Putting the "Student" Back in Student Accountability:

A Longitudinal Study of the Impact of the Delaware Student Accountability Plan on Delaware Students

November 2000

Lisa A. Banicky, Ph.D.
Sr. Associate for Policy Analysis

Delaware Education Research and Development Center
University of Delaware
Newark, DE 19716

Publication T00.017.1

ACKNOWLEDGEMENTS

This study was a collaborative effort of various members of the Accountability Research Team of the Delaware Education Research and Development Center. The author extends special thanks to:

Eileen Malone
Audrey Noble
Jennifer Parisella

The author also extends a special thanks to the following individuals who provided the data and assistance necessary to complete this study:

Terry Anderson
Ann Case
Shirley Dear
Theresa Kough
Katie Semmel
Tommy Tao
Robin Taylor

This research was made possible through the support of the Delaware State Board of Education.

TABLE OF CONTENTS

Executive Summary

Recent calls for increased accountability in education have brought with them recommendations for monitoring systems to determine their effect on students and schools. Although a definitive model for monitoring and evaluating high-stakes accountability systems has yet to emerge, experts agree that the intended and unintended consequences that emerge in high-stakes settings must be examined.

Therefore, the goal of this study is to monitor the impact of Delaware's Student Accountability Plan on students in the First State. Students enrolled in 3rd or 6th grade during the 1997-98 school year were selected as cohorts and followed over a three year period (1997-2000).

Based on the recommendations of several national educational organizations and existing research, several issues were examined in light of the accountability plan. The issues examined and the results to emerge from the study are summarized in the following table:

Issue	**What the Research Says**	**Results of Delaware Study**
Special Education Placements	Instituting high-stakes assessments has been linked to increases in the number of students placed in special education. Many view this as an attempt to manipulate the system and exclude some students from testing.	Across the three years of study less than 4% of regular education students were later re-classified as special education students. However, large percentages of special education students were later re-classified as regular education.
Student Achievement	Many states have found student test scores to increase over time but to decrease when a new testing instrument is implemented. In addition, some researchers have found reported gains in test scores to represent inflated scores.	SAT9 reading and DSTP writing scores did not improve over time but the SAT9 mathematics scores showed slight improvements for most student groups in the 3rd grade cohort. With respect to standards-based scores, an examination of student performance levels indicated that more students were performing at or above the standard in reading and math and fewer students were performing at or above the standard in writing in 2000 as compared to 1998.

Issue	What the Research Says	Results of Delaware Study
Student Behavior	Out-of-school suspension is one of the least effective strategies for improving student behavior but it is widely used. In high-stakes settings, removing a student from school is considered particularly problematic because it can limit opportunity to learn. In addition, minority students are often over-represented in out-of-school suspension rates.	For students in the 3rd grade and 6th grade cohorts, the odds of being given an out-of-school suspension was greater for male, low income, and minority students. Suspended and non-suspended students did not differ in terms of their DSTP performance.
Retention	Retention, as typically practiced, is considered to be an ineffective and harmful strategy for dealing with low performing students. Previous research also indicates that minority, low income, male students are more likely to be retained.	Consistent with previous research, the odds of being retained increased for certain DE student groups. This finding was qualified by whether a student had actually been retained or would have been retained if the consequences associated with the DSTP had been in effect. When considering potential retentions, the odds of being retained were greater for low income status, minority, and male students. The DSTP scores for both actual and potential retainees indicated that a majority of these students continued to perform below the standard.

Introduction

States and school districts across the nation are implementing high-stakes assessment systems to gather information about student achievement for the purpose of holding schools, teachers, and students accountable. For many, accountability offers the promise of improving education through the use of external rewards and sanctions. But it is the inclusion of these rewards and sanctions that have led many educational experts to warn system developers of unintended consequences that can occur in high-stakes systems. Some of these unintended consequences include manipulation of the system (i.e. classifying more students as special education students, excluding students from testing that are expected to decrease performance) or increases in student retention rates and drop-out rates.[1]

Educational experts at the national level have begun conversations around the importance of monitoring and evaluating the impact of accountability systems on all students. The National Research Council made the following recommendation with respect to this issue: "*high stakes testing programs [should] routinely include a well-designed evaluation component. Policymakers should monitor both the intended and unintended consequences of high-stakes assessment on all students and on significant sub-groups of students including minorities, English language learners, and students with disabilities*" (p. 281).[2] More recently the American Educational Research Association (AERA) offered a similar recommendation in their position statement on high-stakes testing in PreK-12 education.[3]

In keeping with the recommendations of the National Research Council and AERA, this study is being conducted at the request of the Delaware State Board of Education with the intent of monitoring Delaware's student accountability plan and providing ongoing information about its effects on students.

Context and Purpose of the Current Study

Over ten years of educational reform in Delaware have culminated in a performance-based accountability system composed of rigorous content standards, a statewide assessment system, and consequences for performance. According to the policymakers involved in the development of the system, Delaware's Student Accountability plan was designed to:

- Improve student achievement by providing a system for measuring student performance against content standards;
- Ensure that all children can achieve by establishing an educational system that expects more and provides more;
- Prepare the workforce by motivating educators toward continuous improvement as professionals; and,

- Focus the educational system on student outcomes by motivating changes in performance and behavior through the use of external rewards and sanctions.[4]

Although the intention of many accountability systems is to benefit students, the extent to which this is actually occurring is often not examined. Therefore, the goal of this study is to put the focus back on students and examine the impact of Delaware's Student Accountability Plan on students in the state.

This study is one of three designed to monitor the effect of the accountability legislation on students in Delaware. The first study examined Delaware policymakers' original intentions for the student accountability plan, and served as a reference point for understanding the original goals of the plan. The translation of these goals at the school level is currently being examined through a 3-year case study designed to examine how Delaware schools are changing over time.[5]

Guiding Questions

Although many experts agree upon the importance of monitoring and evaluating high-stakes systems, a clear picture of what the monitoring efforts should entail has yet to emerge. Given that previous research has found increased rates of special education placements, retentions, and drop-outs in high-stakes systems, these issues were examined in the current study.

In addition to these issues, student behavior was also examined. Researchers in the area of school discipline have found a link between students' academic self-concept and discipline problems. Children who believe they are poor students often have behavior problems and these problems worsen as student achievement declines.[6] Although researchers have not directly examined the link between high-stakes systems and students' academic self-concept, it is reasonable to assume that an increased focus on student outcomes could result in some students (i.e. those unable to meet the standards) viewing themselves as poorer learners.

Based on the research cited above and the data available from the state database, the longitudinal study was designed to address the following issues and questions:

1. **Special Education Placement Rates**
   a. Are the special education placement rates changing within the two cohorts over time, and if so, in what ways?

2. **Student Achievement**
   a. Are students' test scores improving over time?
   b. Do test score gains differ as a function of gender, race, or low income status?
   c. How much change is there in performance levels on the DSTP from one test administration to the next?

3. **Behavioral Effects**
    a. What are the demographic characteristics of students who are given out-of-school suspension?
    b. What is the relationship of out-of-school suspension to performance on the DSTP?

4. **Retention**
    a. What are the demographic characteristics of students who are retained or who would have been retained if the consequences for performance on the DSTP had been in effect?
    b. What is the relationship of retention to performance on the DSTP?
    c. After being retained, how do students fare with respect to behavioral effects and special education status?

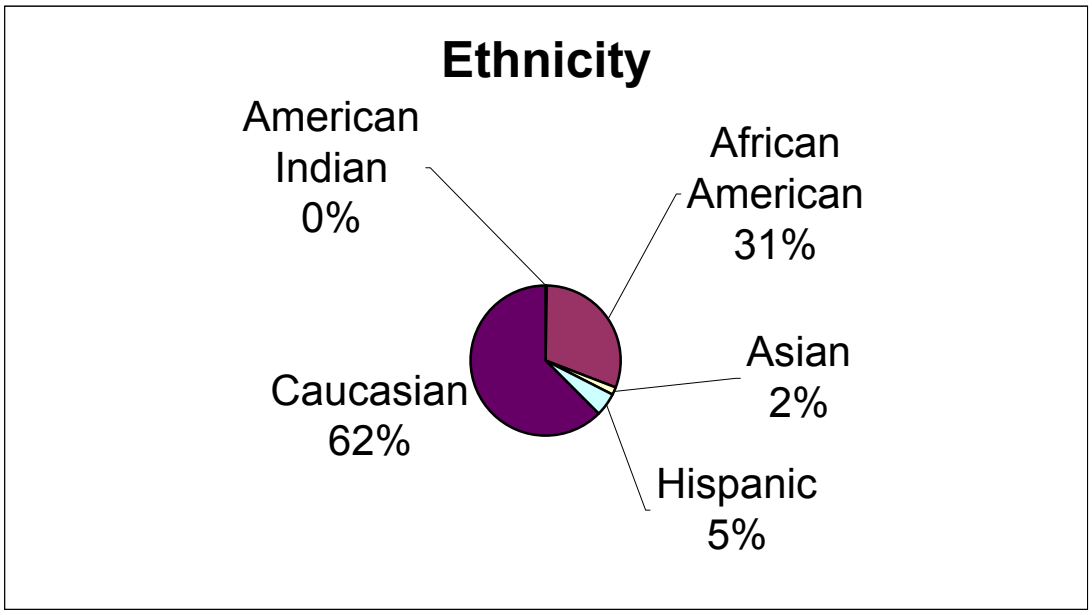
5. **Completion Rates (Not addressed at this time)**
    a. What are the demographic characteristics of students who dropout?
    b. What is the relationship of dropping-out to performance on the DSTP?

To address these issues, two cohorts of students were selected for the study. The first group, hereafter referred to as the elementary cohort, was composed of students enrolled in 3rd grade as of the 1997-98 school year. The second group, hereafter referred to as the secondary cohort, was composed of students enrolled in 6th grade as of the 1997-98 school year. The 1997-98 school year was chosen because this was the first year in which the Delaware Student Testing Program (DSTP) was administered to students. For an overview of the timeline for the longitudinal study and the available data see ***Appendix A.***

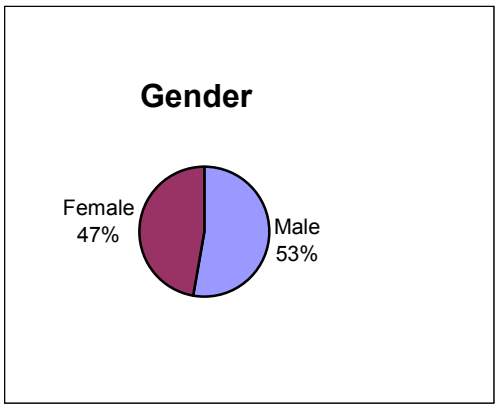
The results of this first longitudinal study are organized around these issues and attempt to answer each of the questions with the data currently available. A summary of the findings for each of these issues and the policy considerations associated with the findings appear at the end of the report in the section entitled "Policy Considerations".

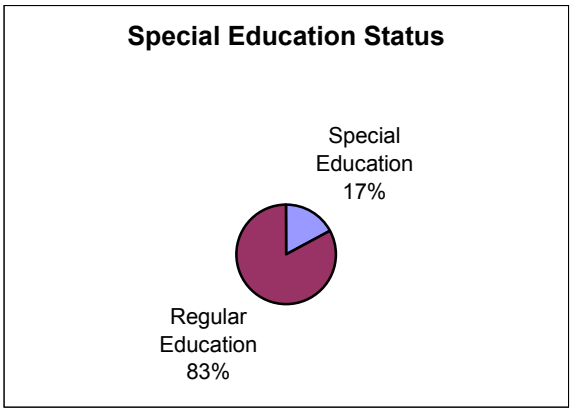
Elementary Cohort Profile

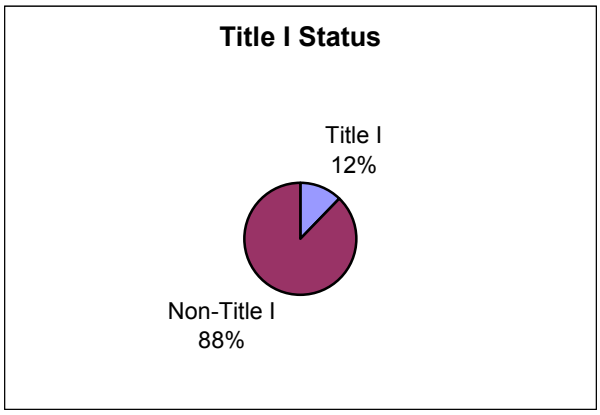
The following profile of the elementary cohort is based on data received as of the September 30th count for the 1997-98 school year. As of that point in time the total number of students in the elementary cohort was 8408. The race, gender, Title I, LEP, and Low Income status of these students were as follows:

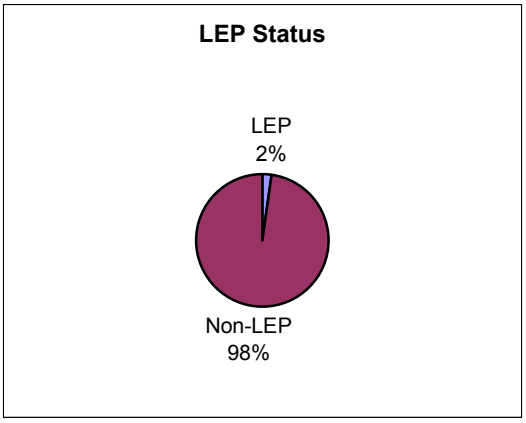


Note: American Indians constituted .3% of students.

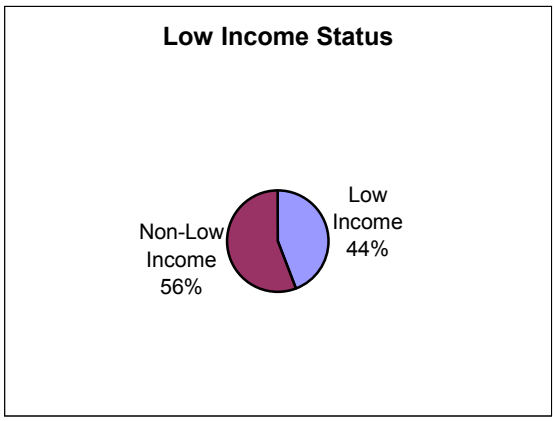








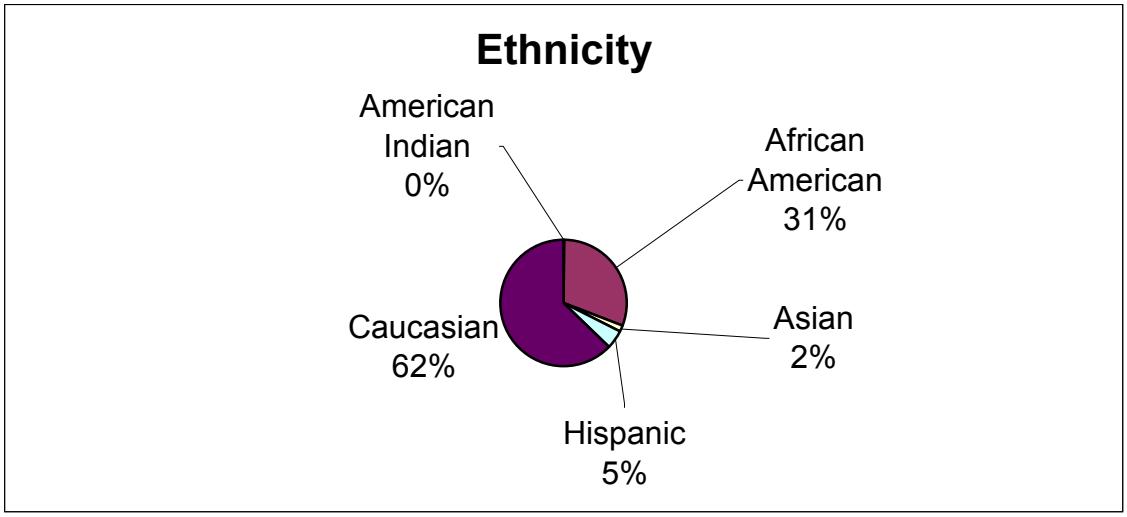
Note: Title I classification includes students eligible for either Title I Math or Reading, or students eligible for both.



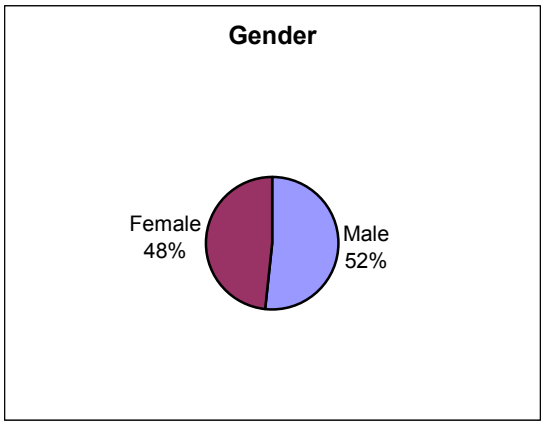
Note: Low income classification is determined by the number of students eligible for free or reduced lunch.

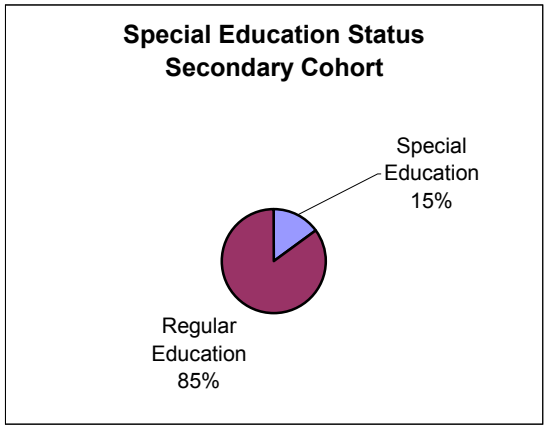
Secondary Cohort Profile

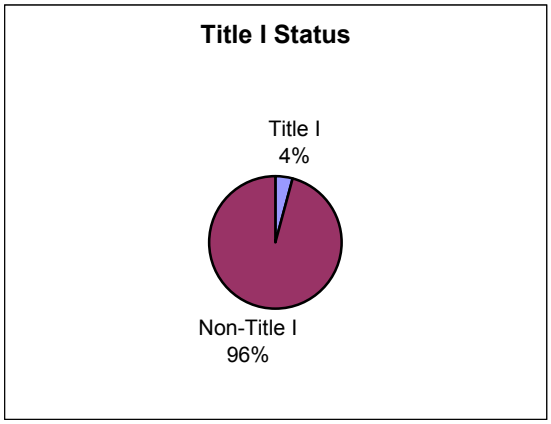
The following profile of the secondary cohort is based on data received as of the September 30th count for the 1997-98 school year. As of that point in time the total number of students in the secondary cohort was 8732. The race, gender, Title I, LEP, and Low Income status of these students were as follows:

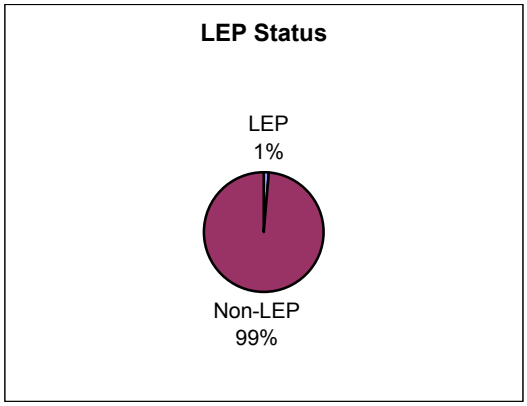


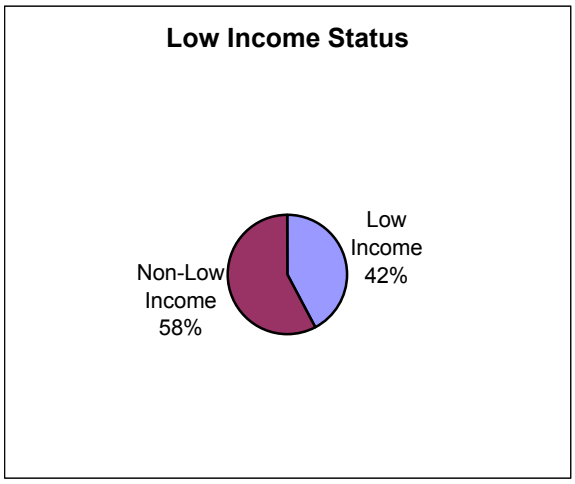
Note: American Indians constituted .3% of students in the secondary cohort.










Low Income Status
Low Income 42%
Non-Low Income 58%

Issue 1: Special Education Placement Rates

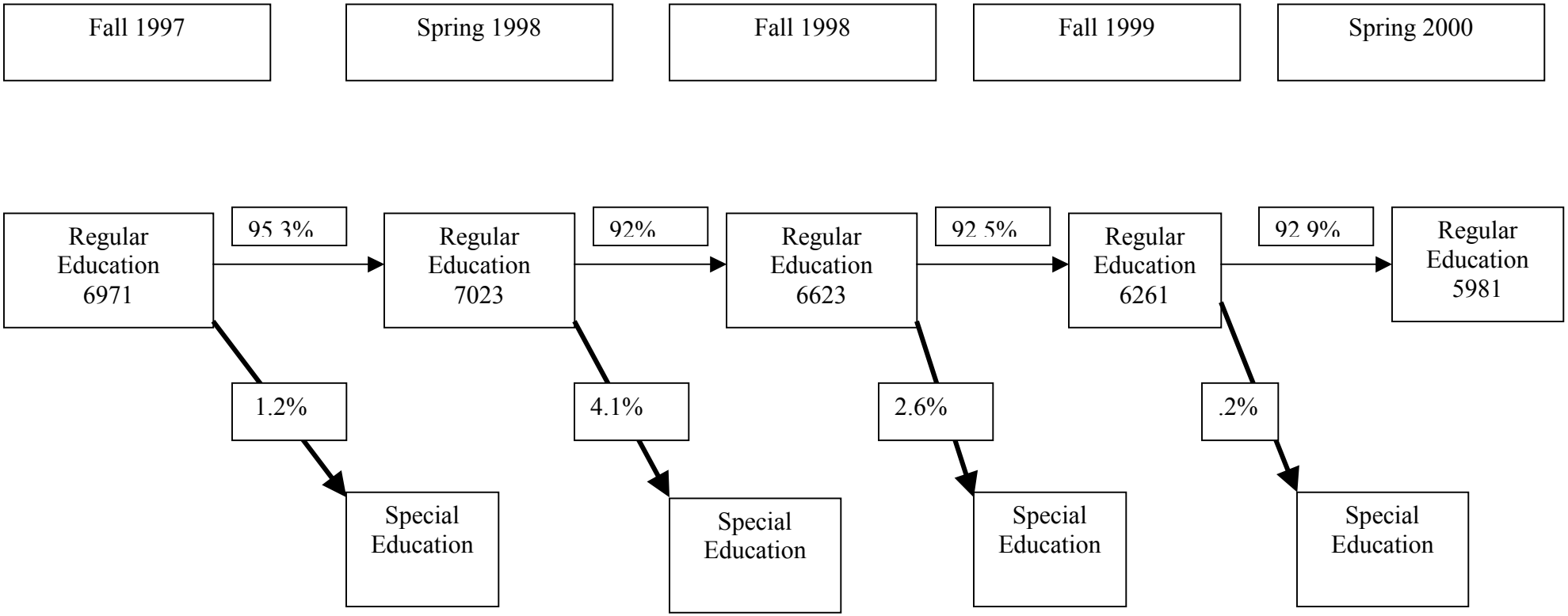
Previous research indicates that using assessments for high stakes purposes can lead to increases in the incidences of classification into special education.[7] Such an increase may represent an attempt to manipulate the system and exclude students from testing that are expected to perform poorly.

To determine how special education enrollment rates are changing as the state prepares to attach consequences to performance, the special education status of the elementary and secondary cohorts were examined for changes over time within each cohort.

Question 1a: Are the special education placement rates changing within the two cohorts over time, and if so, in what ways?

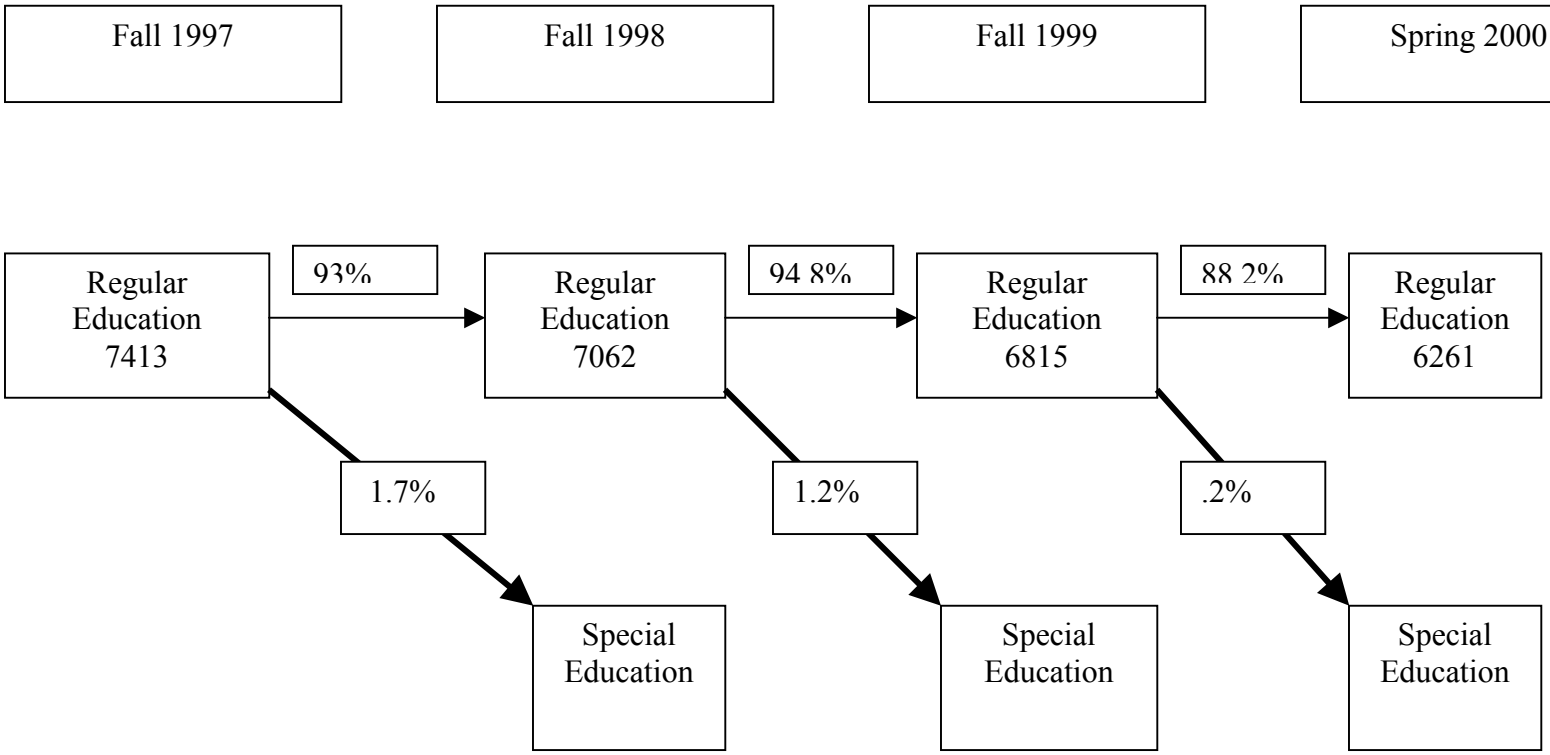
The following diagrams provide an overview of the changes in special education placements for both cohorts. Placements are reported as of the September 30th count and as of the administration of the DSTP in the Spring.

Elementary Cohort: Percentage of Regular Education Students Later Re-classified as Special Education*



* The percentages do not sum to 100% because the percentage of missing students at each point in time was not included.

Secondary Cohort: Percentage of Regular Education Students Later Re-classified as Special Education*



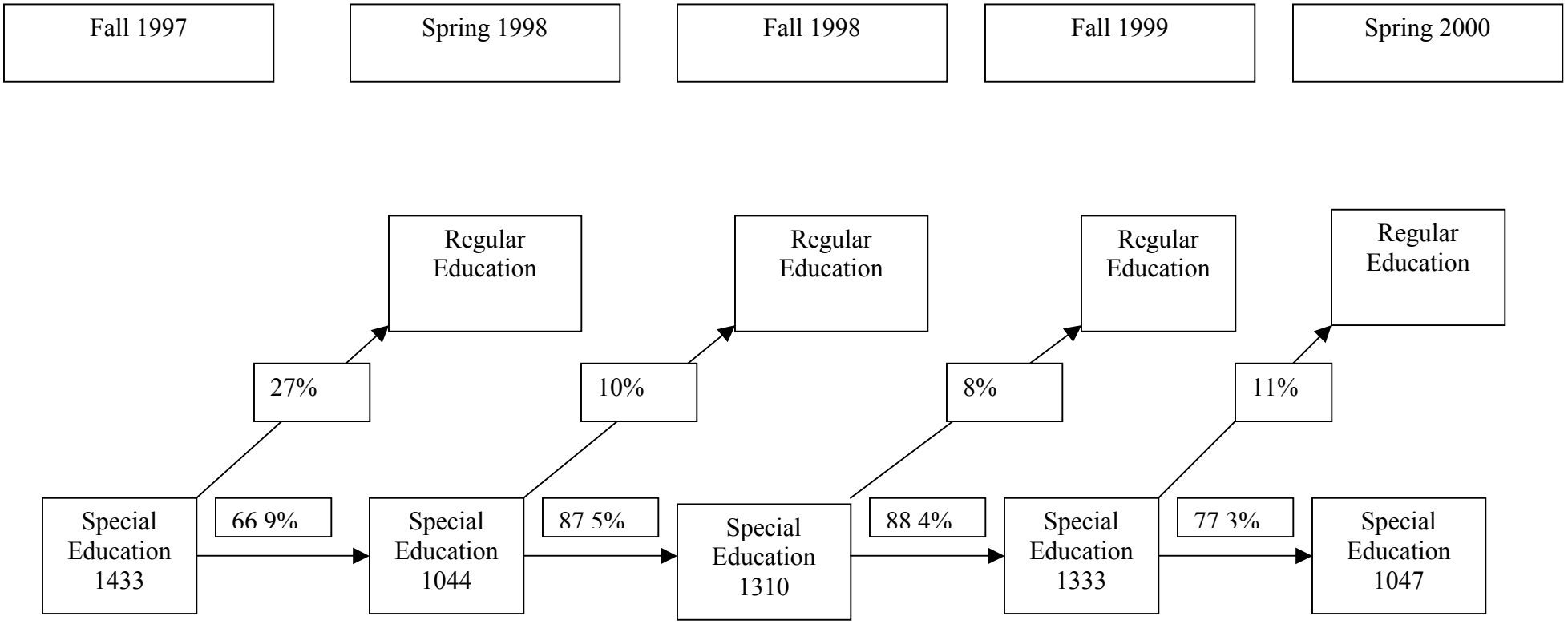
* The percentages do not sum to 100% because the percentage of missing students at each point in time was not included.

It should be noted that there are no rules of thumb as to what constitutes an over classification into special education. However an examination of the elementary cohort diagram revealed that the largest change in classification from regular education to special education occurred from the spring of 1998 to the fall of 1998. Roughly 4% of students classified as regular education in the spring of 1998 were classified as special education in the fall of 1998.

For the secondary cohort, the largest change in classification from regular education to special education occurred between the Fall of 1997 and the Fall of 1998. Roughly 1.7% of students classified as regular education in the fall of 1997 were classified as special education during that time period.

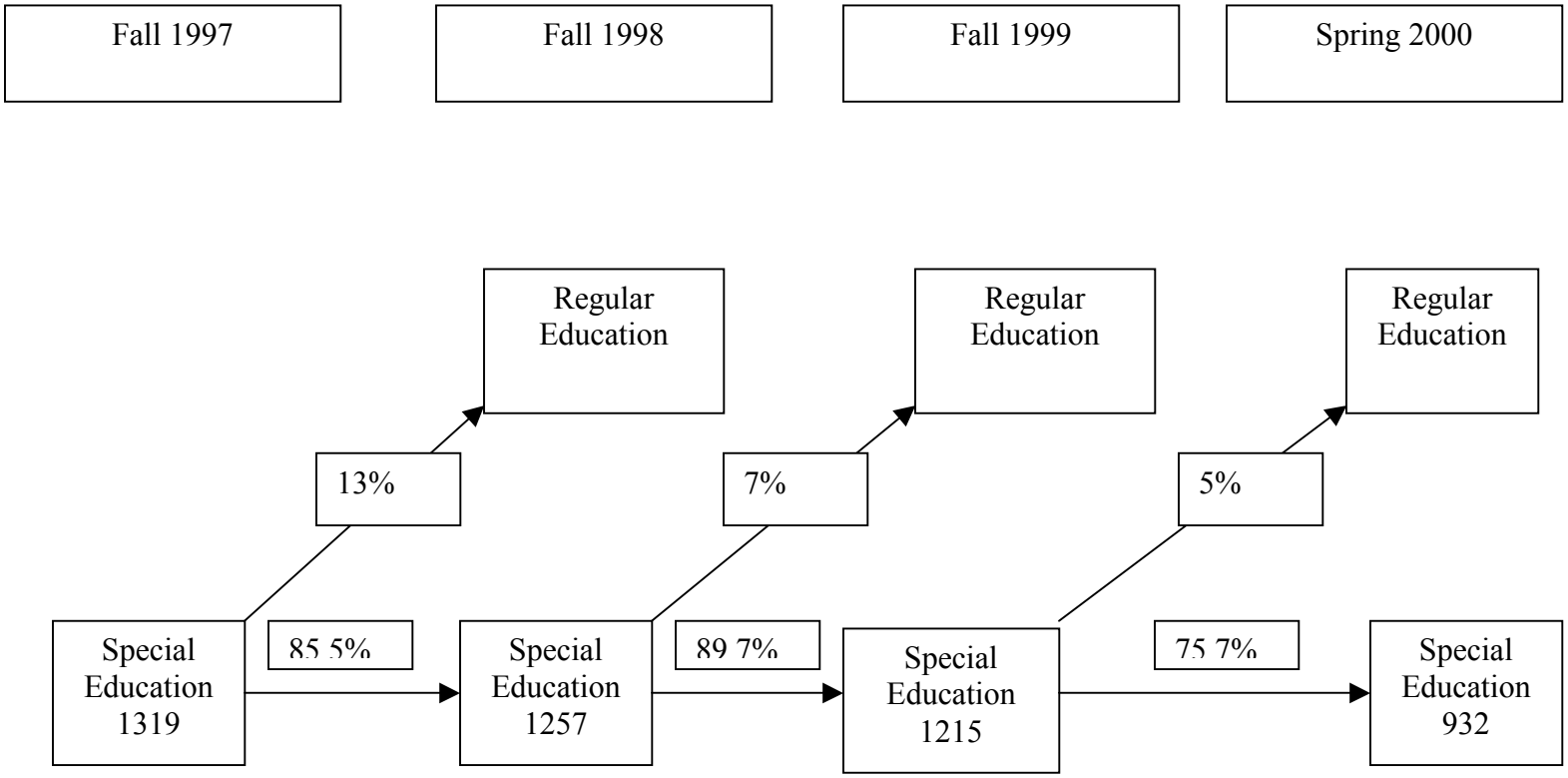
An unexpected finding to emerge from the enrollment file was the number of special education students whose classification status changed to regular education. When comparing the changes, more special education students were classified as regular education than vice versa. The following figures display the changes that occurred in special education placements over time. This finding may require further investigation.

Elementary Cohort: Percentage of Special Education Students Later Re-classified as Regular Education*



* The percentages do not sum to 100% because the percentage of missing students at each point in time was not included.

Secondary Cohort: Percentage of Special Education Students Later Re-classified as Regular Education*



* The percentages do not sum to 100% because the percentage of missing students at each point in time was not included.

Issue 2: Student Achievement

One of the overriding goals of any accountability system is to improve student learning. In most systems the tool used to measure learning is some form of a statewide assessment. The Delaware Student Testing Program (DSTP) is designed to measure students' progress toward reaching the Delaware content standards.

Technical Details Concerning the DSTP

To address the student achievement issue, a few technical details concerning the DSTP should be addressed. The DSTP is composed of multiple choice, short answer, and extended response items. Results are reported out in the form of national percentile ranks, standards-based scores, and performance levels.

Percentile Ranks

The national percentile rankings are based on abbreviated versions of the reading comprehension and the mathematical problem solving subsets of the Stanford Achievement Test series, 9th Edition (SAT9). The SAT9 is a norm-referenced test published by Harcourt Brace Educational Measurement.

Although percentile ranks can provide useful information by referencing student performance against set norms, percentile ranks cannot be manipulated mathematically because there are not equal intervals between them. For example, the difference between a percentile rank of 5 and 10 is not the same as the difference in achievement as the difference between a percentile rank of 50 and 55. This point is worth noting not only for the analyses that follow but also for the purpose of avoiding incorrect conclusions based on cursory examinations of data.

In order to be used in statistical analyses the national percentile rankings must be converted to another metric, in this case normal curve equivalents (NCEs). NCEs can range from 1 to 99 and provide an equal-interval scale which makes them amenable to mathematical manipulation. For the purposes of the current study, students' scores on the SAT9 portion of the DSTP will be reported in NCE units.

Standards-based Scores

The standards-based score reported for the DSTP ranges from 150 to 800 and is based on students' responses to items developed in Delaware and a subset of the SAT9 items that are considered to be aligned with the Delaware content standards. According to the 2000 DSTP Executive Summary, "students in the earlier grades should tend to score towards the lower part of the scale, while students in the upper grades should tend to score towards the higher part of the scale".[8]

The expectation that students in the early grades will have lower scores than students in the upper grades is a function of the manner in which the standards-based scores are

scaled. A vertical scaling system has been applied to the scores which means that a score of 400 in 3rd grade is not equivalent to a score of 400 in 5th grade. This also means that if a student's score is the same in both 3rd grade and 5th grade, they have actually done worse, instead of holding steady the lack of change in scores would represent a decline.

Vertical scaling becomes an issue when attempts are made to track student improvement longitudinally. Since there is an expected amount of increase from year to year, an increase reflects the manner in which the scores are scaled, not true gains. Hypothetically, a score of 400 in 3rd grade may be equivalent to a score of 450 in 5th grade. In this case, if a student did receive these scores, it would appear to represent a 50 point increase, but because of the vertical scaling the students achievement is actually unchanged from 3rd to 5th grade. For this reason, the questions related to student achievement over time can only be examined by using the SAT9 data.

A Word of Caution About Statistical Interpretations

The goal of many statistical analyses is to show that there is some difference between sets of observations, and that the difference is due to something other than chance factors. For example, when examining the elementary cohort's average SAT9 math score from 1998 (Mean=53.88) to their average SAT9 math score in 2000 (Mean=57.35) there is an increase of 3.43. Finding such a difference does not necessarily mean that it is a meaningful difference. This difference may simply reflect the amount of variability in the data. Statistical analyses are set up in such a way as to compare the difference found, in this case 3.43, to a measure of how much of a difference is expected simply due to chance factors. The extent to which the observed difference (3.43) is over and above the amount of difference expected simply due to chance determines whether or not a result is statistically significant. Therefore a statistically significant result simply means that an outcome, in this example a difference of 3.43, is unlikely to be due to chance factors and instead may represent an actual improvement in scores.

Recently many researchers have argued that significance tests can be misleading because with very large sample sizes, even the smallest difference between two sets of observations can result in a significant finding. Therefore, testing for statistical significance is often viewed as the first step in data analysis with the second step focused on the size of the "effect".[9]

To use an analogy, testing for statistical significance is like using a magnifying glass to locate an object. The size of the sample determines the "magnification" of the lens. Consequently, larger samples result in even the smallest difference appearing quite large. Effect size can be thought of as a ruler that researchers use to measure the size of their findings. In the case of the magnifying glass analogy, we may locate an object that appears to be quite large, but when the ruler (i.e. effect size) is placed next to it under the magnifying glass the size of the object is placed in a more meaningful context.

Conventional rules of thumb indicate that an effect size of .2 is small, an effect size of .5 is medium, and an effect size of .8 is large.[10] For the purposes of this study, a statistically

significant finding with an effect size of less than .20 is not considered to represent a real difference or change in scores.

With these caveats in mind, the analyses reported in the following sections provide information as to the statistical significance of the findings as well as the size of the effect found for each analysis.

Question 2a: Are students' test scores improving over time?

Method

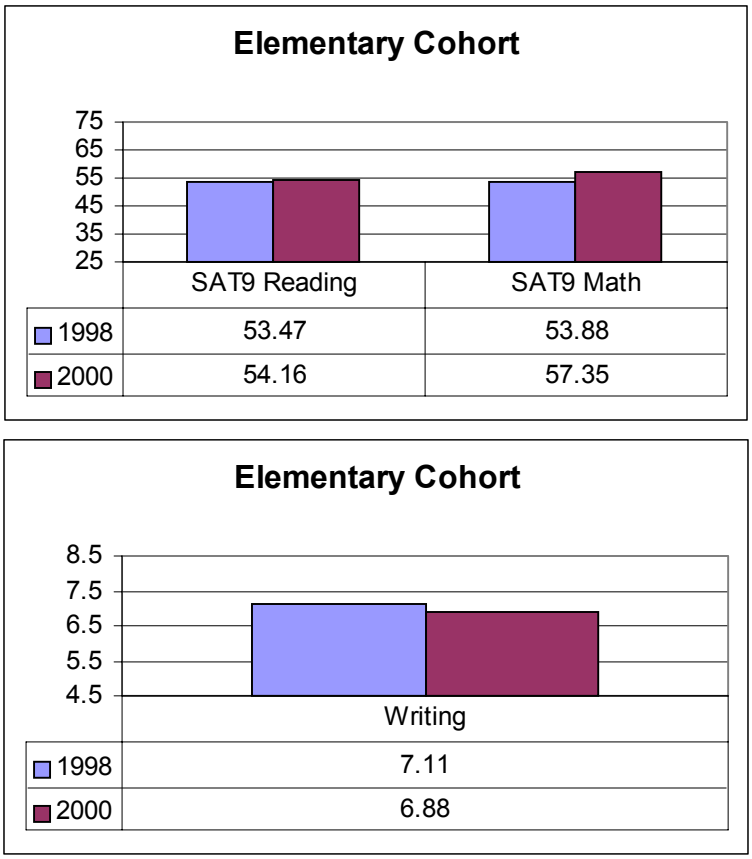
An examination of improvements over time was conducted through use of paired samples t-tests. In this analysis, students' scores from the spring of 1998 were compared to their scores in the spring of 2000. This analysis required that a student have a score at both points in time in order to be included.

When conducting statistical analyses researchers select a probability value indicating how unlikely an outcome needs to be to consider it as resulting from something other than chance factors. In most cases, .05 is chosen as the probability value. If an outcome of an analysis has a probability of occurring that is less than this value then it is considered statistically significant and the researcher can be 95% confident that the outcome reflects a true difference and not simply chance factors.

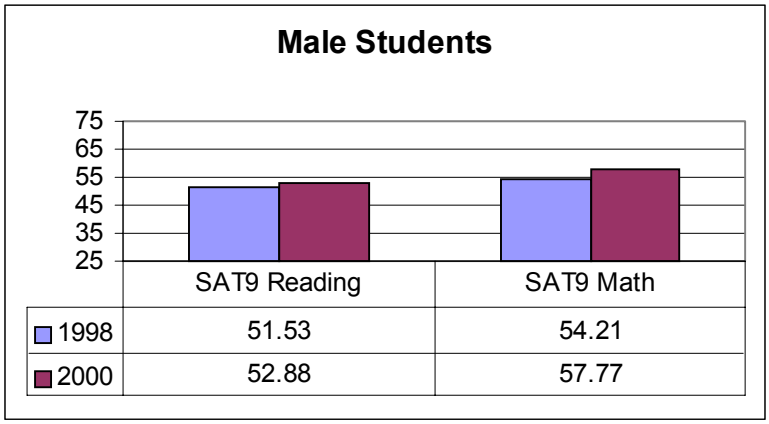
When multiple analyses are conducted on the same set of data, in this case multiple dependent t-tests, adjustments to the probability level must be made. This adjustment is necessary to prevent the researcher from capitalizing on chance factors. It works to limit the likelihood that a researcher will falsely conclude that a difference is statistically significant. One method of adjusting the probability value is to divide it by the number of planned comparisons. For example, when looking at the relationship of SAT9 Math scores to Gender, one dependent t-test was conducted for males and one for females. In this case the probability value (.05) was divided by two resulting in a probability value of .025. In this case, an outcome had to have a probability less than .025 to be considered statistically significant. Similar adjustments were made for each of the dependent t-tests reported.[11]

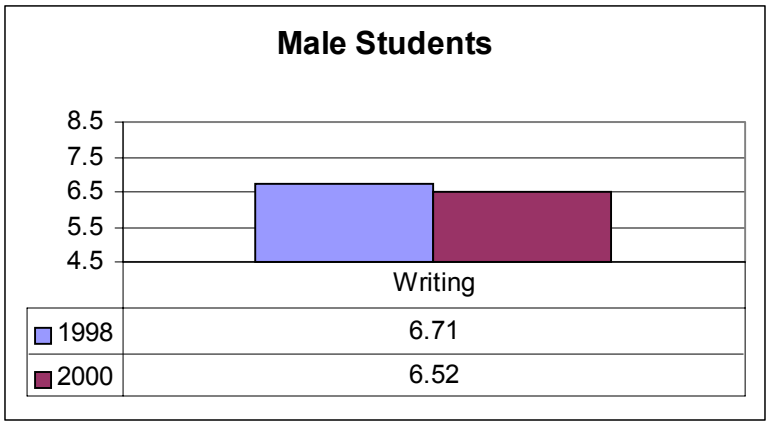
Results

When examining the elementary cohort overall, the results revealed statistically significant increases for SAT9 reading and mathematics scores and a statistically significant decrease for writing scores. The effect sizes for the reading and writing results were negligible ($d<.20$) and the effect size for the math finding was small ($d=.20$), indicating a slight improvement in SAT9 math scores over time.

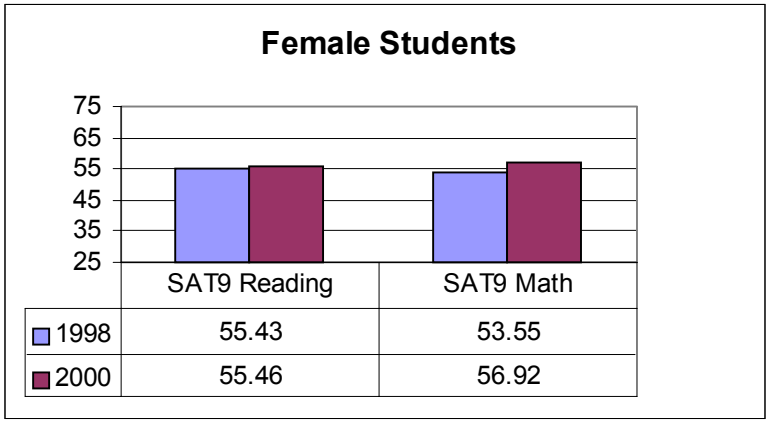


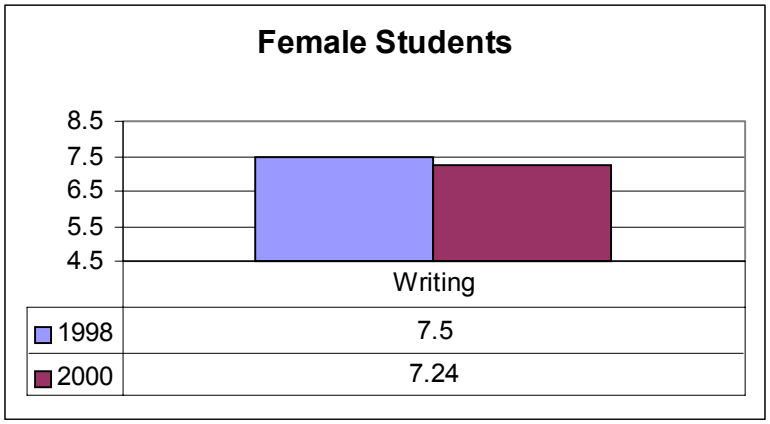
Male students in the elementary cohort showed statistically significant improvements in reading and mathematics and statistically significant decreases in writing scores. As with the elementary cohort as a whole, the effect sizes for males on the reading and writing results were negligible and the effect size for the mathematics finding indicated a slight improvement in scores (d=.21).



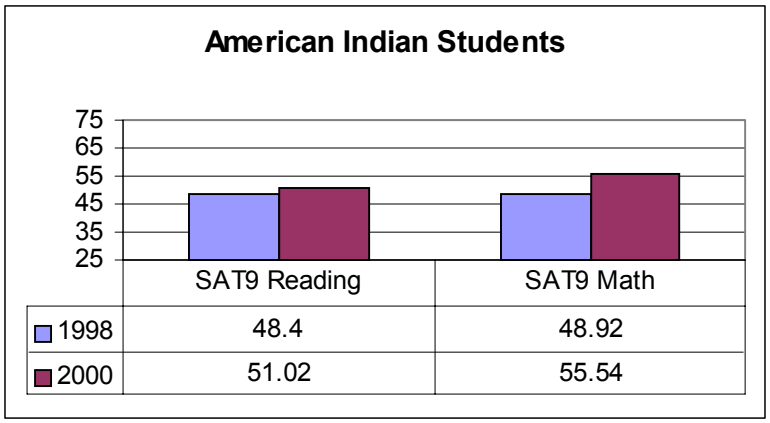


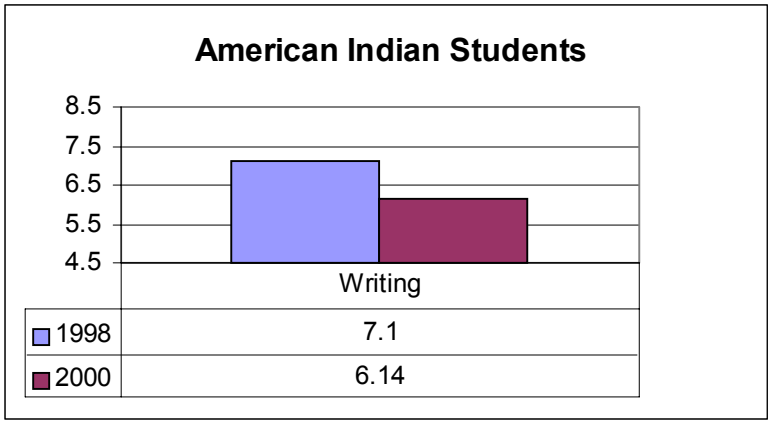
Female students in the elementary cohort showed statistically significant improvements in mathematics and statistically significant decreases in writing scores over time. The effect size for females on the writing results were negligible and the effect size for the mathematics result indicated a slight improvement in scores (d=.20).



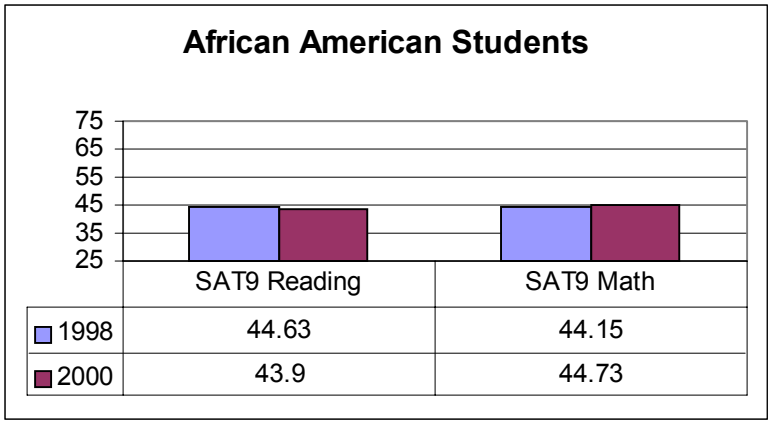


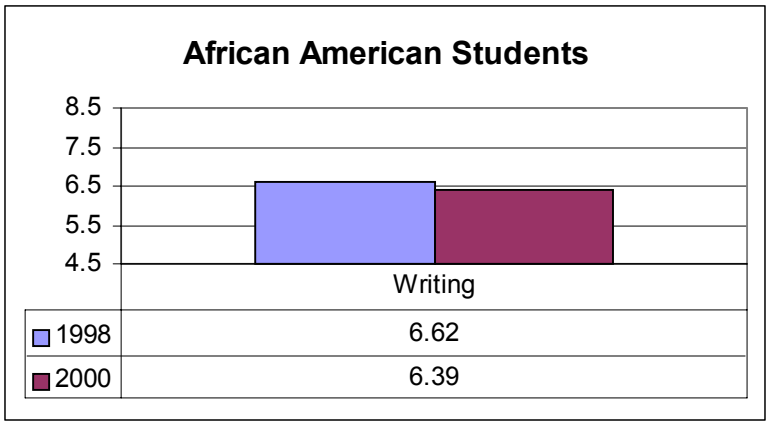
For American Indian students, there were no statistically significant differences for reading, mathematics, or writing.



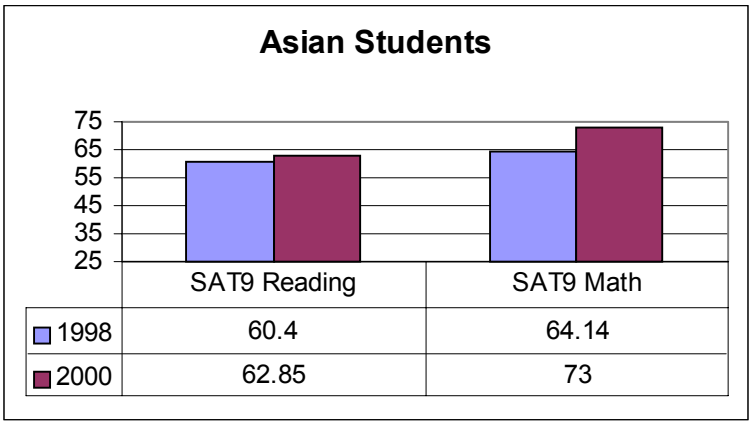


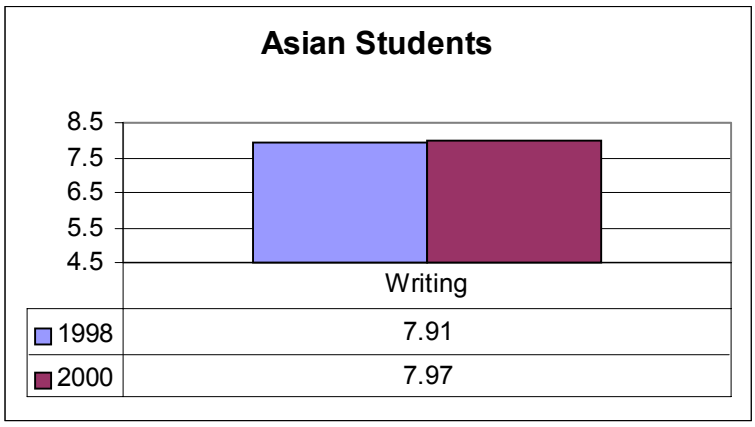
For African American students, there was a statistically significant decrease in writing scores over time but the effect size was negligible.



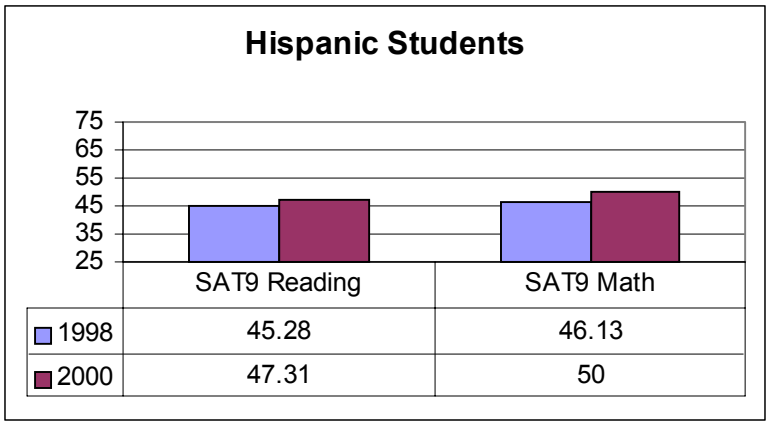


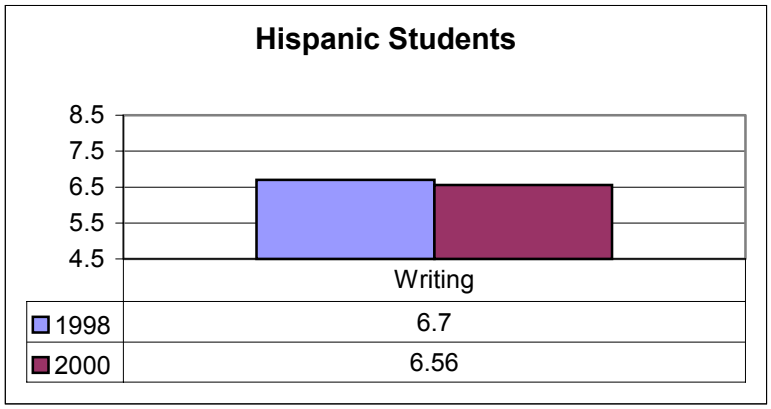
For Asian students, there was a statistically significant increase in mathematics scores over time with a medium effect size (d=.58), indicating a moderate improvement in mathematics scores.



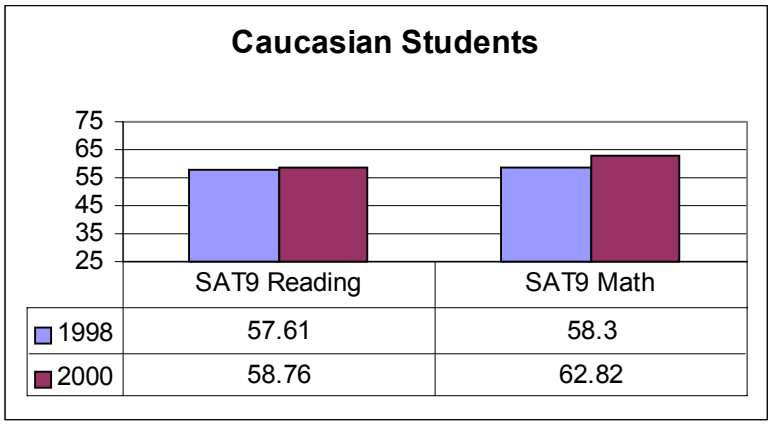


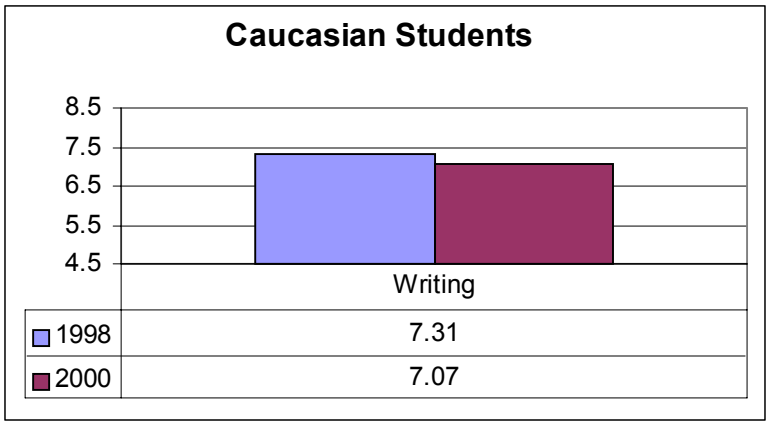
For Hispanic students there was a statistically significant increase in mathematics scores over time. The effect size for this result was small (d=.24), indicating a slight improvement in mathematics scores.



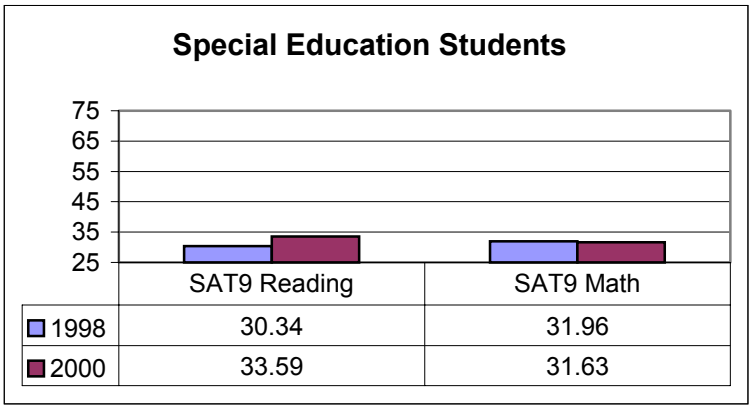


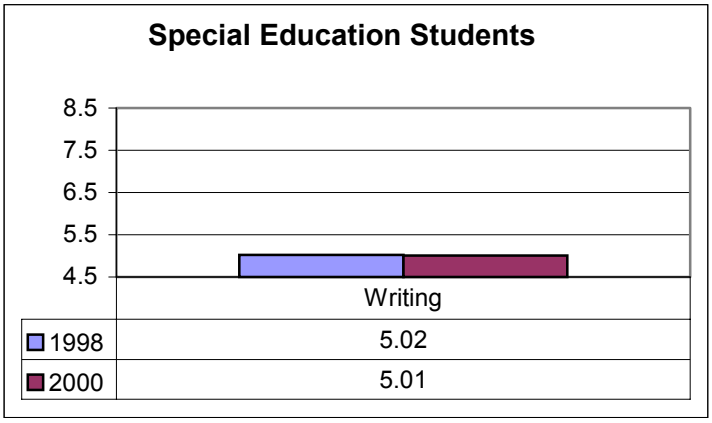
Caucasian students showed statistically significant improvements in reading and mathematics and statistically significant decreases in writing scores. As with the elementary cohort as a whole, the effect sizes for Caucasians on the reading and writing results were negligible and the effect size for the mathematics finding revealed a slight improvement in scores (d=.28).



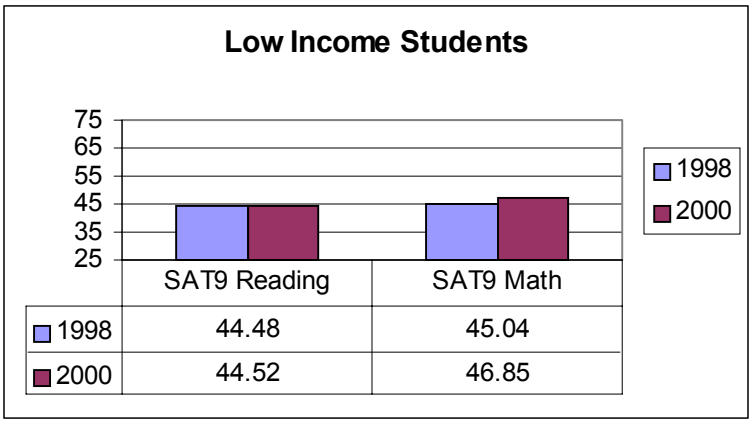


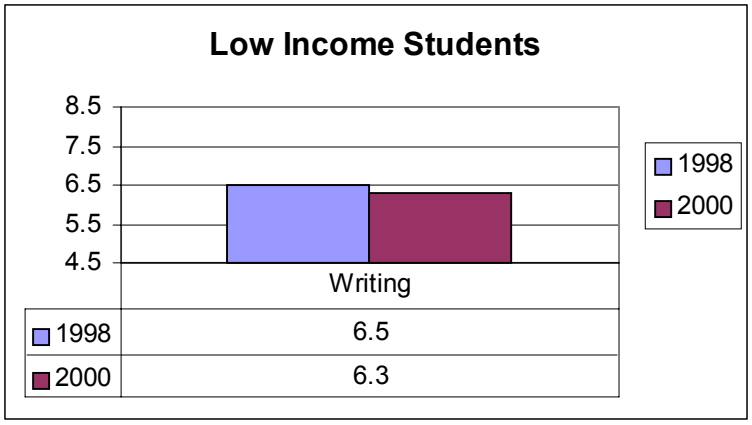
Special education students showed statistically significant improvements in reading scores but the effect size was negligible.



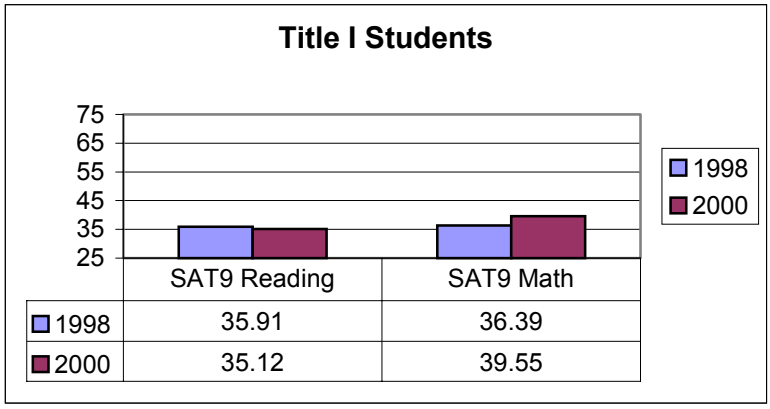


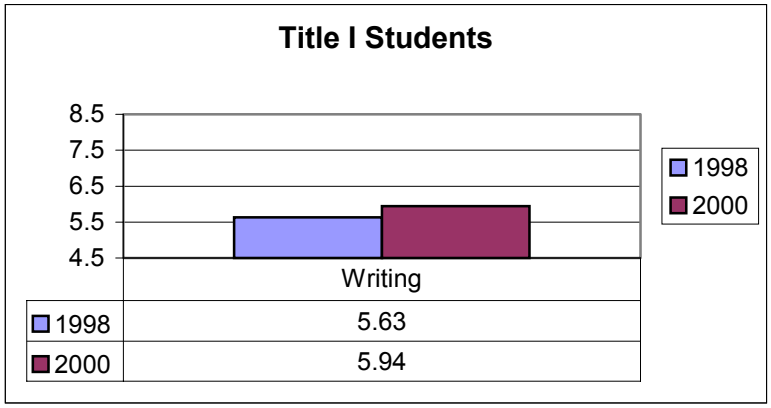
Low income students evidenced a statistically significant improvement in mathematics scores and a statistically significant decrease in writing scores. The effect sizes for both of these results were negligible.



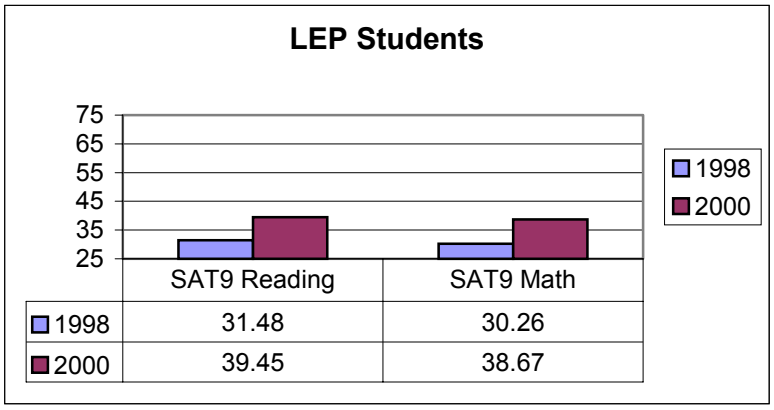


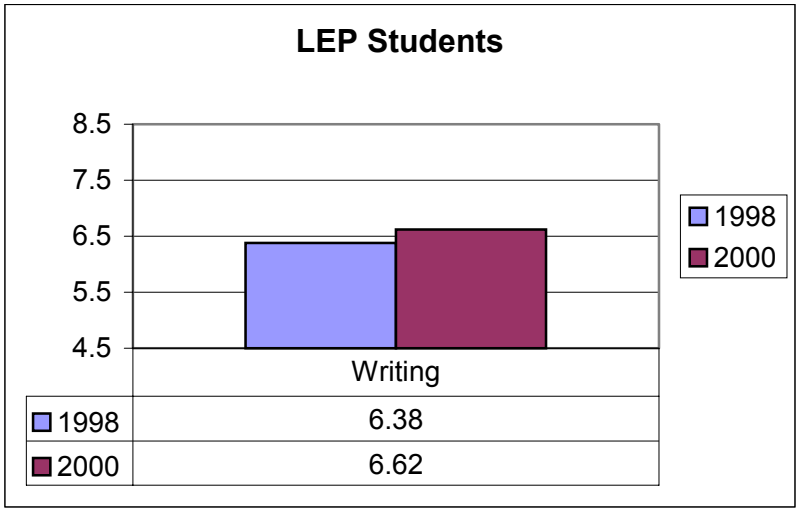
Title I students showed a statistically significant improvement in mathematics and a statistically significant increase in writing. The effect sizes for both of these results were negligible.





Limited English Proficient (LEP) students showed a statistically significant improvement in reading and mathematics scores. The effect size associated with the reading result was of a medium size (d=.70) while the effect size associated with the mathematics finding was large (d=.79).





In summary, the results of the paired samples t-tests revealed that most student groups showed small but statistically significant improvements in SAT9 mathematics scores from 1998 to 2000. Exceptions to this finding included American Indians, African Americans, Special education students, low income students, and Title I students. For these groups the scores showed no real change from 1998 to 2000. Also, the results for Asian students and LEP students revealed medium to large statistically significant improvements in SAT9 mathematics scores. LEP students also showed a medium size statistically significant improvement in SAT9 reading scores from 1998 to 2000.

Question 2b: Do test score gains differ as a function of gender, race, or income status?

The previous question, 2a, compared the performance differences between 1998 and 2000 *within* each student group. This second question compares the change in performance *between* student groups.

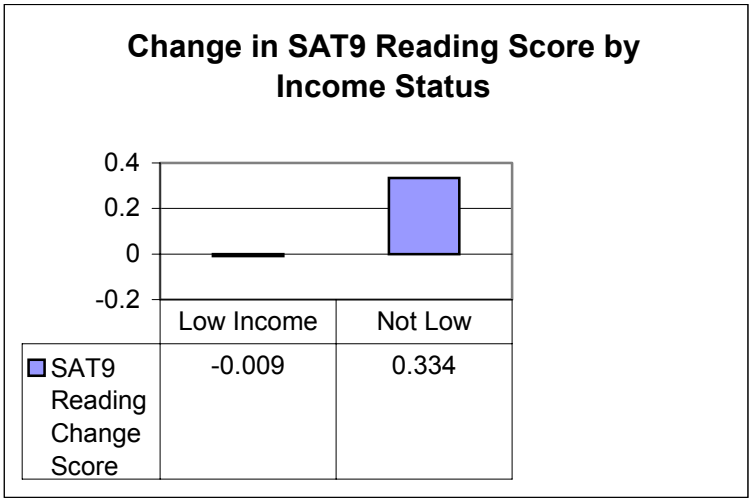
Originally test score gains were to be examined as a function of gender, race, income status, special education status, LEP status, and Title I classifications. However, several of these categories have too few students in them to allow for a reasonable test of the question under consideration. Therefore, test score gains were examined as a function of gender, low income status, and race. For the purpose of this analysis, only African Americans and Caucasians were included in the analysis because of the small number of students represented within the other racial categories.

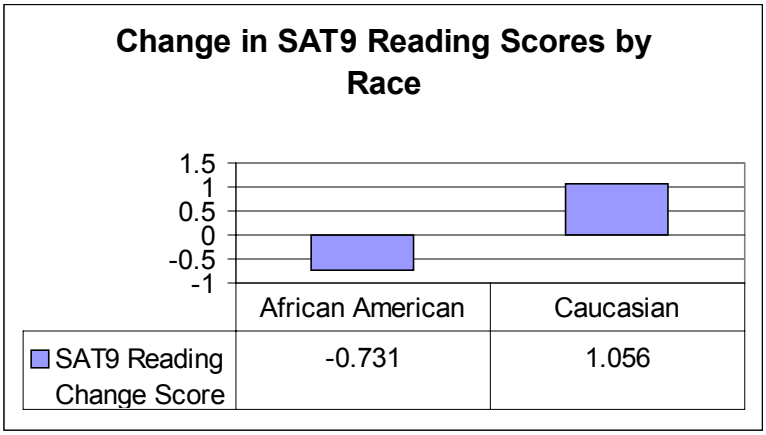
Method

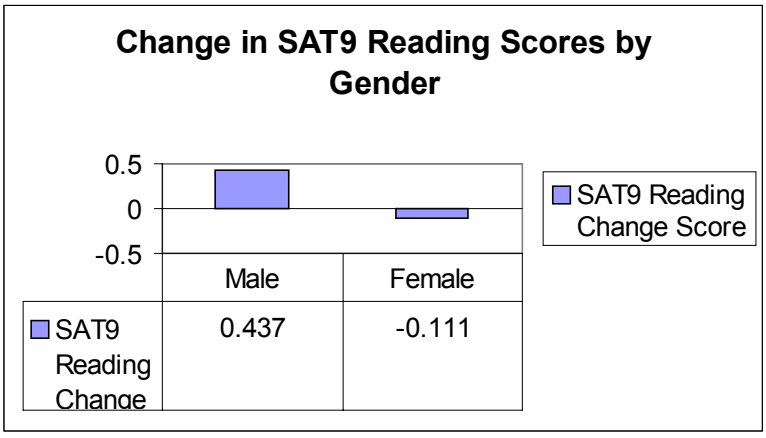
Test score gains were examined by first computing a difference score for each student by subtracting the student's score in 1998 from their score in 2000. Three separate 2 (gender: male vs. female) X 2(race: African American vs. Caucasian) X 2 (income status: low income vs. not low income) factorial analyses of variance (ANOVA) were conducted. The gender, race, and income status of the student were used as categories for comparing change scores. This analysis explored differences between low income and non-low income students, between males and females, and between African Americans and Caucasians. The analysis also allowed the researcher to determine if the variables under examination had a combined influence, or interactive effect.

Changes in Reading Performance Between 1998 and 2000

The results of the factorial ANOVA examining the rate of change in SAT9 reading scores revealed a statistically significant effect for race such that Caucasian students had larger change scores than African Americans. The size of this effect, however, was negligible. No differences were found for income status, gender or the interaction of these variables. The mean change score for each student subgroup is presented on the following pages.



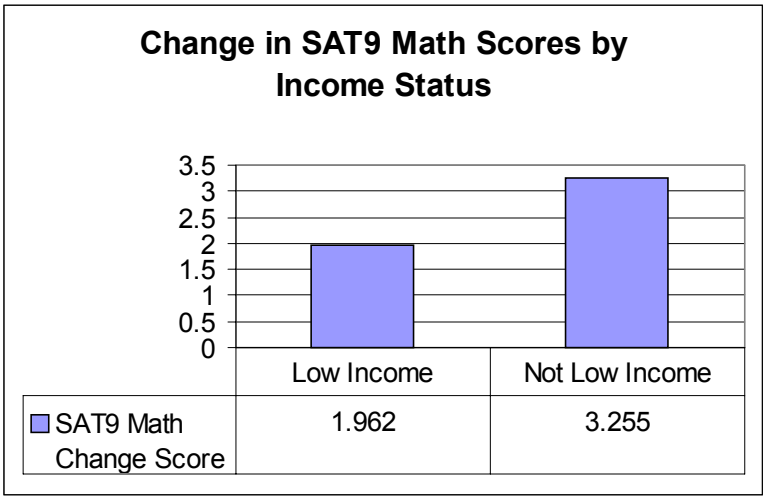


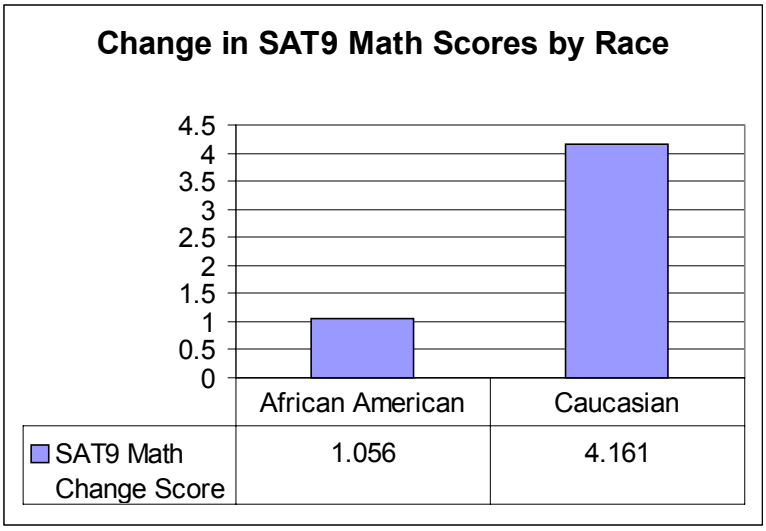


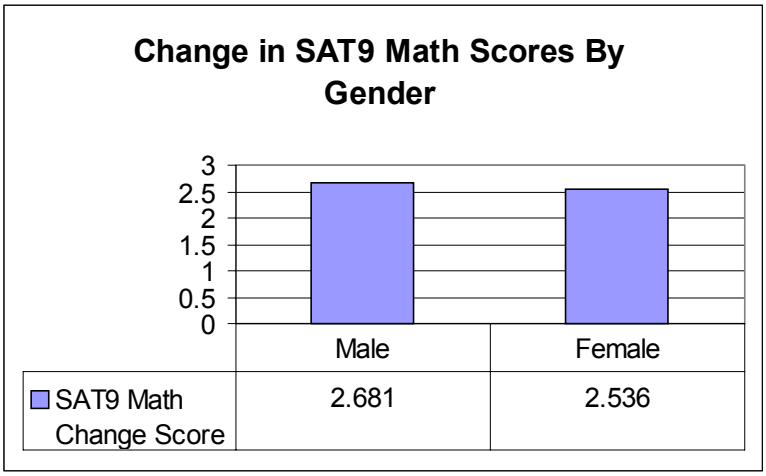
Changes in Mathematics Performance Between 1998 and 2000

The results of the factorial ANOVA examining the rate of change in SAT9 mathematics scores revealed a statistically significant effect for race and low income status such that Caucasian students had larger change scores than African Americans and non-low income students had

larger change scores than low income students. The size of these effects, however, were negligible. No differences were found as a function of gender or the interaction among the variables. The mean change score for each student subgroup is presented on the following pages.

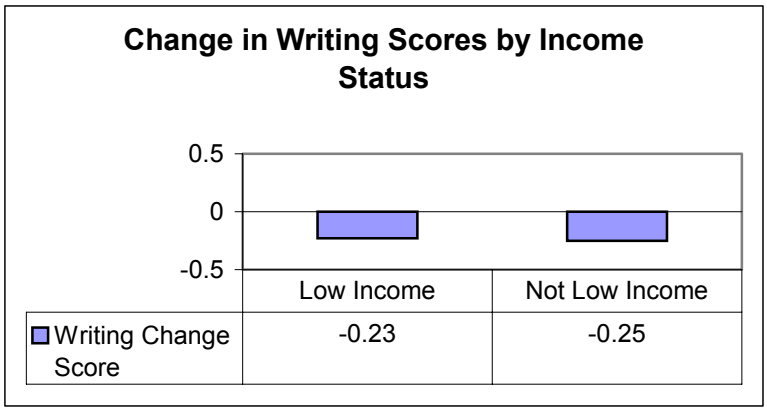


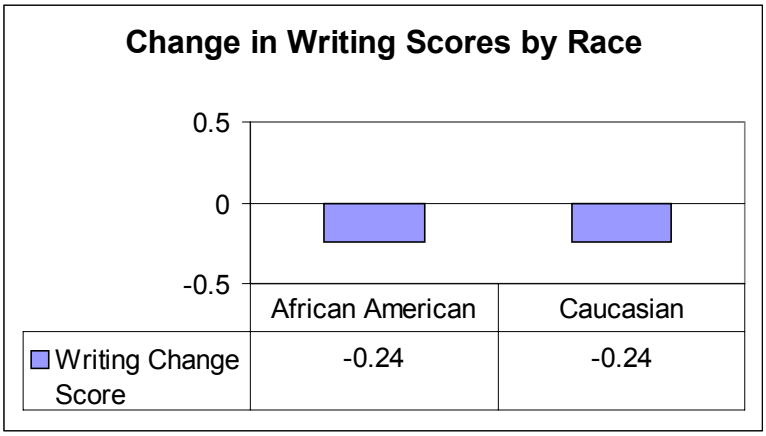


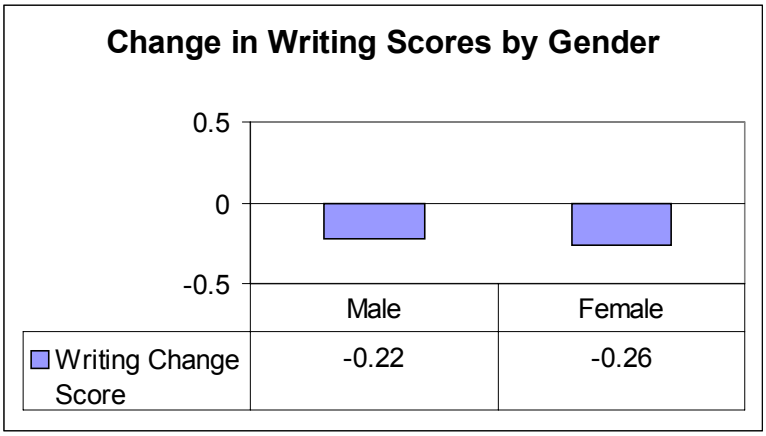


Changes in Writing Performance Between 1998 and 2000

The results of the factorial ANOVA examining the rate of change in writing scores did not result in any statistically significant effects. The rate of change was the same across all study subgroups. All student groups evidenced a decline in writing scores from 1998 to 2000.





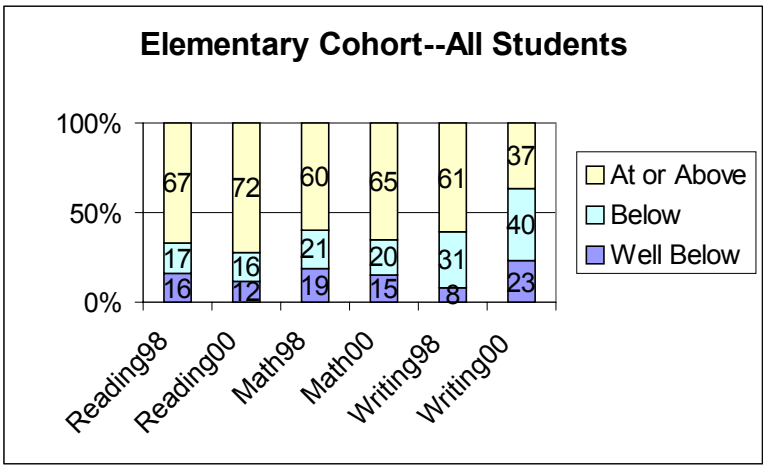


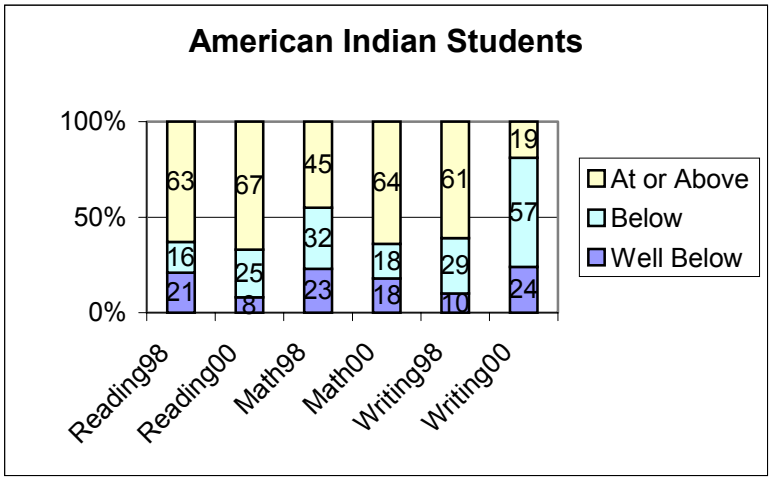
In summary, the results of the factorial ANOVAs revealed no real differences in change scores across student groups. It may be worth noting that changes in the SAT9 reading scores of African Americans and female students was in the negative direction and all students evidenced a decline in writing scores over time.

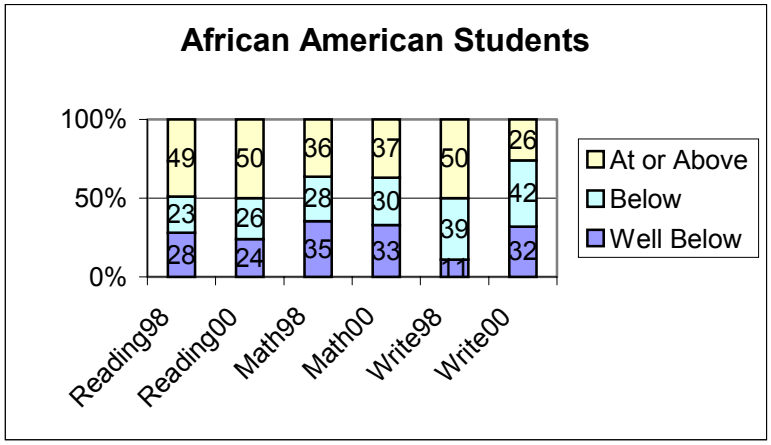
Question 2c: How much movement is there in performance levels on the DSTP from one test administration to the next?

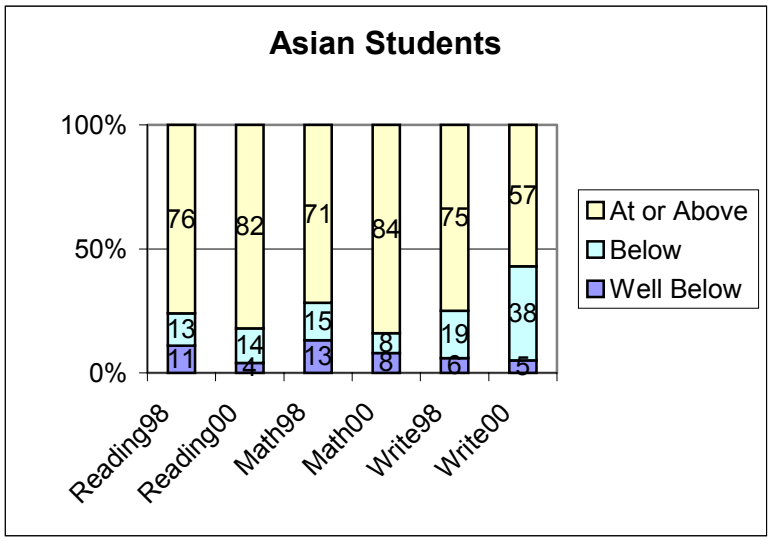
Descriptive statistics were used to examine how much movement in DSTP performance levels occurred from 1998 to 2000. Students were classified as performing at or above the standard, below the standard, or well below the standard based on the cut-scores associated with each content area on the DSTP.[12] The following graphs show the percentage of students performing at each of these three levels for the reading, math, and writing portions of the DSTP. Overall, more students were performing at or above the standard in reading and mathematics in 2000 compared to 1998 and fewer students were performing at or above the standard in writing in 2000 as compared to 1998.

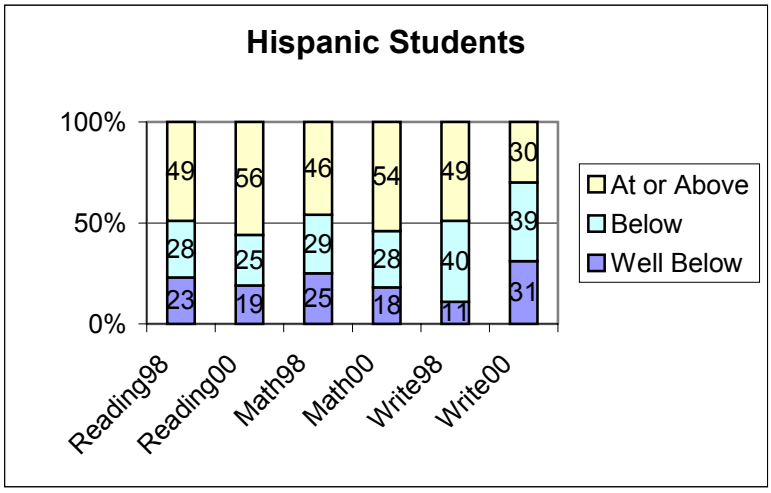
Elementary Cohort: Performance Levels for 1998 and 2000

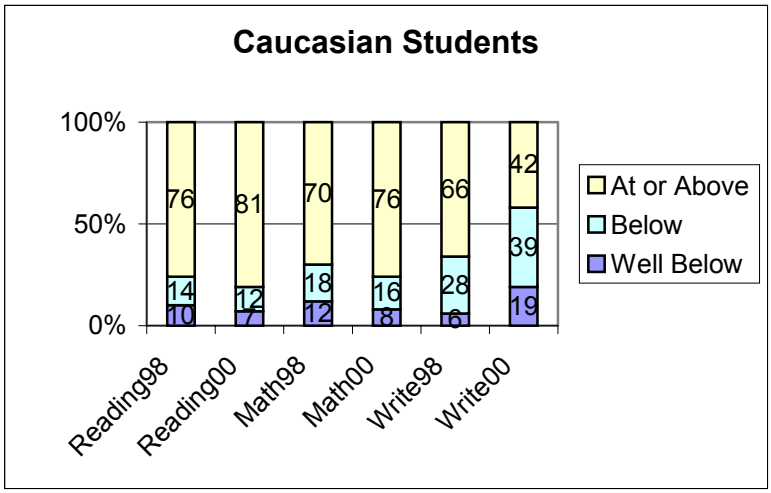


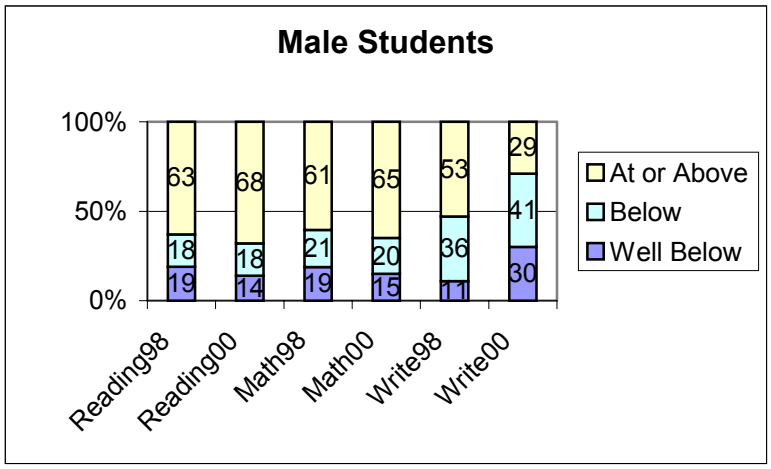


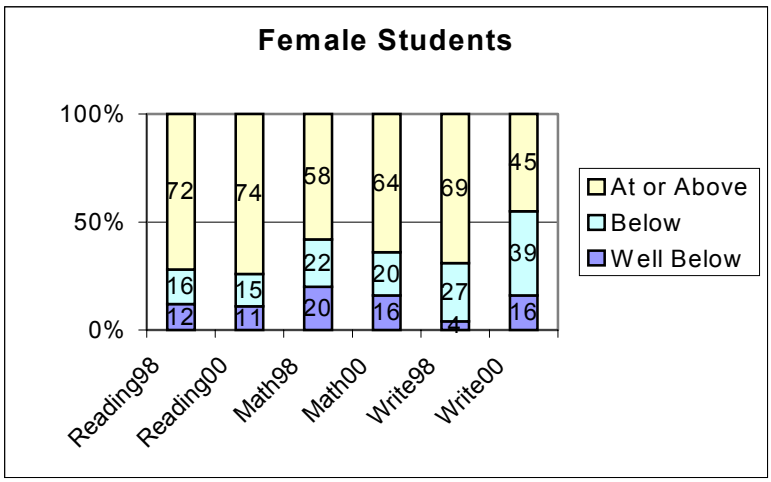

African American Students
100%
50%
0%
49
23
28
50
26
24
36
28
35
37
30
33
50
39
11
26
42
32
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

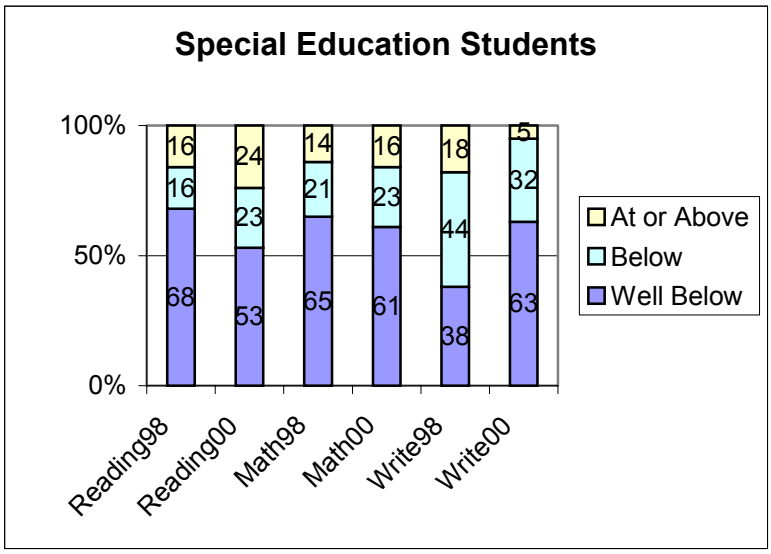

Asian Students
100%
50%
0%
76
13
11
82
14
4
71
15
13
84
8
8
75
19
6
57
38
5
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

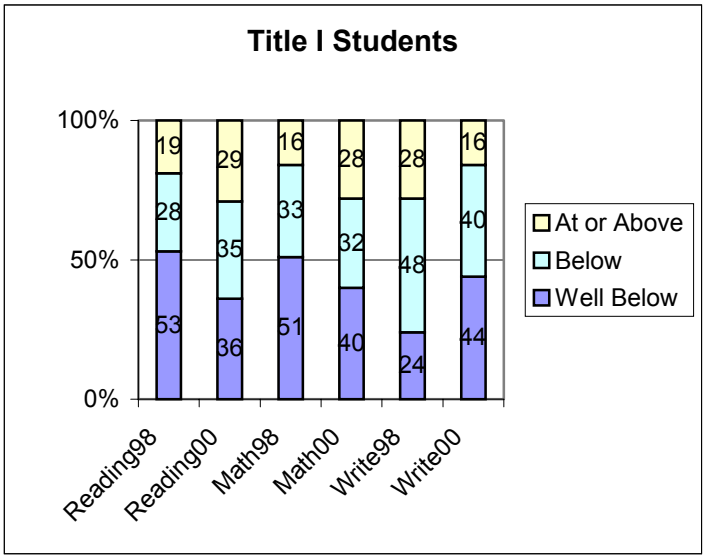

Hispanic Students
100%
50%
0%
49
28
23
56
25
19
46
29
25
54
28
18
49
40
11
30
39
31
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

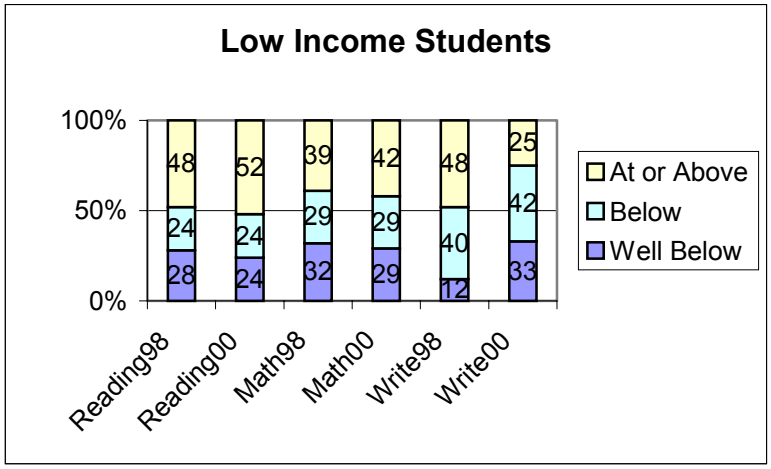

Caucasian Students
100%
50%
0%
76
14
10
81
12
7
70
18
12
76
16
8
66
28
6
42
39
19
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

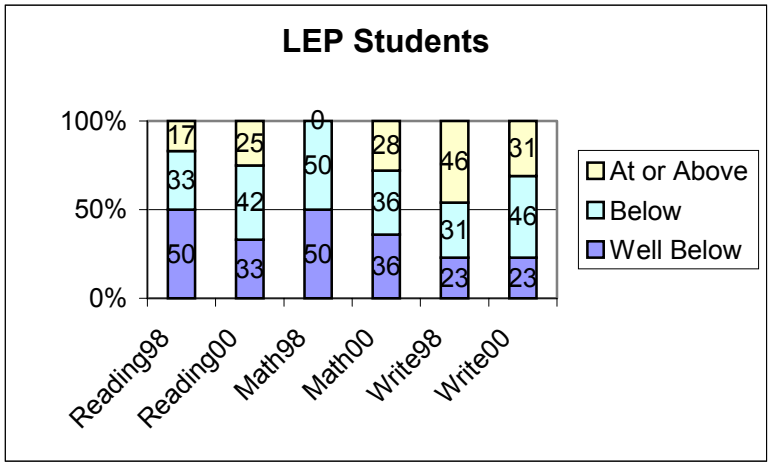

Male Students
100%
50%
0%
63
18
19
68
18
14
61
21
19
65
20
15
53
36
11
29
41
30
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00


Female Students
100%
50%
0%
72
16
12
74
15
11
58
22
20
64
20
16
69
27
4
45
39
16
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00


Special Education Students
100%
50%
0%
16
16
68
24
23
53
14
21
65
16
23
61
18
44
38
5
32
63
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

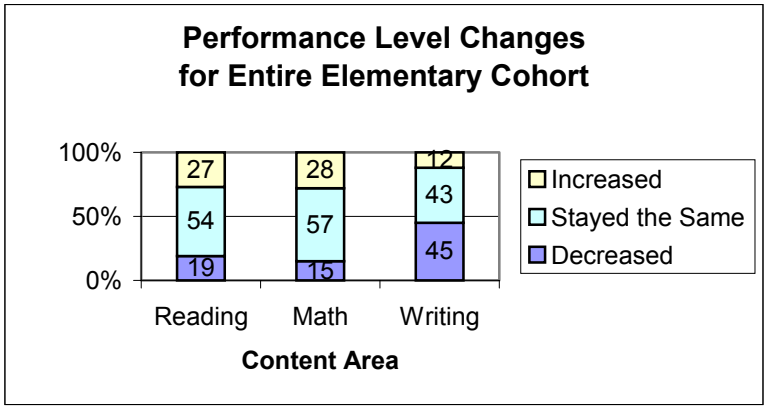

Title I Students
100%
50%
0%
19
28
53
29
35
36
16
33
51
28
32
40
28
48
24
16
40
44
At or Above
Below
Well Below
Reading98
Reading00
Math98
Math00
Write98
Write00

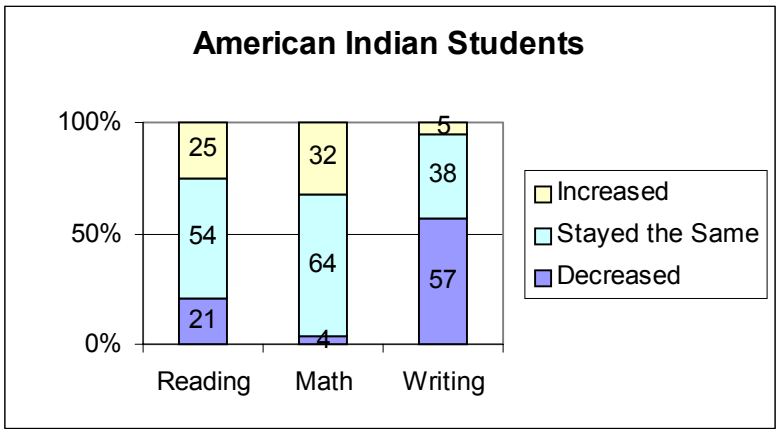


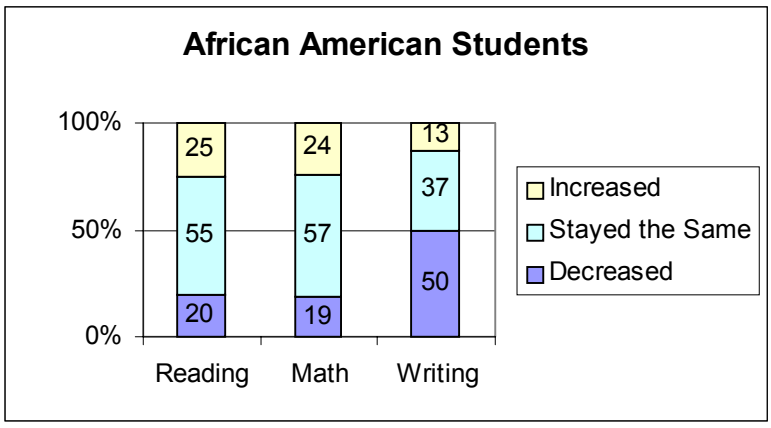


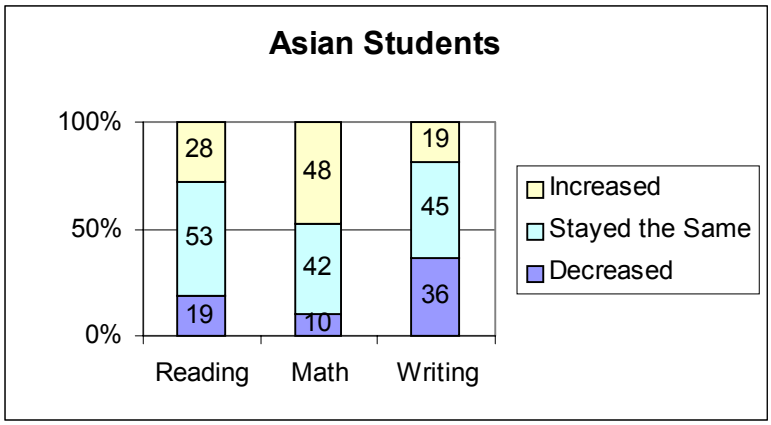
Elementary Cohort: Performance Level Changes between 1998 and 2000

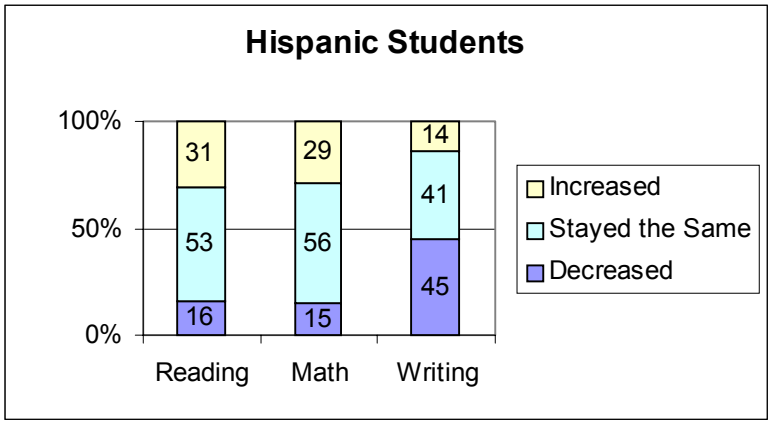
The previous analyses provide information on the percentage of students classified into each of the various performance levels in 1998 and 2000. These analyses do not address how much movement occurred on an individual basis. In order to examine this issue, simple counts were made of the number of students who performed at the same level both times, the number of students who declined by at least one performance level, and the number of students who improved by at least one performance level. This information is displayed on the following pages.

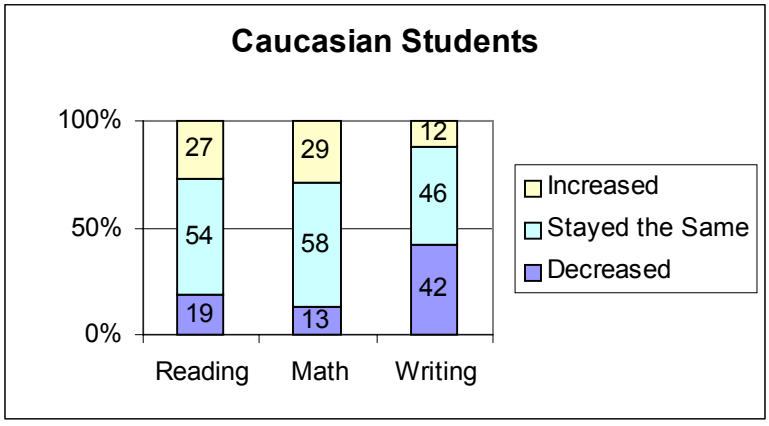

Performance Level Changes
for Entire Elementary Cohort
100%
50%
0%
27
54
19
28
57
15
12
43
45
Reading
Math
Writing
Content Area
Increased
Stayed the Same
Decreased

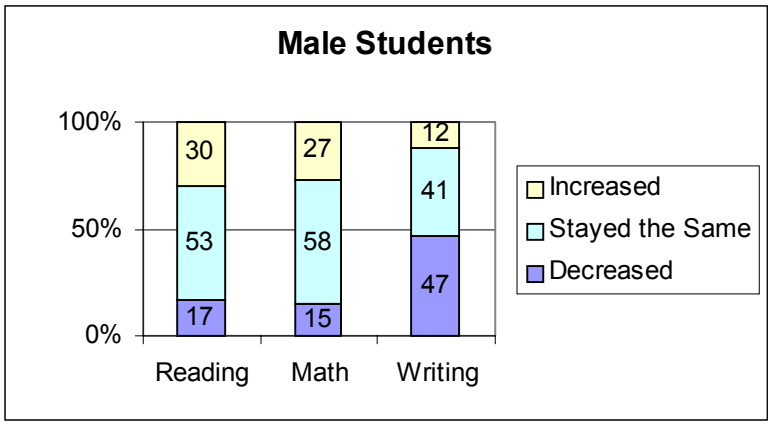

American Indian Students
100%
50%
0%
25
54
21
32
64
4
5
38
57
Reading
Math
Writing
Increased
Stayed the Same
Decreased

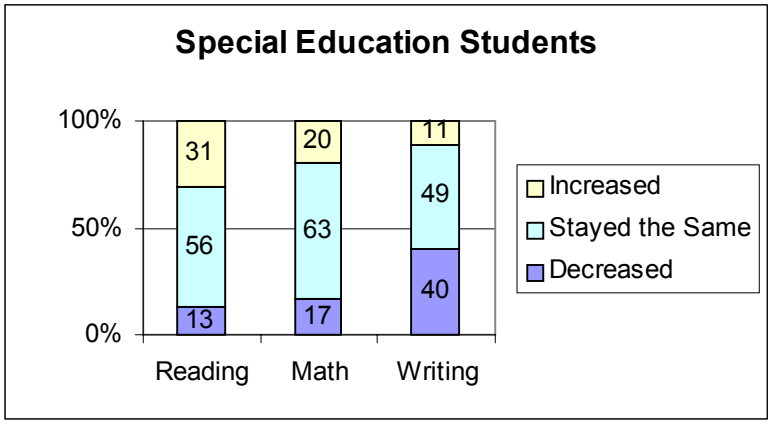

African American Students
100%
50%
0%
25
55
20
24
57
19
13
37
50
Reading
Math
Writing
Increased
Stayed the Same
Decreased

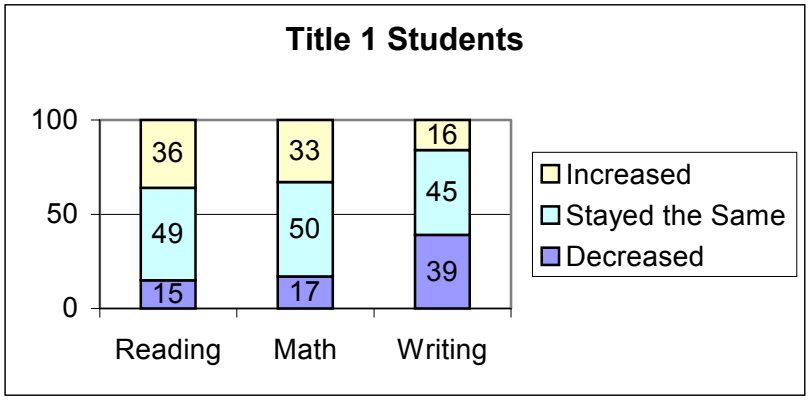

Asian Students
100%
50%
0%
28
53
19
48
42
10
19
45
36
Reading
Math
Writing
Increased
Stayed the Same
Decreased

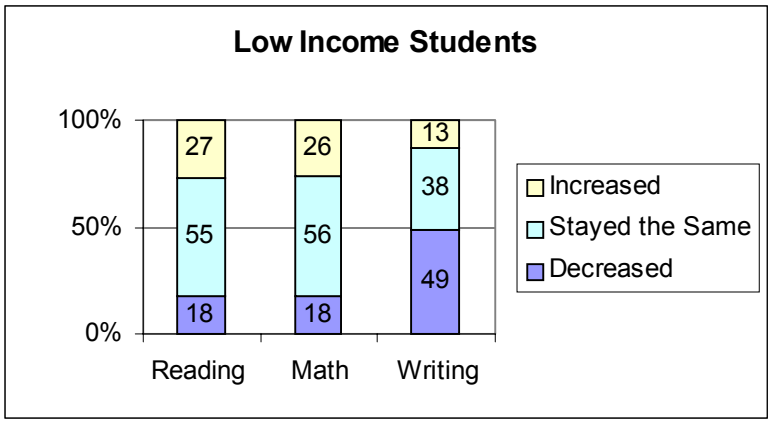

Hispanic Students
100%
50%
0%
31
53
16
29
56
15
14
41
45
Reading
Math
Writing
Increased
Stayed the Same
Decreased

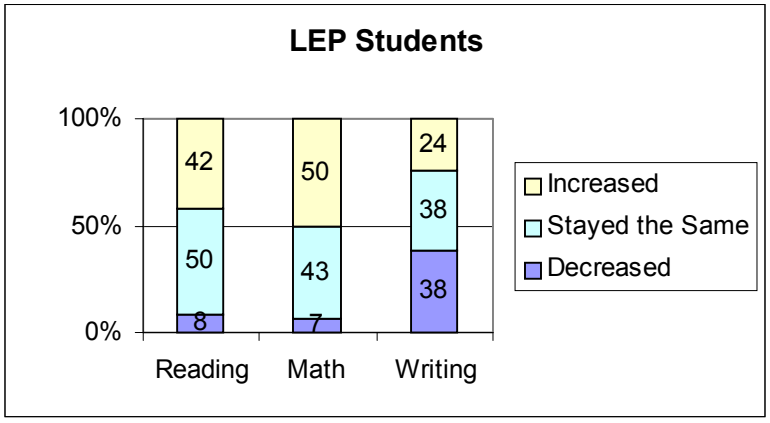

Caucasian Students
100%
50%
0%
27
54
19
29
58
13
12
46
42
Reading
Math
Writing
Increased
Stayed the Same
Decreased


Male Students
100%
50%
0%
30
53
17
27
58
15
12
41
47
Reading
Math
Writing
Increased
Stayed the Same
Decreased


Special Education Students
100%
50%
0%
31
56
13
20
63
17
11
49
40
Reading
Math
Writing
Increased
Stayed the Same
Decreased


Title 1 Students
100
50
0
36
49
15
33
50
17
16
45
39
Reading
Math
Writing
Increased
Stayed the Same
Decreased

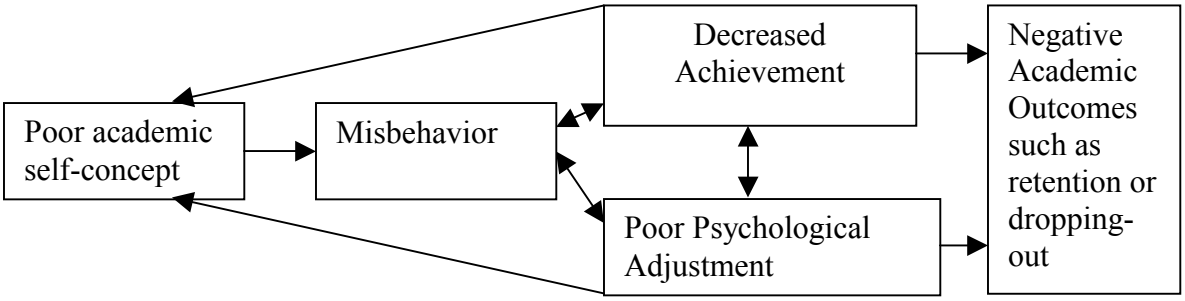




The results of this analysis revealed that for reading and mathematics, a majority of students performed at the same performance level in 1998 as they did in 2000. The two exceptions were Asian students and LEP students who both evidenced increases in the number of students who improved their performance levels in mathematics from 1998 to 2000. In addition, for all student groups, writing was the content area in which the largest number of students performed at a lower level in 2000 as compared to 1998.

Issue 3: Behavioral Effects

Research in the area of school discipline indicates that student conduct problems in the classroom are often a precursor to later school dropout and other negative social outcomes such as poorer psychological adjustment and poorer academic self-concepts. There is some research to indicate that children who believe they are poor students often have behavior problems and these problems worsen as student achievement declines.[13] All of these factors could combine to create a negative perpetuating cycle which could be conceptualized as follows:

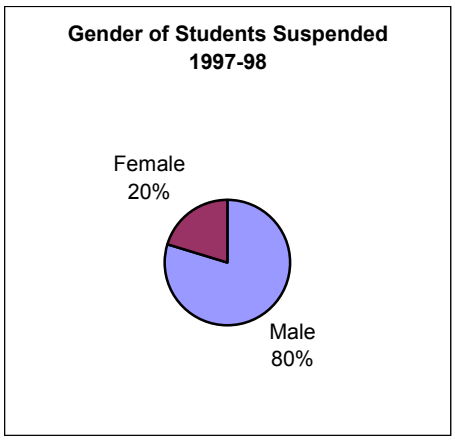


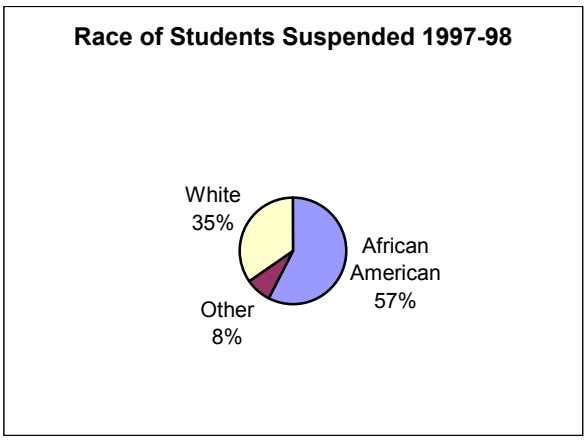
Although this model is not formally tested in the current study, the link of achievement to student behavior was investigated by examining the incidence of out-of-school suspension and its relationship to performance on the DSTP. In addition, previous research on out-of-school suspension indicates that minority students are often over-represented in out-of-school suspension rates.[14] Therefore the demographic characteristics of students given out-of-school suspensions were also examined.

Question 3a: What are the demographic characteristics of students who are given out-of-school suspension?

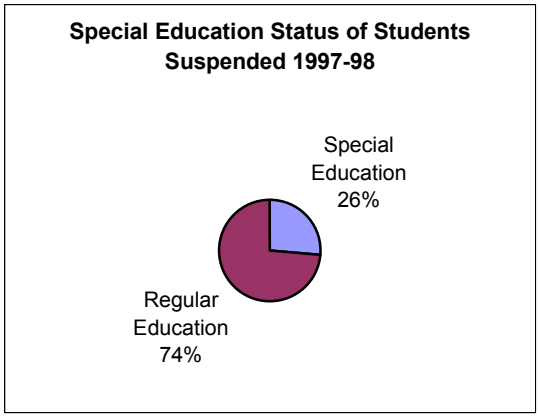
Elementary Cohort: Suspensions During the 1997-98 School Year

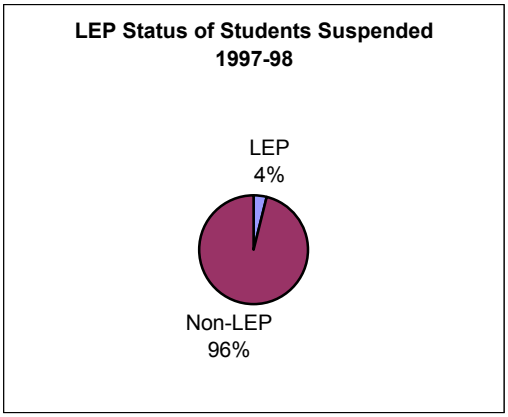
Within the ***elementary cohort***, 345 students were suspended during the 1997-98 school year. The total number of out-of-school suspensions received by any single student ranged from 1 to 15 with most students receiving only one out-of-school suspension. The total number of days spent out of school ranged from 1 to 68 days, with 90% of students spending 6 days or fewer out of school. The demographic characteristics of the students given out-of-school suspension were as follows:

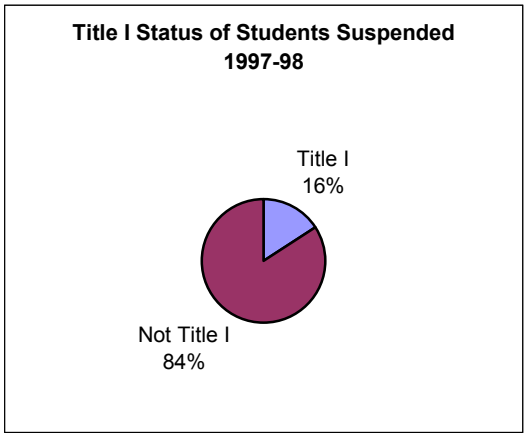


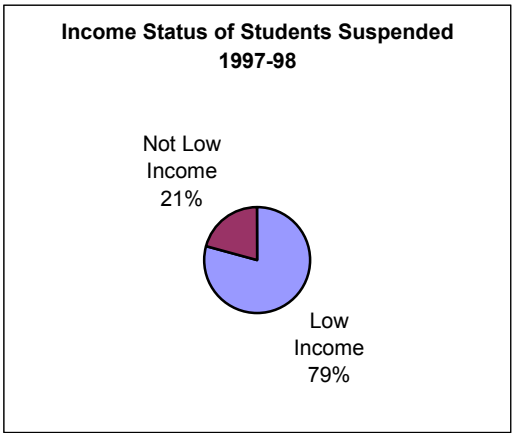


Note: American Indians, Asians, and Hispanics were combined into the "other" category.





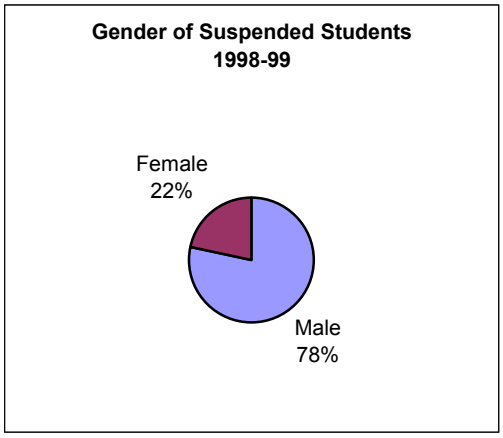


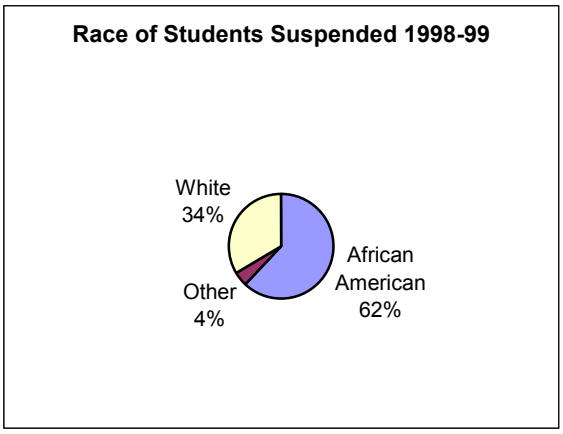


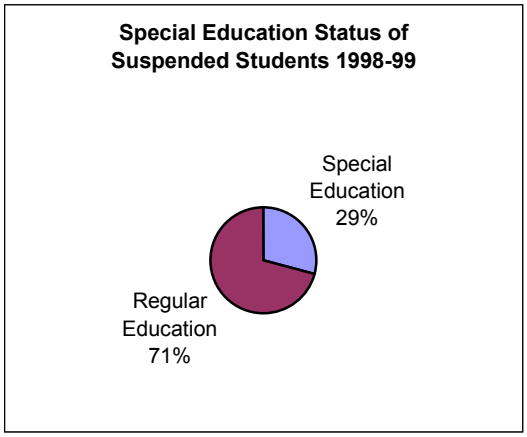
Elementary Cohort: Suspensions During the 1998-99 School Year

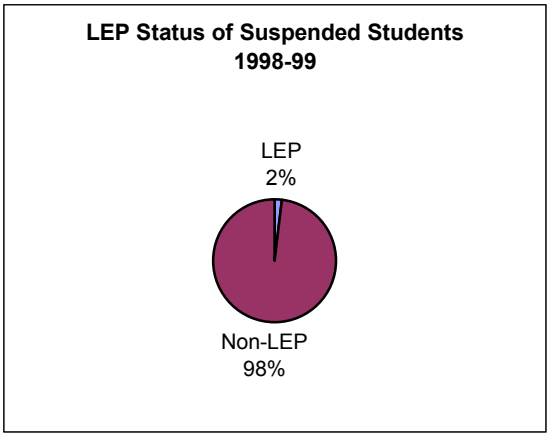
Within the ***elementary cohort***, 454 students were suspended during the 1998-99 school year. The total number of out-of-school suspensions received by any single student ranged from 1 to 10 with most students receiving only one out-of-school suspension. The total number of days spent out of school ranged from 1 to 26 days, with 90% of students

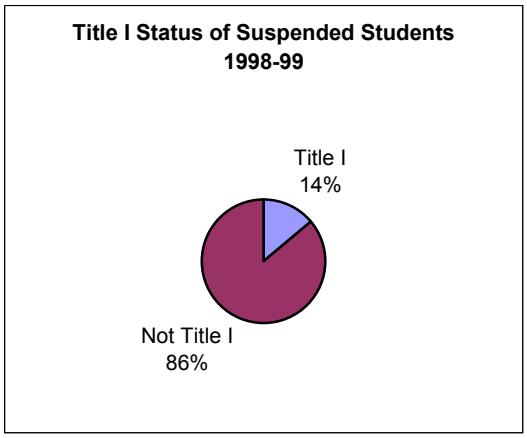
spending less than 7 days out of school. The demographic characteristics of the students given out-of-school suspension were as follows:

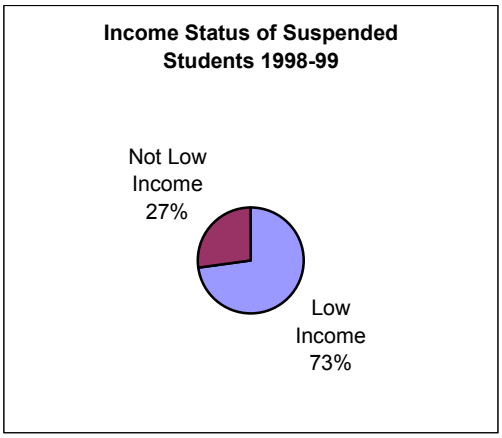






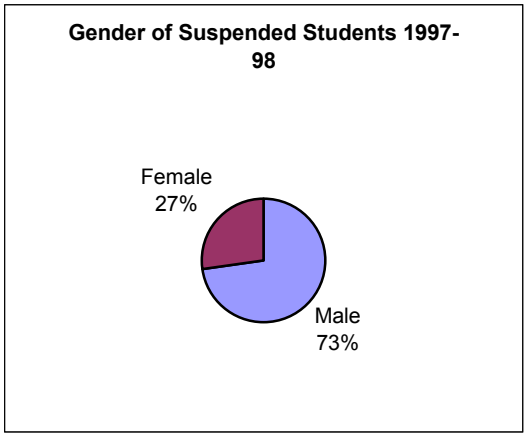


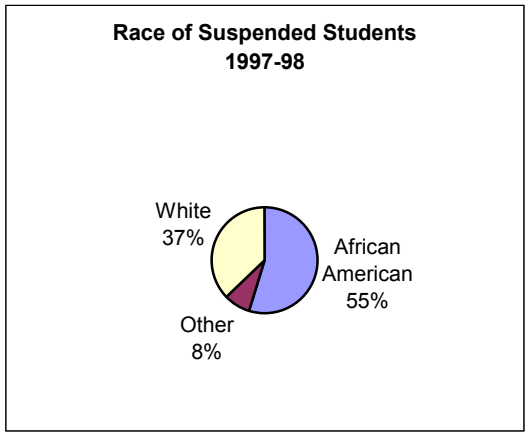


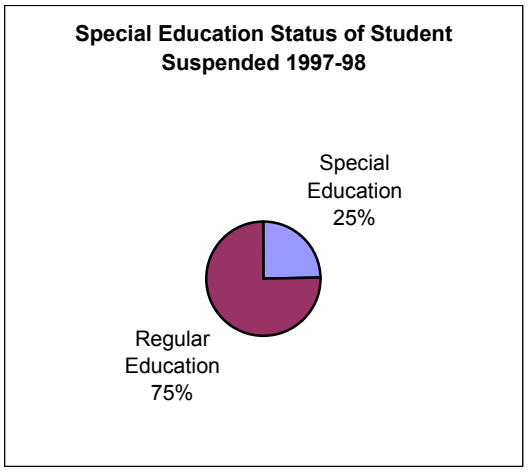


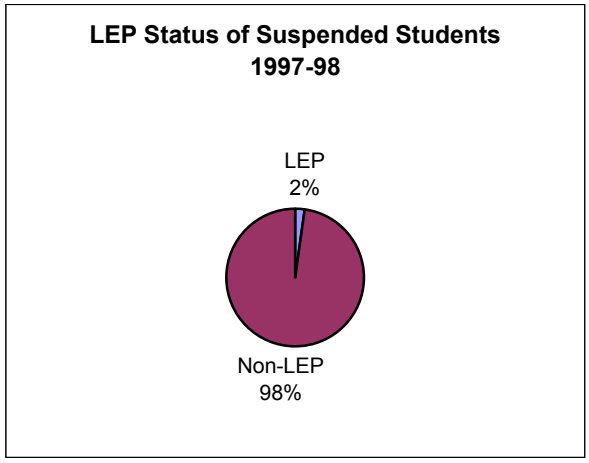
Secondary Cohort: Suspensions During the 1997-98 School Year

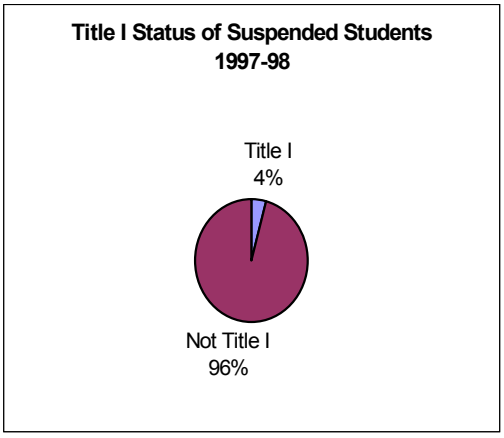
Within the ***secondary cohort***, 1039 students were suspended during the 1997-98 school year. The total number of out-of-school suspensions received by any single student ranged from 1 to 15 with most students receiving only one out-of-school suspension. The total number of days spent out of school ranged from 1 to 38 days, with 90% of students spending 9 days or fewer out of school. The demographic characteristics of the students given out-of-school suspension were as follows:

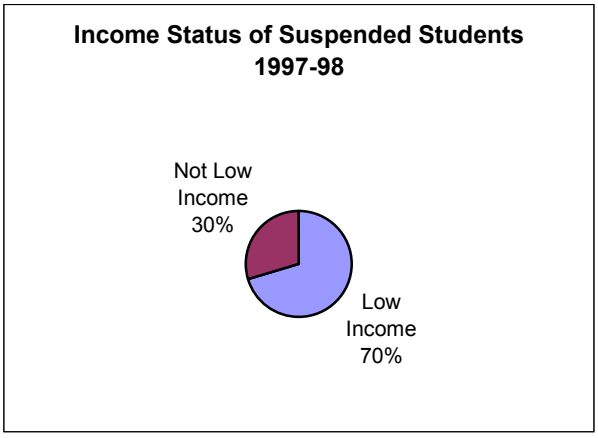






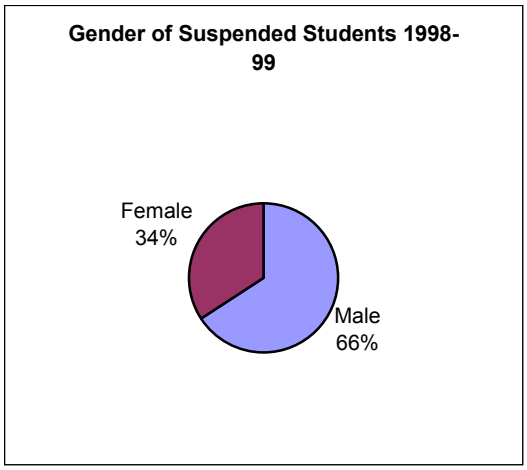


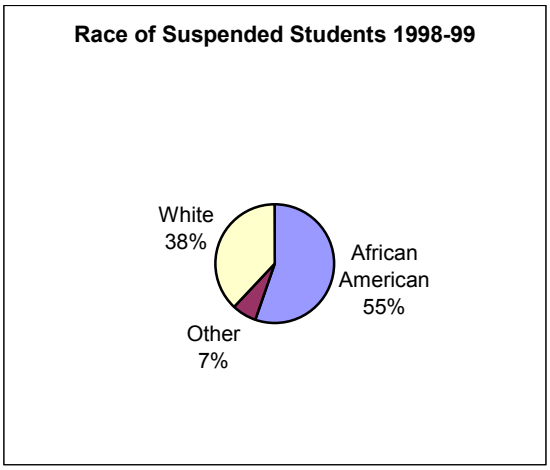


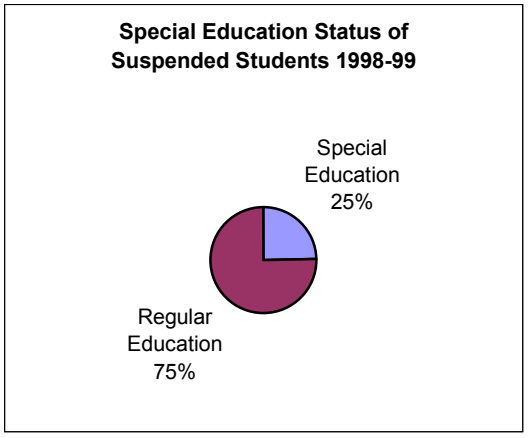


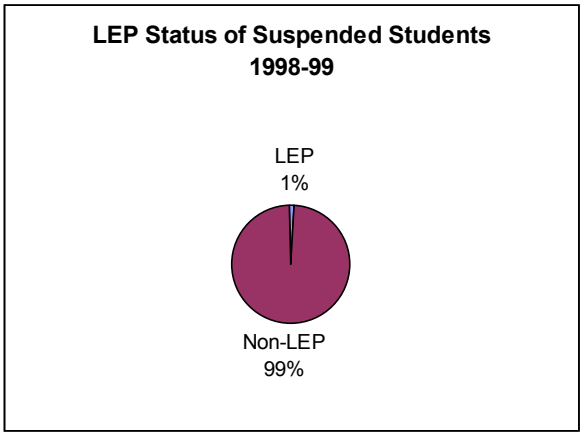
Secondary Cohort: Suspensions During the 1998-99 School Year

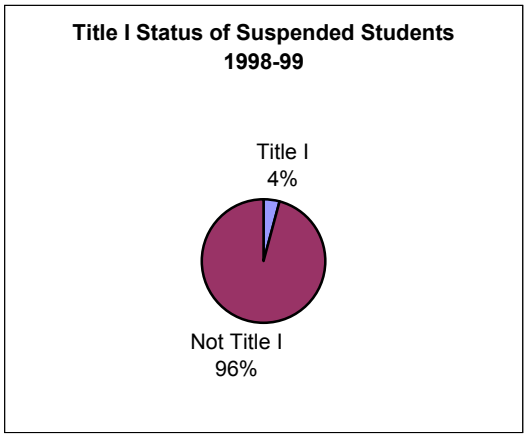
Within the ***secondary cohort***, 1663 students were suspended during the 1998-99 school year. The total number of out-of-school suspensions received by any single student ranged from 1 to 19 with most students receiving only one out-of-school suspension. The total number of days spent out of school ranged from 1 to 189 days, with 90% of students spending 13 days or fewer out of school. The demographic characteristics of the students given out-of-school suspension for the 1998-99 school year were as follows:

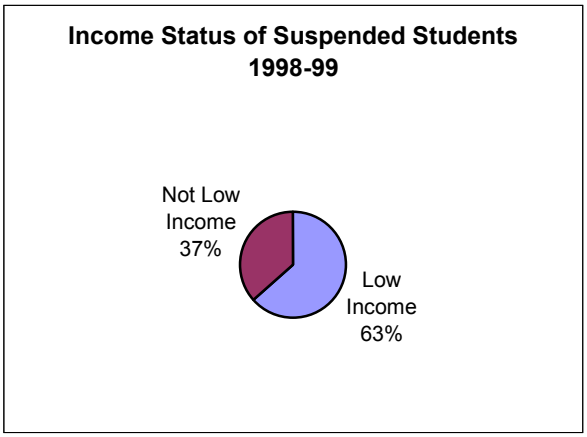












A Closer Examination of the Demographic Characteristics Associated with OSS

An examination of the demographic characteristics associated with out-of-school suspension suggested that students who are low income, of a minority status, or male may be over-represented in suspensions. In order to address this issue statistically, logistic regression analyses were performed.

Logistic regression is a statistical procedure for estimating the relationship between one or more predictor variables and the likelihood that an individual is a member of a particular group.[15] For the purposes of the current investigation, group membership was defined as whether or not a student had been given a suspension during the school year. Gender, minority status (minority vs. non-minority), and income status (low income vs. not low income) were used to predict whether or not a student was suspended.

For the elementary cohort, the results of the logistic regression revealed that income status, gender, and minority status were statistically significant predictors of suspensions given during the 1997-98 school year. According to the logistic regression, the odds of being suspended were 1.93 times greater for minority students compared to non-minorities, 3.69 times greater for males than females, and 3.79 times greater for low

income students than students who were not low income. The logistic regression analysis examining suspensions for the elementary cohort during the 1998-99 school year yielded similar results.

For the secondary cohort, the results of the logistic regression revealed that income status, gender and minority status were statistically significant predictors of suspensions given during the 1997-98 school year. The results also revealed that the odds of being suspended were 2.22 times greater for minority students compared to non-minorities, 3.08 times greater for males than females, and 2.83 times greater for low income students than non-low income students. The logistic regression analysis examining suspensions for the secondary cohort during the 1998-99 school year yielded similar results.

Question 3b: What is the relationship of out-of-school suspension to performance on the DSTP?

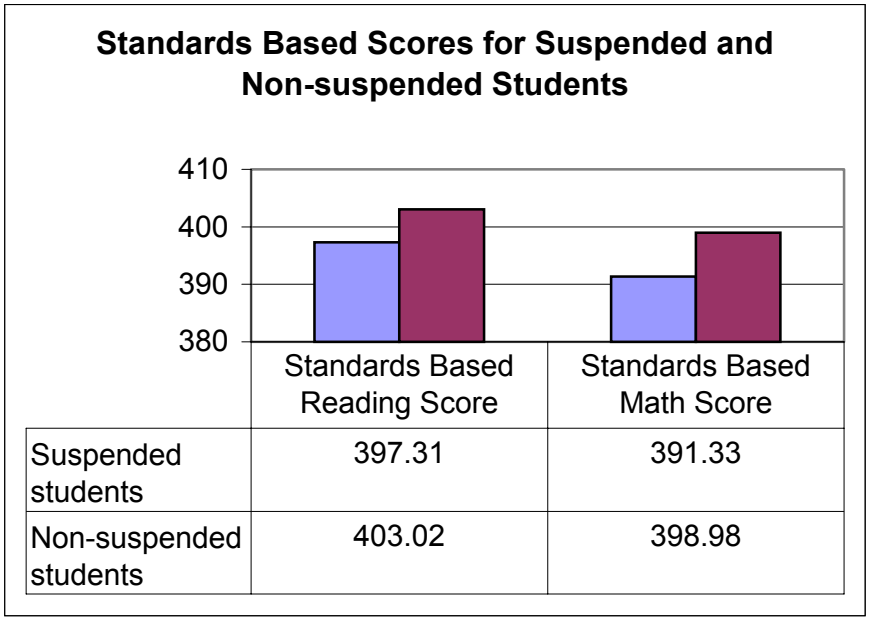
Given that out-of-school suspension may limit a student's opportunity to learn, it is important to examine the relationship between suspension and performance on the DSTP. At the time of this report, only suspension data from the 1997-98 and 1998-99 school year were available. Therefore the question of the relationship between out-of-school suspension and performance on the DSTP could only be examined through use of the elementary cohort.

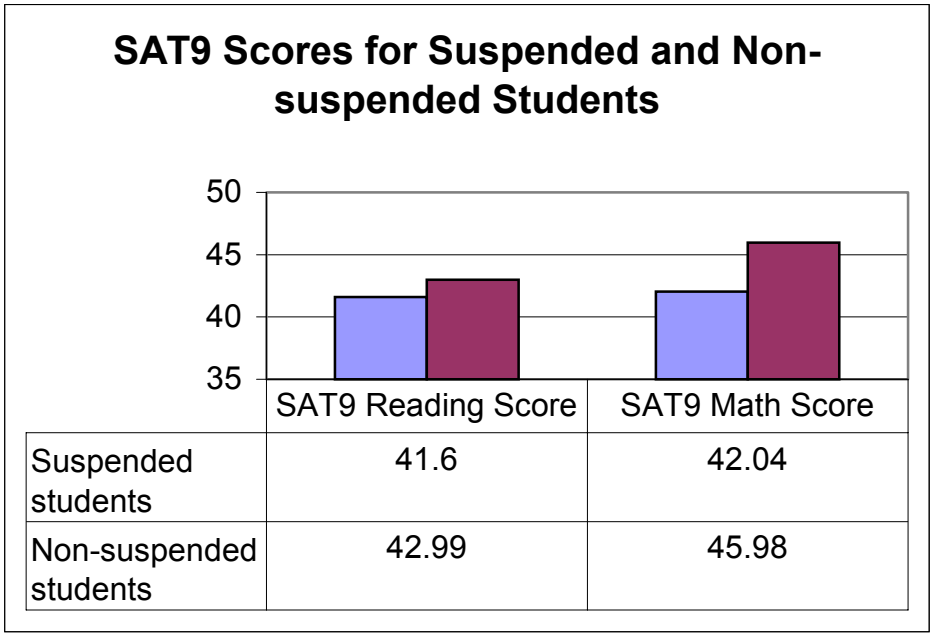
Method

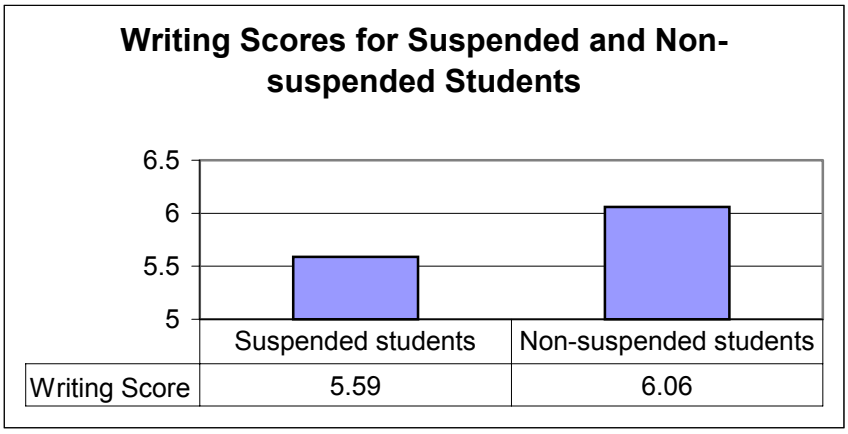
To address this question, students from the elementary cohort who were suspended during the 1997-98 school year were matched with non-suspended students on the basis of race and income status. These two student characteristics were used for matching purposes because of the demographic patterns that emerged in question 3a.

Paired sample t-tests were computed to compare suspended and non-suspended students on their DSTP performance from the spring of 1998. The analyses examined the standards based scores as well as the SAT9 reading and mathematics scores.

The results of the paired sample t-tests revealed statistically significant differences between suspended and non-suspended students in terms of their standards based mathematics score, the SAT9 mathematics score, and the writing portion of the DSTP. In each case suspended students performed at a lower level, however the effect sizes of these effects were all negligible.







The results indicated that suspended students did not evidence any real performance differences from students who were not suspended. However, the finding that the odds of receiving a suspension were greater for minority students coupled with the finding that African Americans were one of the few student groups who did not evidence a slight improvement in SAT9 math scores from 1998 to 2000 does raise some concern (see section 2a).

Issue 4: Retention

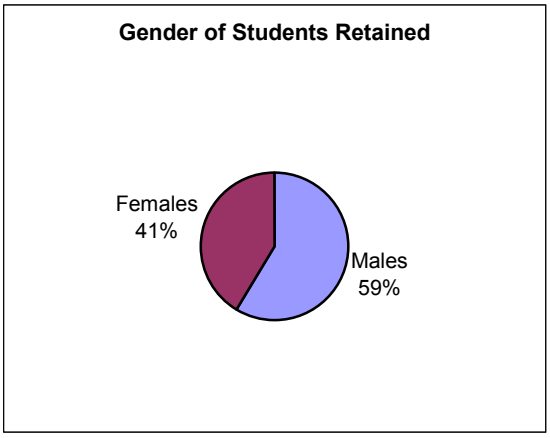
Research on high-stakes assessment systems reveals that retention rates often increase when such systems are implemented. There is also a great deal of evidence that grade retention, as typically practiced, is ineffective if not harmful and that certain groups of students are more likely to be retained. For example, retention is twice as likely to occur among African American students as among Caucasians, is more likely to affect students from low-income families, and is more prevalent among boys.[16] In addition, students who are retained often exhibit signs of poorer personal adjustment and feel stigmatized by the retention. One study indicates that students rank grade retention as the third most feared life experience behind blindness and the death of a parent.[17]

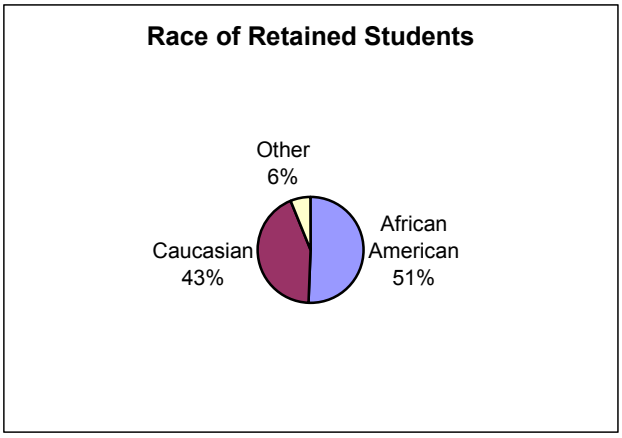
For these reasons, the incidence of retention and the demographic characteristics of retained students were examined in two ways. First, the demographic characteristics of students who were enrolled in the same grade in 1998-99 as they were in 1997-98 were examined. Second, students ***who would have been retained*** if the consequences associated with their DSTP performance were in place were also examined. For the elementary cohort, the latter group was composed of those students performing at the lowest performance level on the reading portion of the DSTP. For the secondary cohort, the latter group was composed of those students performing at the lowest performance level on the reading and mathematics portion of the DSTP.

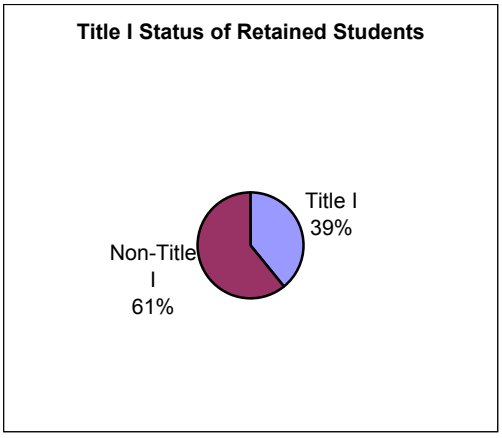
Question 4a: What are the demographic characteristics of students who are retained or would have been retained if the consequences for performance had been in effect?

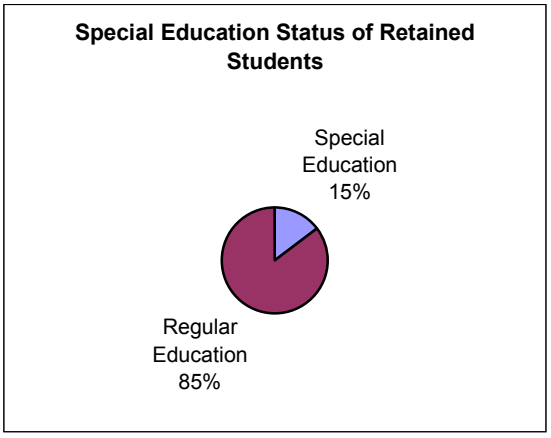
Elementary Cohort: Actual Retainees

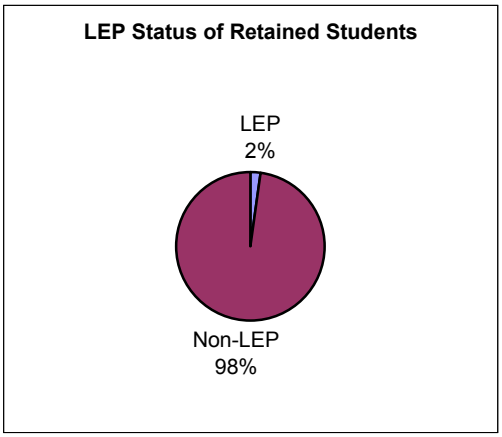
The demographic characteristics of students in the elementary cohort who were enrolled in grade 3 during the 1997-98 ***and*** 1998-99 school year (n=184) were as follows:

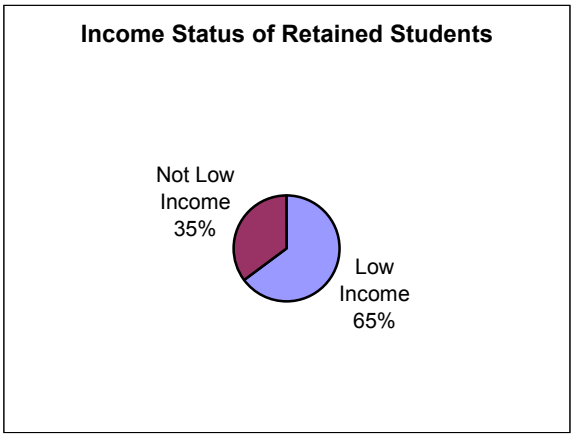








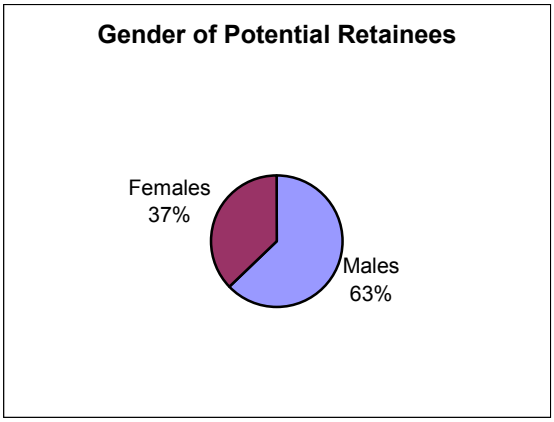


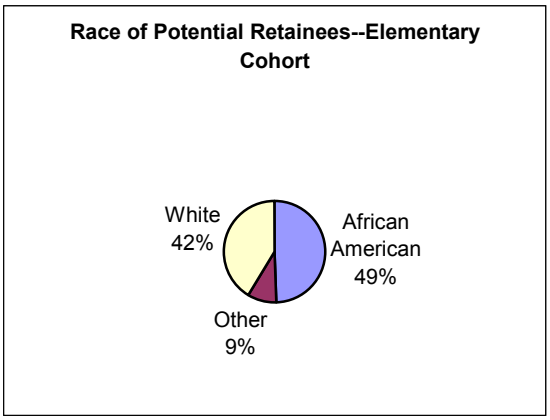


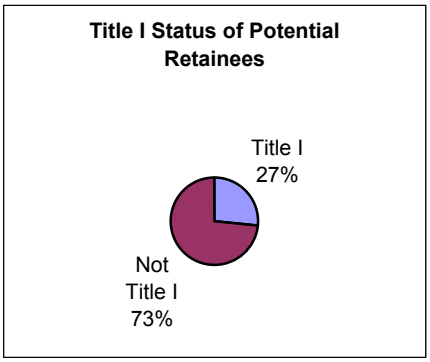
An examination of the demographic characteristics associated with retention suggested that students who were of low income and minority status may have been over-represented in retentions. A logistic regression analysis was performed in which income status, gender and minority status were used to predict who was retained in 3rd grade. This analysis revealed that minority status was a useful predictor of retention while gender and income status was not. The results revealed that the odds of being retained were 2.18 times greater for minority students compared to non-minorities.

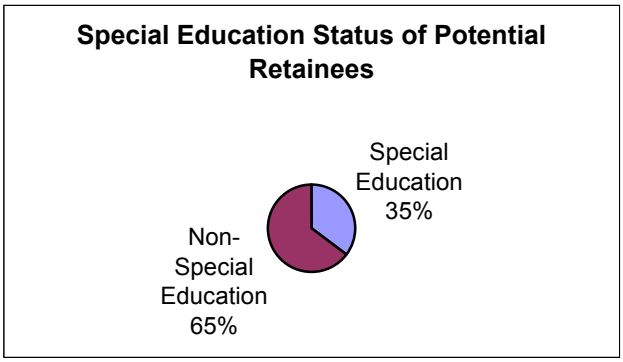
Elementary Cohort: Potential Retainees

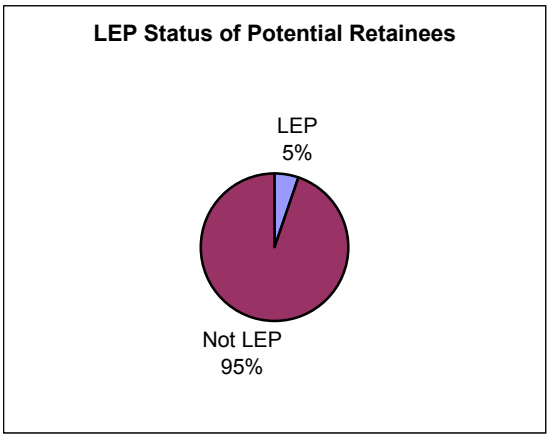
In the future, decisions about retaining students will be based primarily on students' performance on the DSTP. Third grade students scoring well below the standard in reading who do not meet the standard before the start of the following year may be retained. The demographic characteristics of students in the elementary cohort ***who might have been retained if the consequences had been in effect*** (n=1506), were as follows:

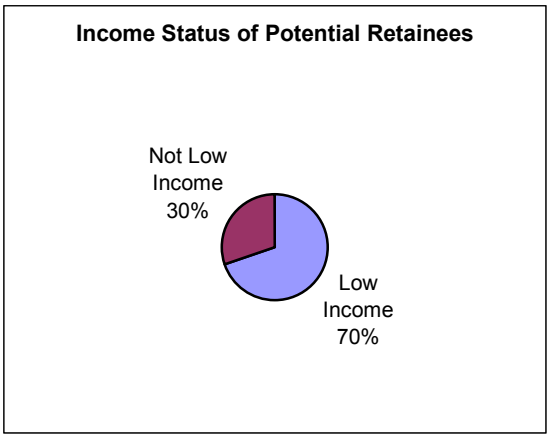

Gender of Potential Retainees
Females
37%
Males
63%


Race of Potential Retainees--Elementary Cohort
White
42%
African American
49%
Other
9%


Title I Status of Potential Retainees
Title I
27%
Not
Title I
73%


Special Education Status of Potential Retainees
Special
Education
35%
Non-
Special
Education
65%

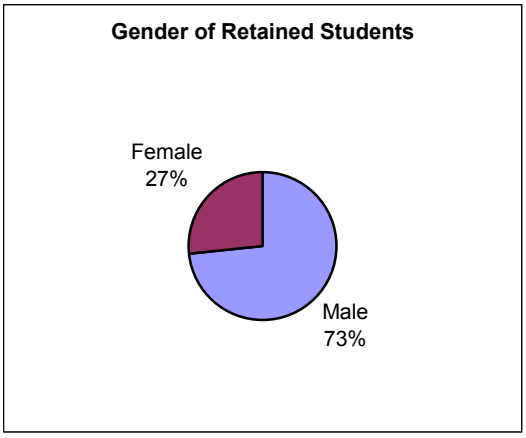

LEP Status of Potential Retainees
LEP
5%
Not LEP
95%

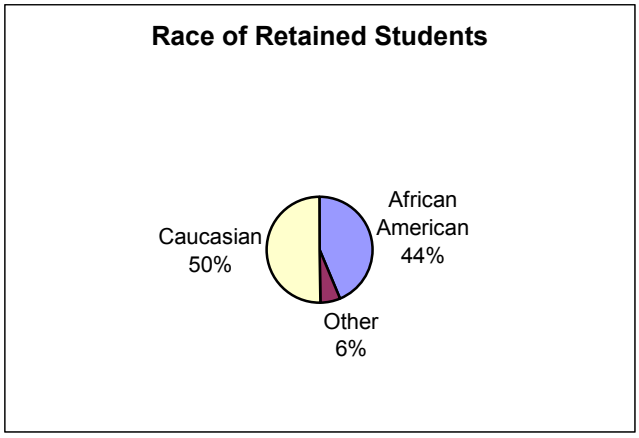

Income Status of Potential Retainees
Not Low
Income
30%
Low
Income
70%

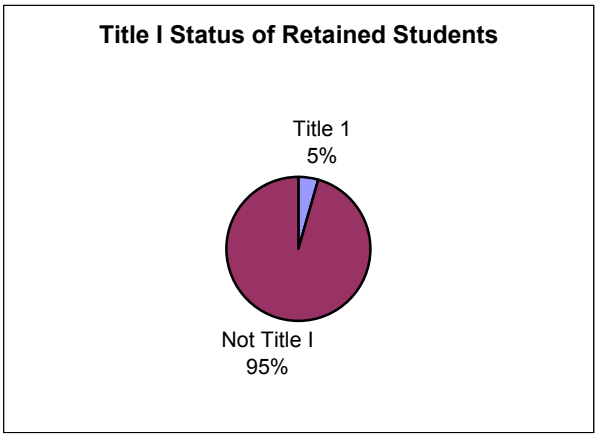
A logistic regression analysis was performed in which income status, gender and minority status were used to predict who would have been retained if the DSTP consequences had been in effect. The analysis revealed that gender, minority status and low income status were all statistically significant predictors of retention. The results revealed that the odds of being retained were 2.05 times greater for minority students compared to non-minorities, 3.2 times greater for low-income students than non-low-income students, and 1.86 times greater for males than females.

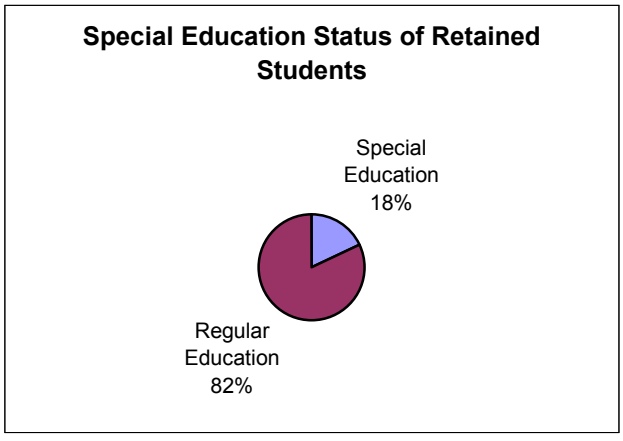
Secondary Cohort: Actual Retainees

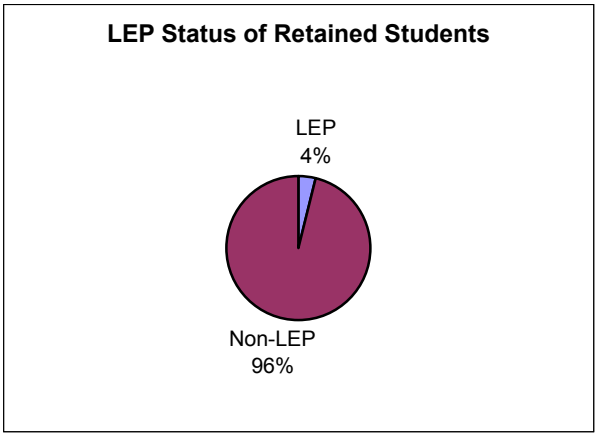
The demographic characteristics of students in the secondary cohort who were enrolled in grade 6 during the 1997-98 ***and*** the 1998-99 school year (n=266) were as follows:

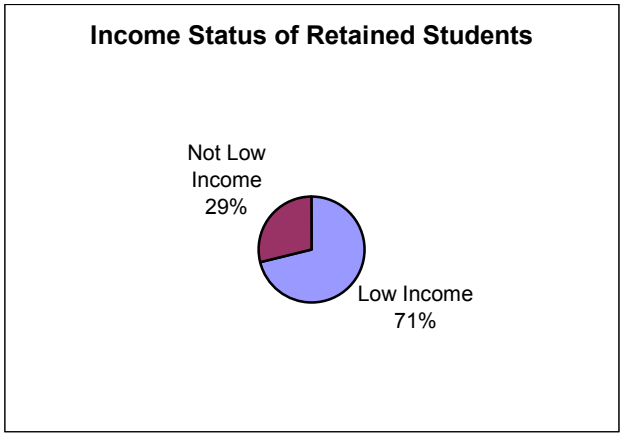


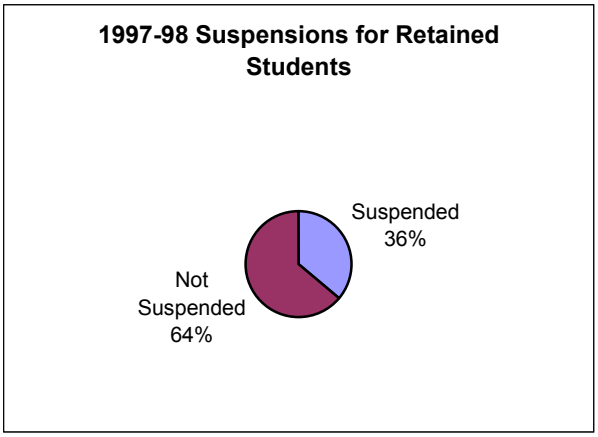


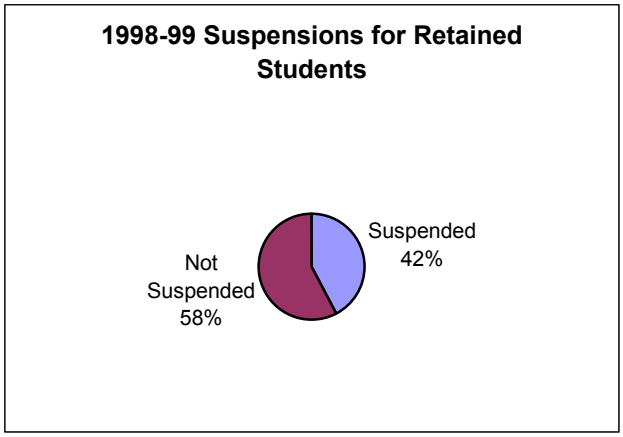








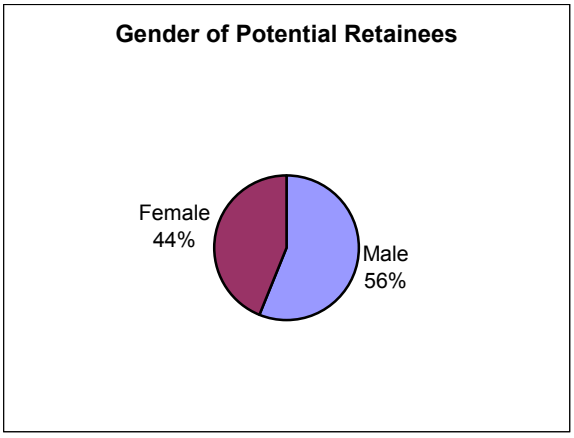


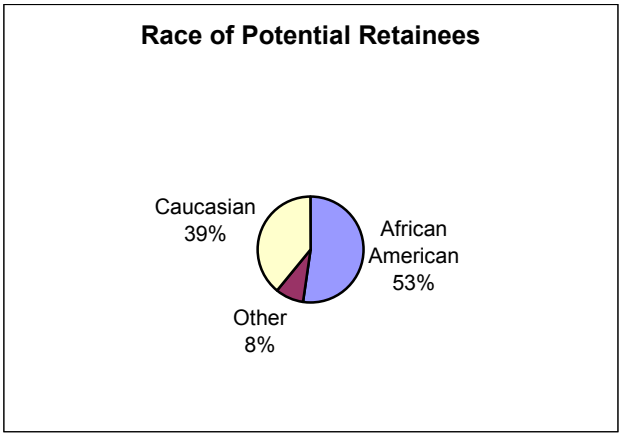


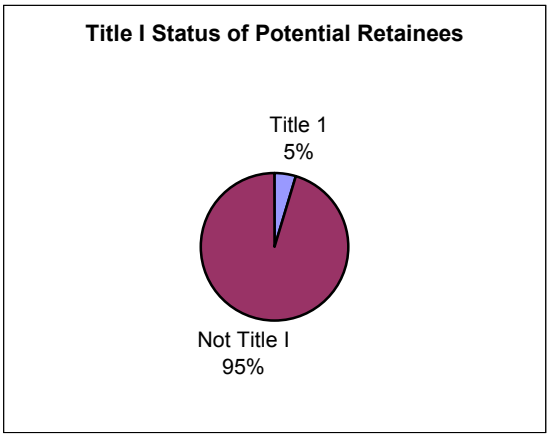
A logistic regression analysis was conducted in which income status, gender and minority status were used to predict retention. This analysis revealed that gender and income status were both statistically significant predictors of retention, but minority status was not. The results revealed that the odds of being retained are 3.66 times greater for low-income students than non-low-income students and the odds are 2.73 times greater for males than females.

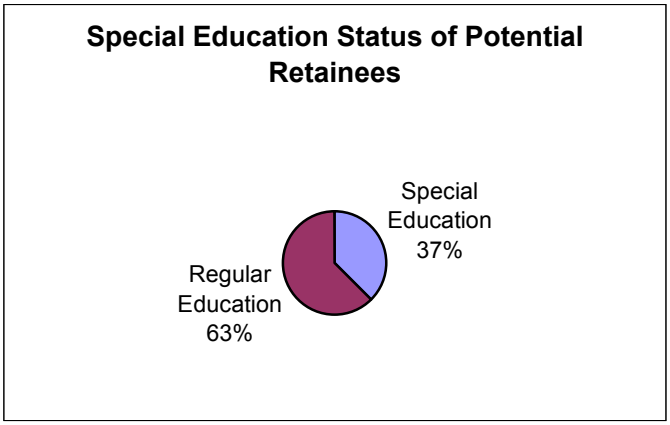
Secondary Cohort: Potential Retainees

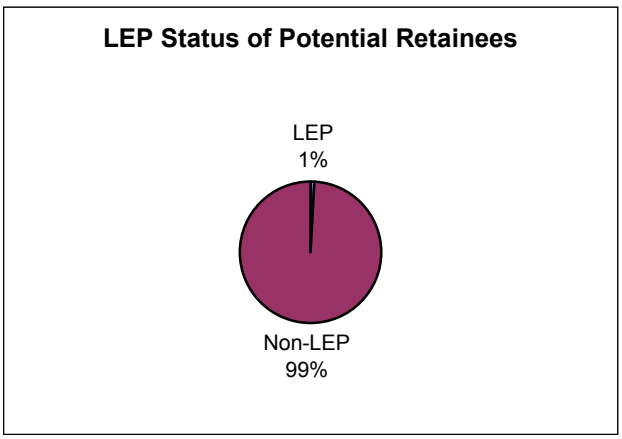
The demographic characteristics of students in the secondary cohort ***who would have been retained had the consequences associated with the DSTP been in effect*** (n=814), were as follows:

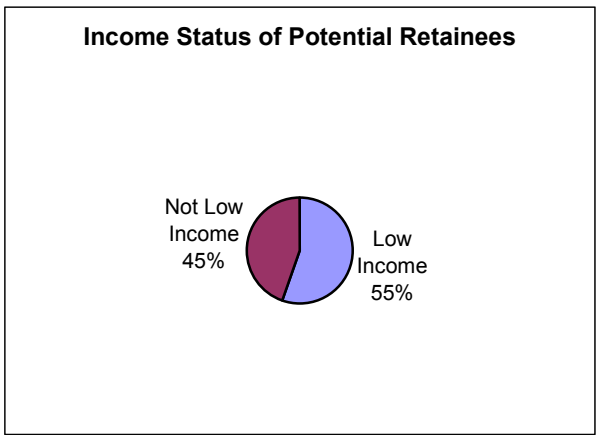












A logistic regression was performed in which income status, gender and minority status were used to predict who would have been retained if the DSTP consequences had been in effect. This analysis revealed that gender, minority status and low income status were all statistically significant predictors of retention. The results revealed that the odds of being retained are 2.61 times greater for minority students compared to non-minorities, the odds are 2.70 times greater

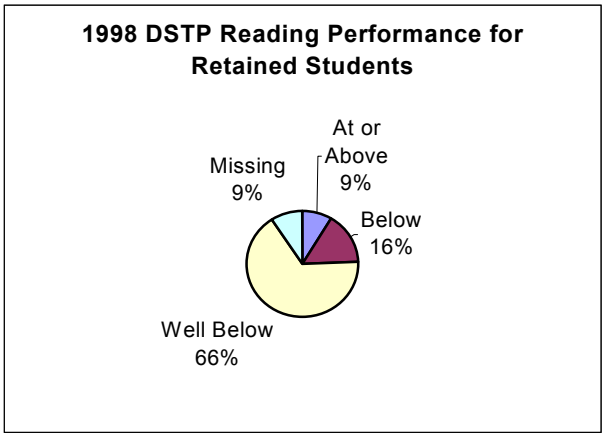
for low-income students than non-low-income students, and the odds are 1.53 times greater for males than females.

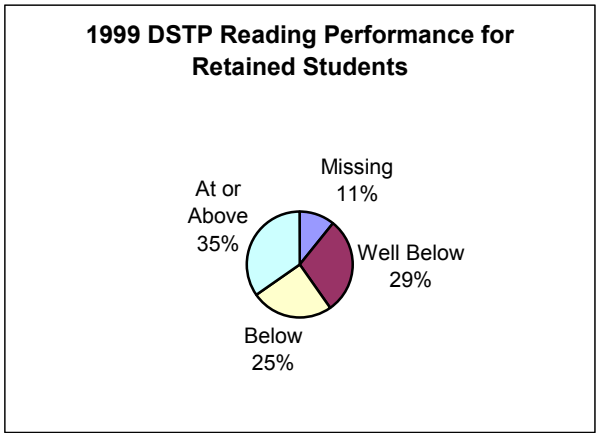
Question 4b: What is the relationship of retention to performance on the DSTP?

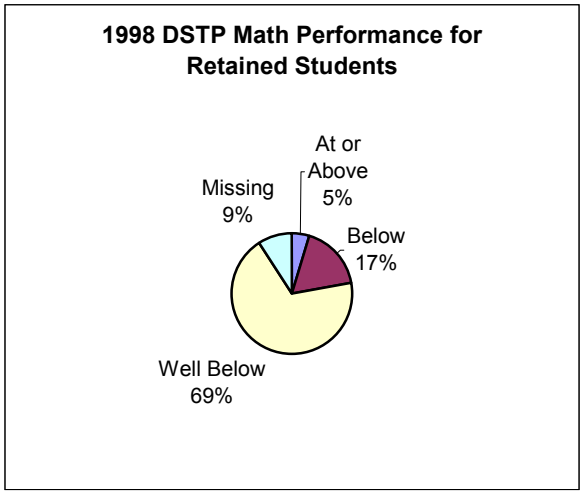
At this point in time, the relationship of retention to performance on the DSTP can only be examined through use of the elementary cohort. The secondary cohort could not be used to address this question because the only year of DSTP data available for the secondary cohort was from the Spring 2000 administration and the current enrollment files (2000-01) were not available.

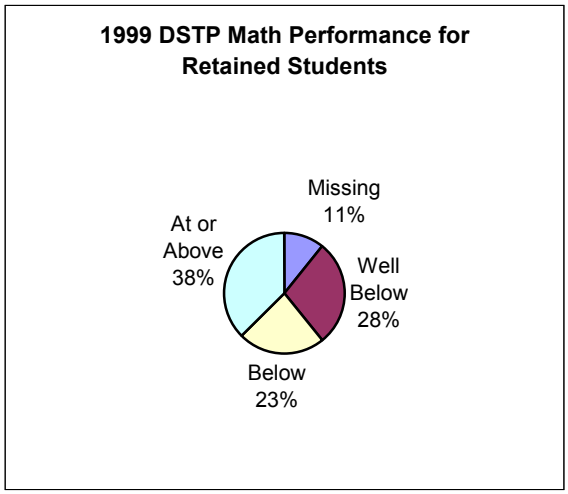
DSTP Performance Levels for Retained Students

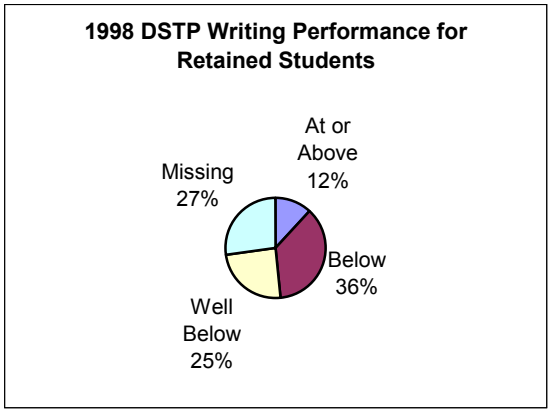
Students who were enrolled in 3rd grade in both the 1997-98 and 1998-99 school year should have completed the 3rd grade version of the DSTP in the spring of 1998 an the spring of 1999. The performance levels of the students retained were as follows:

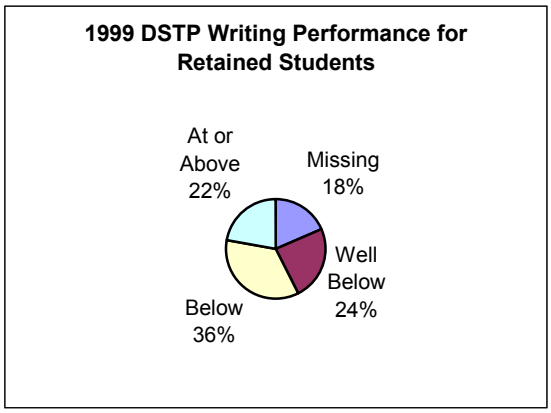








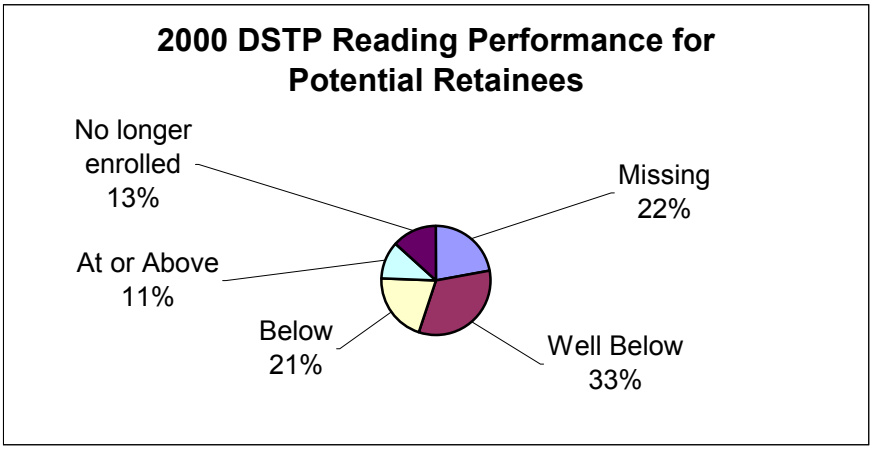


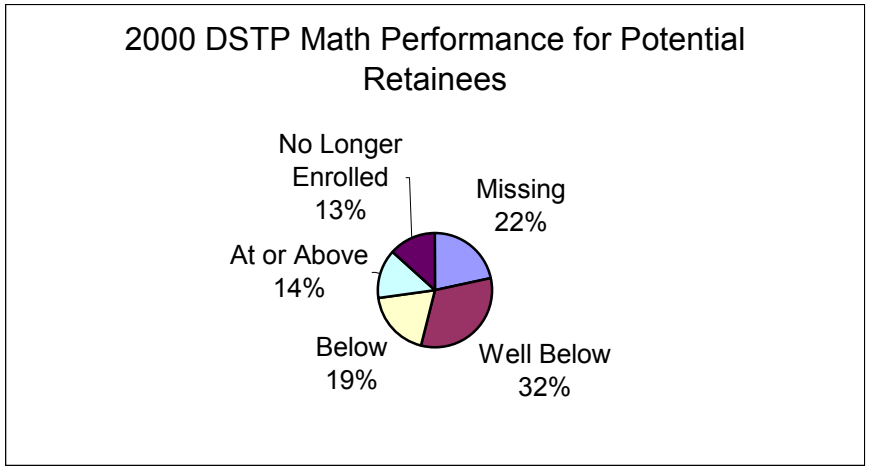


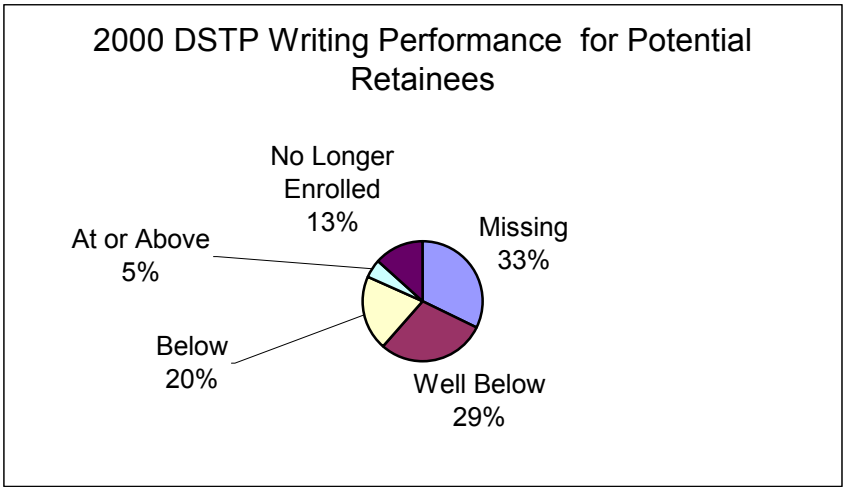
An examination of the data revealed that a little more than a third of the retained students were able to meet or exceed the reading and math standard when taking the DSTP for a second time. Less than one-quarter of the retained students were able to meet the writing standard. For each of the content areas a majority of students who were retained continued to perform below/well below the standard.

DSTP Performance Levels for Potential Retainees

Because the consequences associated with DSTP performance are not yet in place, it is possible to examine what happens to students who performed poorly on the DSTP but were not retained. The following charts display the spring 2000 DSTP performance for students who scored well below the reading standard in 1998.







According to the data, as many as 165 students who were performing well below the standard in reading in 3rd grade were meeting or exceeding the reading standard in 5th grade. These are students who would have been unnecessarily retained in 3rd grade and possibly have suffered some of the negative consequences associated with retention.

Since retention policies typically have a negative impact on student achievement and psychological adjustment, some believe that social promotion is a better solution. However, research clearly indicates that neither retention, as typically practiced, nor social promotion are effective strategies for low-performing students.[18]

The ineffectiveness of social promotion is clearly seen in the number of students who continue to perform well below the standard in the elementary cohort. As many as 497 students who were performing well below the standard in reading as 3rd graders continued to perform well below the standard in reading as 5th graders.

Question 4c: After being retained, how do students fare with respect to behavioral effects and special education status?

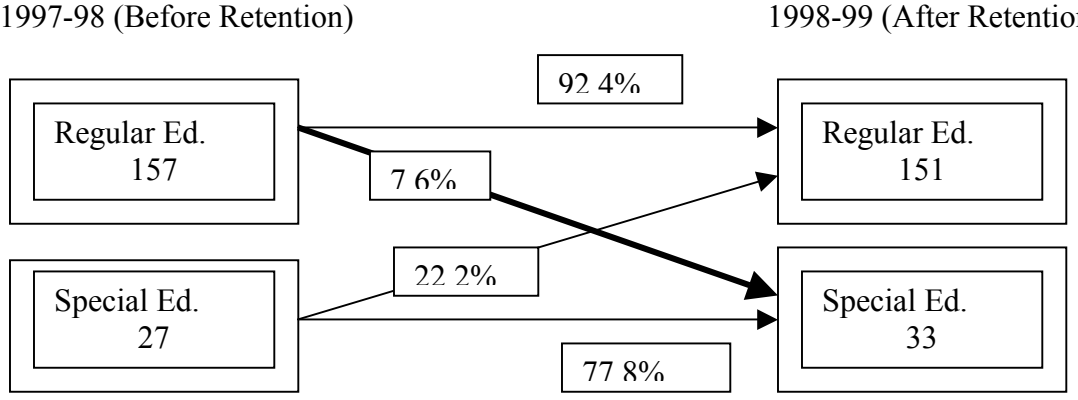
In the section entitled "Issue 3: Behavioral Effects" a model was proposed that linked academic self-concept to student misbehavior. Given the stigma associated with retention, it was expected that being retained might affect students' views of themselves as learners which might in turn increase the likelihood of student misbehavior.

In order to examine this issue, a logistic regression was performed using retention status to predict whether or not a student was suspended. For the elementary cohort, whether a student was enrolled in 3rd grade or 4th grade during the 1998-99 school year was used to predict whether or not the student was suspended during the 1998-99 school year. The results of the logistic regression analysis revealed that retention was a significant predictor of suspension and that the odds of being suspended were 1.95 times greater for students who had been retained than students who had been promoted.

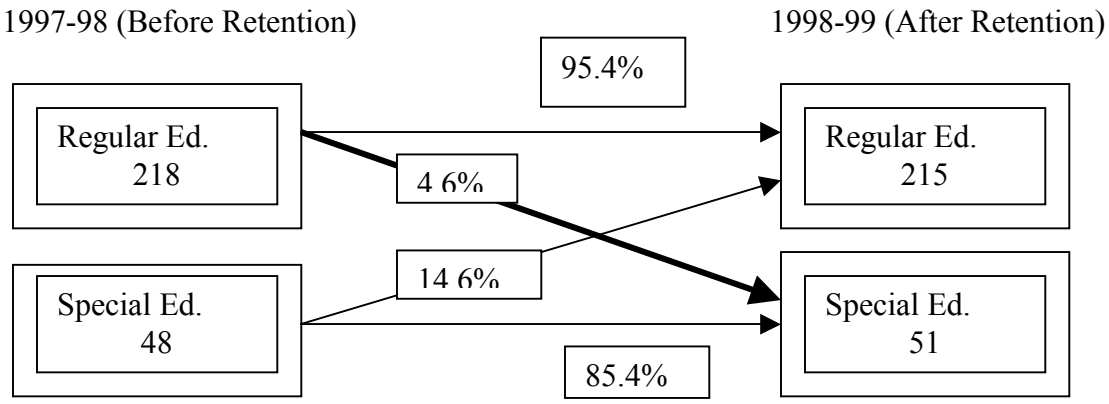
For the secondary cohort, whether a student was enrolled in 6th grade or 7th grade during the 1998-99 school year was used to predict whether or not the student was suspended during the 1998-99 school year. The results of the logistic regression analysis revealed that retention was a significant predictor of suspension and that the odds of being suspended were 3.06 times greater for students who had been retained than students who had been promoted.

In addition to examining the link between retention and student behavior, changes in special education placements following retention were also examined. Although there is currently no research evidence to suggest that special education placement rates increase after retention, previous research has found that retained students are often placed in transitional programs that become little more than dumping grounds for under-performing students.[19]

The special education placements before and after retention for the ***elementary cohort*** were as follows:



The special education placements before and after retention for the ***secondary cohort*** were as follows:



An examination of the special education placements before and after retention reveal a pattern similar to that reported in the section of the report entitled "Issue 1: Special Education Placements". In both cohorts, the percentage of special education students whose classification status changed to regular education was greater than the percentage of regular education students later classified as special education.

Summary and Policy Considerations

Issue 1: Special Education Placement Rates

Many critics of high-stakes assessment systems believe that motivating change through external rewards and sanctions places undue pressure on individuals to manipulate the system. One form of manipulation that has occurred in high-stakes systems is increased special education placements for students. In the current study, an examination of placement changes within the elementary and secondary cohort revealed that across the three years of study (1997-2000) less than 4% of regular education students were later reclassified as special education students. One unexpected finding was that a larger percentage of special education students were later reclassified into regular education programs.

While tracking special education placement rates over time is a good starting place for monitoring the system, a further examination should include an investigation of the testing conditions for special education students. With the re-authorization of the Individuals with Disabilities Education Act (IDEA), all students with disabilities are required to be included in state or district assessments or be given an alternative examination. Many state accountability systems do test students with disabilities but do not include their scores for accountability purposes. In Delaware, special education students may test under one of three conditions: without an accommodation, with an accommodation that allows for aggregation, or with an accommodation that does not allow for aggregation.

Policy Considerations related to Special Education Issues:

- How many special education students currently test under each of the three conditions (without an accommodation, with an accommodation that allows for aggregation, or with an accommodation that does not allow for aggregation)?
- To what extent are present state and district assessment programs inclusive of all students?
- What safeguards are in place to prevent the overuse of accommodations that prohibit aggregation?

Issue 2: Student Achievement

The goal of nearly every accountability system is to improve education and help students learn at higher levels. The extent to which this is occurring in Delaware schools was examined in this study.

For the elementary cohort, nearly every student group evidenced slight improvements in SAT9 mathematics scores from 1998 to 2000 with Asian and LEP students demonstrating moderate improvements. LEP students also evidenced moderate improvements in SAT9 reading scores.

The analyses of student achievement further indicated that test score gains from 1998 to 2000 did not vary as a function of gender, race, or income status. However it should be noted that African American students and female students evidenced a negative change in SAT9 reading scores from 1998 to 2000 and all students evidenced a decline in writing scores over time.

An examination of improvements in the standards-based scores were not possible given the scaling method used for that portion of the test. However, the performance levels associated with the standards based scores were examined. The results indicated that more students were performing at or above the standard in reading and math in 2000 compared to 1998 and fewer students were performing at or above the standard in writing in 2000 compared to 1998.

The fact that reading and writing scores did not improve over time and mathematics scores generally showed only slight improvements does not qualify the reform effort in Delaware as a failure. Many states have found student test scores to increase over time but to decrease when a new testing instrument is implemented. In addition, many educational experts also indicate that reported gains in test scores often represent score inflation.[20] Furthermore, it may be unrealistic to expect large improvements over a short period of time and over a period during which the system is continuing to develop.

However, the pattern that emerged with respect to the writing scores does raise some concern. Given that all groups of students statewide evidenced a decline in writing from 1998 to 2000 seems to point to a measurement problem with the writing portion of the DSTP. The decline in writing scores also suggests that using the writing portion of the DSTP as a proxy for reading performance may be particularly problematic.
Policy Considerations related to Student Achievement:

- What policy or curriculum changes may have contributed to the improvements in mathematics scores?

- What are the ramifications of the vertical scaling of the standards-based portion of the DSTP? How can student improvement on Delaware standards be tracked over time?

- How is the state attempting to investigate and address the decreases in writing scores found statewide?

Issue 3: Behavioral Effects

Out-of-school suspension is considered to be one of the least effective strategies for improving student behavior but it is widely used. In high-stakes settings, removing a student from the school is particularly problematic; it can limit the student's opportunity to learn which can have legal ramifications.

An examination of the demographic characteristics of students receiving out-of-school suspension revealed that the odds of being given an out-of-school suspension were greater for male, low income, and minority students.

Suspended students did not evidence any real performance differences on the DSTP compared to their non-suspended peers. However, the finding that the odds of receiving a suspension were greater for minority students coupled with the finding that African Americans were one of the few student groups who did not evidence a slight improvement in SAT9 math scores from 1998 to 2000 does raise some concern. In addition, although no differences were found when examining DSTP performance, a number of the suspended students did not have a valid score for at least one of the content areas of the DSTP.

Policy considerations related to Student Behavior:

- Are low income, minority, male students receiving adequate opportunities to learn in Delaware schools?

- What attempts are being made to monitor the rate of out-of-school suspensions during the DSTP testing period?

- In high stakes accountability environments, ineffective models of out-of-school suspension become particularly problematic. What other means of dealing with misconduct should be promoted that would have fewer negative effects on student performance?

Issue 4: Retention

Research in the area of retention indicates that retention policies typically have a negative impact on student achievement and psychological adjustment. Also, in keeping with previous research, the results of the current investigation also revealed that retention was more likely to occur to certain student groups. An examination of the demographic characteristics of elementary students who were retained during the 1998-99 school year indicated the odds of being retained were greater for minority students. Within the secondary cohort the odds of being retained were greater for minority and male students.

An interesting finding to emerge in this study was that the extent to which student characteristics were useful predictors of retention depended on how the retention decision was made. When the future consequences associated with the DSTP were used to identify students who would have potentially been retained, the results indicated that all of the student characteristics (gender, minority status, income status) were significant predictors of retention. The results revealed that the odds of being retained were greater for low income, minority, and male students. However, the actual retentions, based on criteria other than the DSTP did not show the same disparate impact on all student groups.

The current study was also able to shed some light on the ineffectiveness of both retention and social promotion in remediating under-performing students. In the elementary cohort, there were a number of students who repeated 3rd grade and were given the DSTP a second time. A majority of these students continued to perform below the standard. For students who would have been retained if the DSTP consequences were in effect but were promoted, a little over a

tenth of them were meeting the standard in 5th grade but a majority were still performing below the standard in 5th grade.

The link between retention and poor social adjustment reported in the literature was also examined. The results indicated that odds of being suspended were greater for students who had been retained than students who had not been retained.

Policy Considerations related to Retention:

- What are the legal implications of potentially retaining a disproportionate number of students who primarily come from minority and/or low income families?

- In the future, students performing well below the standard will be required to attend summer school and re-take the DSTP before the start of the next school year. Given that a majority of students repeating an entire year of school were still performing below the standard when given the DSTP the following spring, how likely is it that a six-week program will be an effective method of remediation? What other strategies could be employed to assist under-performing students?

- In the future, students performing below the standard will be given an Individualized Improvement Plan and promoted to the next grade. How does the state plan to evaluate the effectiveness of the IIPs? How can the state ensure that retention becomes something more than a simple repetition of a year's worth of material?

This report represents an initial attempt to monitor the impact of Delaware's student accountability plan on students. While previous research in the area of high-stakes assessment can point to obvious areas for monitoring, a comprehensive design for monitoring systems is still emerging. A review of the educational literature also suggests other, less obvious areas (i.e. student behavior) that need to be included when monitoring a high-stakes accountability system. It is likely that the process of monitoring the system will evolve along with the accountability system and yield more questions or issues for consideration.

Endnotes

1. Allington, R., & McGill, Franzen, A. (1992). Unintended Effects of Educational Reform in New York. *Educational Policy,* 6, 397-414. Behn, R. (1997). Linking Measurement and Motivation: A Challenge for Accountability. http://www.cse.ucla.edu/CRESST/pages/reports.htm (15 June 1999) Ladd, H. (1996). Catalysts for Learning. *The Brookings Review*, *14*, 14-17.
2. National Research Council. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation.* Washington, DC: National Academy Press.
3. AERA Position Statement Concerning High-Stakes Testing in PreK-12 Education. http://www.aera.net/about/policy/stakes.htm (03 November 2000)
4. Noble, A.J., & Banicky, L.A. (2000). *Synchronizing the Accountability Clocks: A Policy Study of Delaware's Student Accountability Plan.*
5. Banicky, L.A., Noble, A.J., & Siach-Bar, Y. (2000). *Navigating Accountability: Delaware Schools' Response to the State's Student Accountability Plan.*
6. For a summary see: Giancola, S., & Banicky, L. (1998). Discipline. Delaware Education Research and Development Center, College of Human Services, Education, and Public Policy. Education Policy Brief, Volume 3, September 1998.
7. Fuhrman, S. (1999). The New Accountability. (CPRE Policy Brief No. RB-27). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education. Allington, R., & McGill, Franzen, A. (1992). Unintended Effects of Educational Reform in New York. *Educational Policy,* 6, 397-414.
8. 2000 State Summary Report. http://www.doe.state.de.us/aab/DSTP_publications
9. See Aron, A., & Aron, E. (1994). *Statistics for Psychology,* pp. 239-240. Upper Saddle River, NJ: Prentice Hall.
10. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum.
11. See Aron, A., & Aron, E. (1994). *Statistics for Psychology,* pp. 239-240. Upper Saddle River, NJ: Prentice Hall.
12. Report and Recommendations to the Delaware State Board of Education for Establishing Proficiency Levels for the Delaware Student Testing Program in Reading, Writing, and Mathematics. http://www.doe.state.de.us/aab/DSTP_publications.html
13. Branch, C., Purkey, W., & Damico, S. (1977). Academic self concepts of disruptive and nondisruptive middle school students. *The Middle School Journal, 7,* 15-16. Finn, J. (1993). *School Engagements and Students at Risk.* Washington, DC: National Center for Education Statistics. (ERIC Document Reproduction Service No. ED 362 322) Myers, D., Milne, A., Baker, K., & Ginsburg, A. (1987). Student discipline and high school performance. *Sociology of Education, 60*, 18-33. Weishew, N., & Peng, S. (1993). Variables predicting students' problem behaviors. *Journal of Educational Research, 87*, 5-17.
14. Doyle, W. (1989). Classroom Management Techniques. In O. Moles, (Ed.), *Strategies to Reduce Student Misbehavior*. Washington, DC: Office of Educational Research and Improvement. Slee, R. (1986). Integration: The Disruptive Student and Suspension. *The Urban Review, 18,* 87-103.
15. Hosmer, D., & Lemeshow, S. (1989). Applied Logistic Regression. New York: John Wiley and Sons.

16. U.S. Department of Education. (1999). *Taking Responsibility for Ending Social Promotion.* Washington, DC: Author.
17. Dill, S.V. (1993). Closing the Gap: Acceleration vs. Remediation and the Impact of Retention in the Grade on Student Achievement. (ERIC Document Reproduction Services No. 364 938)
18. U.S. Department of Education. (1999). *Taking Responsibility for Ending Social Promotion.* Washington, DC: Author. Shepard, L., & Smith, M. (Eds.). *Flunking Grades: Research and Policies on Retention*. Philadelphia, PA: The Falmer Press.
19. Billman, J. (1988). Two Years of Kindergarten: Ethical and Curricular Considerations. (ERIC Document Reproduction Services No. 297 884)
20. Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12(2),* 8-15, 46-52. Linn, R., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice.* Koretz, D. (2000). Steps Toward Lessened Score Inflation. A report presented at the Annual CRESST Conference, September 14, 2000.

Appendix A

Delaware Student Accountability Monitoring Study Timeline and Data Sources

School Year	Grade of Student		
	Elementary Cohort		Secondary Cohort
1997-98	3		6
1998-99	4		7
1999-00	5		8
2000-01	6		9
2001-02	7		10
2002-03	8		11

Note: Shaded boxes represent on-time test schedules for each cohort.

Data Sources (received annually):

Enrollment files: district code, school code, grade, special education status, LEP status, Title I status, Income status, date of entry into the system, date of exit from the system and student activity codes.

Suspension files: total number of suspensions, total number of days suspended
Suspension data for the 1999-00 school year was not available at the time of reporting.

DSTP files: gender, race, district code, school code, grade, special education status, LEP status, Title I status, Income status, reading standards based score, SAT9 reading score, math standards based score, SAT9 math score, writing score