13.05	9474	222.20	12.90	9087	223.61	13.00	-0.02	0.00	0.16	0.01

Appendix B

Table B1 – Change in mathematics growth index scores between 2001-2002 and 2003-2004 school year, disaggregated by ethnicity

	2001-2002		2003-2004			
Native American						
Grade	**Mean**	**Std Deviation**	**Mean**	**Std Deviation**	**Difference**	**Effect Size**
3	-0.67	7.23	-1.14	7.58	-0.47	-0.06
4	-0.93	7.17	-1.08	7.23	-0.15	-0.02
5	-0.58	7.26	-0.60	6.79	-0.02	0.00
6	-1.87	6.97	-1.85	7.27	0.02	0.00
7	-0.68	7.01	-1.04	7.78	-0.36	-0.05
8	-1.66	7.36	-1.87	7.77	-0.21	-0.03
Asian						
Grade	**Mean**	**Std Deviation**	**Mean**	**Std Deviation**	**Difference**	**Effect Size**
3	0.51	7.29	-0.22	7.24	-0.72	-0.10
4	0.49	6.47	0.02	7.04	-0.46	-0.07
5	0.03	6.27	-0.56	6.86	-0.59	-0.09
6	0.00	6.35	-0.87	6.16	-0.87	-0.14
7	-0.33	6.30	-1.17	6.26	-0.84	-0.13
8	-1.12	7.36	-2.23	6.18	-1.11	-0.15
African-American						
Grade	**Mean**	**Std Deviation**	**Mean**	**Std Deviation**	**Difference**	**Effect Size**
3	-2.39	7.36	-2.54	7.44	-0.15	-0.02
4	-1.70	7.72	-1.51	7.17	0.19	0.02
5	-0.98	7.38	-1.94	7.03	-0.96	-0.13
6	-2.60	6.95	-2.60	7.59	0.00	0.00
7	-2.13	7.58	-2.60	7.65	-0.47	-0.06
8	-2.58	7.25	-3.71	7.54	-1.13	-0.16
Hispanic						
Grade	**Mean**	**Std Deviation**	**Mean**	**Std Deviation**	**Difference**	**Effect Size**
3	-1.36	7.33	-1.28	7.08	0.09	0.01
4	-0.89	6.84	-0.60	7.03	0.30	0.04
5	-0.31	6.95	-0.60	7.14	-0.28	-0.04
6	-1.55	6.89	-1.56	7.18	-0.01	0.00
7	-1.24	6.82	-1.43	7.12	-0.19	-0.03
8	-2.11	6.85	-2.88	6.94	-0.77	-0.11

European-American						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	0.29	7.17	0.10	7.02	-0.19	-0.03
4	0.77	6.88	0.58	6.94	-0.19	-0.03
5	0.34	6.82	0.41	7.02	0.07	0.01
6	-0.55	6.71	-0.78	6.96	-0.23	-0.03
7	-0.35	6.86	-0.91	7.03	-0.56	-0.08
8	-1.48	7.17	-2.06	7.18	-0.58	-0.08

Table B2 – Change in reading growth index scores between 2001-2002 and 2003-2004 school year, disaggregated by ethnicity

	2001-2002		2003-2004			
Native American						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	-1.07	7.88	-1.60	7.63	-0.53	-0.07
4	-1.92	7.59	-2.25	7.43	-0.33	-0.04
5	-2.61	7.26	-1.58	7.34	1.03	0.14
6	-1.44	6.71	-1.86	7.71	-0.42	-0.06
7	-1.96	7.10	-2.00	6.92	-0.03	0.00
8	-0.44	6.87	-0.97	7.24	-0.53	-0.08
Asian						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	0.59	6.97	-0.53	6.66	-1.11	-0.16
4	0.31	6.08	-0.02	6.77	-0.34	-0.06
5	-0.29	6.06	-0.36	6.48	-0.07	-0.01
6	0.13	5.65	0.15	5.93	0.02	0.00
7	0.16	5.96	-0.23	6.06	-0.39	-0.07
8	0.46	5.90	0.18	6.06	-0.28	-0.05
African-American						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	-2.08	9.20	-2.61	8.20	-0.53	-0.06
4	-2.19	8.58	-1.92	7.59	0.27	0.03
5	-2.43	7.62	-2.20	7.81	0.23	0.03
6	-2.30	7.49	-1.73	7.90	0.57	0.08
7	-2.26	8.65	-2.06	7.68	0.20	0.02
8	-1.65	8.40	-2.18	8.33	-0.53	-0.06

Hispanic						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	-1.66	8.79	-1.40	8.28	0.26	0.03
4	-1.35	7.58	-1.01	7.81	0.34	0.05
5	-1.08	7.17	-1.32	7.39	-0.24	-0.03
6	-1.25	7.18	-1.23	7.51	0.02	0.00
7	-0.85	6.84	-1.22	7.59	-0.36	-0.05
8	-0.43	6.33	-1.12	6.93	-0.69	-0.11
European-American						
Grade	Mean	Std Deviation	Mean	Std Deviation	Difference	Effect Size
3	0.58	7.59	0.11	7.39	-0.47	-0.06
4	-0.08	6.85	-0.16	7.12	-0.08	-0.01
5	-0.49	6.50	-0.56	6.71	-0.08	-0.01
6	-0.31	6.65	-0.50	6.74	-0.19	-0.03
7	-0.83	6.59	-0.92	7.06	-0.08	-0.01
8	-0.42	6.64	-0.61	6.88	-0.19	-0.03

Appendix C

Table C1 - Changes in mathematics growth index scores for grades in which the state test is administered, disaggregated by ethnic group

	2002				2004											
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
Grade	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	392	-0.15	268	-1.43	333	-0.82	300	-1.50	7.20	-0.66	-0.09	-0.07	-0.01	0.60	0.08
4	483	-0.97	192	-0.84	561	-1.70	239	0.36	6.97	-0.73	-0.10	1.20	0.17	1.93	0.28
5	169	-1.95	448	-0.07	163	-0.23	493	-0.73	6.95	1.72	0.25	-0.66	-0.09	-2.38	-0.34
6	534	-1.90	90	-1.72	633	-2.05	88	-0.42	6.90	-0.15	-0.02	1.30	0.19	1.46	0.21
7	450	-0.87	178	-0.19	479	-1.33	201	-0.35	7.00	-0.46	-0.07	-0.16	-0.02	0.31	0.04
8	254	-2.48	492	-1.24	57	-2.60	491	-1.79	7.16	-0.12	-0.02	-0.55	-0.08	-0.43	-0.06
Native American	2282	-1.27	1668	-0.82	2226	-1.51	1812	-0.94	7.02	-0.24	-0.03	-0.12	-0.02	0.12	0.02
	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	715	-0.01	358	1.54	879	-1.22	452	1.73	7.20	-1.21	-0.17	0.19	0.03	1.40	0.19
4	449	-0.91	415	2.00	521	-1.94	439	2.35	6.97	-1.03	-0.15	0.36	0.05	1.39	0.20
5	108	-0.12	886	0.05	129	-0.05	1060	-0.62	6.95	0.07	0.01	-0.67	-0.10	-0.73	-0.11
6	235	-0.55	546	0.23	229	-0.48	655	-1.01	6.90	0.07	0.01	-1.24	-0.18	-1.31	-0.19
7	78	0.03	563	-0.37	90	-0.51	663	-1.26	7.00	-0.54	-0.08	-0.88	-0.13	-0.34	-0.05
8	92	-2.57	320	-0.70	138	-2.07	438	-2.28	7.16	0.49	0.07	-1.58	-0.22	-2.07	-0.29
Asian	1677	-0.47	3088	0.36	1986	-1.27	3707	-0.36	7.07	-0.80	-0.11	-0.72	-0.10	0.08	0.01
	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	695	-2.49	357	-2.19	727	-2.67	351	-2.28	7.20	-0.18	-0.03	-0.09	-0.01	0.10	0.01
4	572	-2.04	396	-1.20	657	-1.74	390	-1.13	6.97	0.31	0.04	0.07	0.01	-0.24	-0.03
5	132	-1.00	884	-0.98	157	-0.85	942	-2.12	6.95	0.15	0.02	-1.14	-0.16	-1.29	-0.19
6	450	-2.33	329	-2.98	493	-2.61	332	-2.59	6.90	-0.28	-0.04	0.39	0.06	0.67	0.10
7	58	-0.86	773	-2.23	125	-2.03	796	-2.69	7.00	-1.17	-0.17	-0.46	-0.07	0.71	0.10
8	218	-3.64	142	-0.96	263	-3.86	194	-3.51	7.16	-0.22	-0.03	-2.55	-0.36	-2.33	-0.32
African-American	2125	-2.32	2881	-1.72	2422	-2.38	3005	-2.30	7.05	-0.07	-0.01	-0.58	-0.08	-0.51	-0.07

	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	1984	-1.73	2011	-1.00	2175	-1.75	2172	-0.80	7.20	-0.02	0.00	0.20	0.03	0.22	0.03
4	1594	-1.46	2120	-0.46	1903	-1.04	2336	-0.24	6.97	0.43	0.06	0.23	0.03	-0.20	-0.03
5	554	-0.67	3824	-0.26	625	-0.47	4383	-0.62	6.95	0.20	0.03	-0.35	-0.05	-0.55	-0.08
6	1033	-1.61	2938	-1.52	976	-1.55	3430	-1.56	6.90	0.06	0.01	-0.04	-0.01	-0.10	-0.01
7	288	-1.08	3413	-1.25	525	-1.46	4031	-1.42	7.00	-0.38	-0.05	-0.17	-0.02	0.21	0.03
8	466	-2.53	2281	-2.02	582	-3.32	2789	-2.78	7.16	-0.78	-0.11	-0.76	-0.11	0.02	0.00
Hispanic	5919	-1.57	16587	-1.05	6786	-1.51	19141	-1.25	7.05	0.05	0.01	-0.20	-0.03	-0.25	-0.04
	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	15356	-0.14	3859	2.03	15216	-0.29	3998	1.58	7.20	-0.66	-0.09	-0.07	-0.01	0.60	0.08
4	11227	0.54	7050	1.13	11561	0.22	6795	1.19	6.97	-0.73	-0.10	1.20	0.17	1.93	0.28
5	4828	-0.74	17767	0.64	4824	-0.25	17446	0.59	6.95	1.72	0.25	-0.66	-0.09	-2.38	-0.34
6	12895	-1.15	8105	0.39	12088	-1.31	7940	0.02	6.90	-0.15	-0.02	1.30	0.19	1.46	0.21
7	3609	-0.51	14302	-0.31	3938	-0.98	14446	-0.89	7.00	-0.46	-0.07	-0.16	-0.02	0.31	0.04
8	7804	-1.92	6446	-0.95	9299	-2.54	6730	-1.40	7.16	-0.12	-0.02	-0.55	-0.08	-0.43	-0.06
European-American	55719	-0.56	57529	0.34	56926	-0.81	57355	0.04	7.04	-0.25	-0.04	-0.30	-0.04	-0.05	-0.01

Table C2 - Changes in reading growth index scores for grades in which the state test is administered, disaggregated by ethnic group

	2002				2004											
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
Grade	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	356	-0.50	299	-0.27	233	-1.95	257	-3.14	7.78	-1.45	-0.19	-2.87	-0.37	-1.42	-0.18
4	335	-2.13	393	-3.78	219	-1.60	262	0.04	7.17	0.54	0.07	3.82	0.53	3.28	0.46
5	128	-2.34	118	-1.69	434	-2.69	486	-1.56	6.76	-0.36	-0.05	0.14	0.02	0.49	0.07
6	510	-1.62	628	-1.94	80	-0.33	93	-1.30	6.87	1.29	0.19	0.64	0.09	-0.65	-0.09
7	396	-2.36	414	-2.53	172	-1.06	199	-0.89	6.96	1.30	0.19	1.63	0.23	0.33	0.05
8	252	-1.04	50	-1.70	456	-0.10	412	-0.88	6.78	0.94	0.14	0.82	0.12	-0.12	-0.02
Native American	356	-0.50	299	-0.27	233	-1.95	257	-3.14	7.78	-1.45	-0.19	-2.87	-0.37	-1.42	-0.18
	2002				2004											
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean		Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	359	1.49	402	-0.37	398	-0.23	477	-0.66	7.78	-1.72	-0.22	-0.28	-0.04	1.44	0.18
4	140	0.56	151	0.34	496	0.25	515	-0.13	7.17	-0.31	-0.04	-0.47	-0.06	-0.15	-0.02
5	102	-0.38	135	-0.96	542	-0.27	598	-0.22	6.76	0.11	0.02	0.73	0.11	0.62	0.09
6	229	-0.22	233	-0.45	357	0.36	424	0.49	6.87	0.59	0.09	0.94	0.14	0.35	0.05
7	66	-0.30	77	-1.32	589	0.21	689	-0.11	6.96	0.52	0.07	1.22	0.18	0.70	0.10
8	124	-0.68	170	-1.13	456	0.77	506	0.62	6.78	1.45	0.21	1.75	0.26	0.30	0.04
Asian	1020	0.41	1168	-0.54	2838	0.17	3209	-0.02	7.22	-0.24	-0.03	0.52	0.07	0.76	0.10
	2002				2004											
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	493	-1.52	513	-2.04	367	-2.82	385	-3.36	7.78	-1.30	-0.17	-1.31	-0.17	-0.01	0.00
4	510	-2.47	558	-2.10	402	-1.83	414	-1.68	7.17	0.64	0.09	0.42	0.06	-0.23	-0.03
5	131	-0.12	155	-1.60	790	-2.82	827	-2.31	6.76	-2.69	-0.40	-0.71	-0.11	1.98	0.29
6	466	-1.58	512	-1.68	305	-3.40	284	-1.83	6.87	-1.82	-0.26	-0.15	-0.02	1.67	0.24
7	71	-2.30	124	-2.56	860	-2.26	876	-1.99	6.96	0.04	0.01	0.57	0.08	0.53	0.08
8	260	-2.33	287	-1.89	154	-0.51	221	-2.57	6.78	1.82	0.27	-0.69	-0.10	-2.51	-0.37
African-American	1931	-1.83	2149	-1.95	2878	-2.45	3007	-2.24	7.17	-0.62	-0.09	-0.29	-0.04	0.33	0.05

	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	801	-0.52	991	0.10	2733	-1.99	2879	-1.91	7.78	-1.47	-0.19	-2.01	-0.26	-0.54	-0.07
4	553	-1.50	638	-0.63	2807	-1.32	3234	-1.08	7.17	0.17	0.02	-0.45	-0.06	-0.62	-0.09
5	504	-0.59	581	-1.27	3085	-1.17	3607	-1.33	6.76	-0.57	-0.08	-0.06	-0.01	0.51	0.08
6	1137	-0.85	1066	-1.09	2686	-1.42	3066	-1.28	6.87	-0.57	-0.08	-0.19	-0.03	0.38	0.06
7	265	-1.09	516	-1.17	3350	-0.84	3895	-1.22	6.96	0.25	0.04	-0.05	-0.01	-0.30	-0.04
8	490	-0.53	667	-1.40	2655	-0.41	3236	-1.07	6.78	0.12	0.02	0.33	0.05	0.21	0.03
Hispanic	3750	-0.81	4459	-0.84	17316	-1.18	19917	-1.30	7.09	-0.37	-0.05	-0.46	-0.07	-0.10	-0.01
	2002				2004										
	No State Test Administered		State Test Administered		No State Test Administered		State Test Administered		Pooled Standard Deviation	No State Test Administered		State Test Administered		Difference Y/N	
	Count	Mean	Count	Mean	Count	Mean	Count	Mean	SD	Diff No	Eff Size	Diff Yes	Eff Size	Diff Both	Eff Size
3	11239	0.19	11247	-0.21	6038	1.32	6089	0.71	7.78	-0.66	-0.09	-0.07	-0.01	0.60	0.08
4	7743	-0.24	8170	-0.30	9251	0.06	8970	-0.04	7.17	-0.73	-0.10	1.20	0.17	1.93	0.28
5	4389	-1.08	4357	-0.95	16302	-0.33	16074	-0.46	6.76	1.72	0.25	-0.66	-0.09	-2.38	-0.34
6	12958	-0.63	12270	-0.82	7646	0.24	7349	0.04	6.87	-0.15	-0.02	1.30	0.19	1.46	0.21
7	3664	-1.07	4083	-1.39	14868	-0.78	15667	-0.79	6.96	-0.46	-0.07	-0.16	-0.02	0.31	0.04
8	7995	-0.93	9474	-1.22	8717	0.05	9087	0.02	6.78	-0.12	-0.02	-0.55	-0.08	-0.43	-0.06
European-American	47988	-0.50	49601	-0.73	62822	-0.10	63236	-0.24	7.11	0.40	0.06	0.49	0.07	0.09	0.01