National Assessment Governing Board
National Assessment of Educational Progress

Using the National Assessment of Educational Progress To Confirm State Test Results

A Report of

The Ad Hoc Committee on Confirming Test Results

March 1, 2002

Ad Hoc Committee on Confirming Test Results
Michael Nettles, Chair
Daniel Domenech
Edward Haertel
Nancy Kopp
Debra Paulson
Diane Ravitch
Michael Ward
Marilyn Whirry
Dennie Palmer Wolf

Planning Work Group
Mark Reckase, Chair
Peter Behuniak
David Francis
Paul Holland
Scott Jenkins
Mary Jean LeTendre
Gerry Shelton
Wendy Yen

Governing Board Staff
Ray Fields

"Using NAEP to confirm state test results should be done as simply as possible without being wrong."

Mark Musick, November 2001

Purpose of this Report

With the recent passage of the "No Child Left Behind Act of 2001" (the Act), state and federal policymakers may look to the National Assessment of Educational Progress (NAEP) to perform a new role. This role involves providing information about student achievement in grades 4 and 8 in reading and mathematics that can be used by the U.S. Department of Education as confirmatory evidence about student achievement on state tests.

As explained in more detail below, neither the requirement to perform this role nor how this role is to be performed is specified in legislation. Congress has afforded broad discretion in developing appropriate, feasible, and defensible procedures for using the National Assessment as confirmatory evidence.

The National Assessment Governing Board recognized the need to study the associated technical and policy issues to ensure that NAEP is ready to do the best job possible if called upon to confirm state results. This report describes the work that was undertaken at the Board's direction and sets forth recommendations to guide the use of National Assessment data as confirmatory evidence.

The immediate audience for this report is the Governing Board, to inform policy decisions to be made about the National Assessment. It is hoped that the recommendations and information in this report also may be of assistance to state and federal policymakers.

Background

Shortly after taking office in January 2001, President Bush issued his blueprint for education, "No Child Left Behind." The blueprint provided for increased accountability on the part of states, districts and schools for improving the achievement of all students, especially those who are economically disadvantaged, and closing the gaps in achievement between high and low performing students.

One element of the accountability plan required annual testing in reading and mathematics for every child in grades 3 through 8 using tests developed by states aligned with content and performance standards set by the states. The President's blueprint provided further that "Progress on state assessments will be confirmed by state results on an annual sampling of 4th and 8th grade students on the National Assessment of Educational Progress (NAEP) in math and reading."[1] Confirming state test results represented a new formal purpose for the National Assessment, with many attendant policy and technical issues to address.

[1] "No Child Left Behind," President George W. Bush, January 2001; http://www.whitehouse.gov/news/reports/no-child-left-behind.html#1.

By June 2001, the House and Senate had passed respective versions of No Child Left Behind legislation. The two bills were different in many respects, but both had provisions that would either require or permit participation by states in annual NAEP testing in reading and mathematics in grades 4 and 8, beginning in 2003, as an element of the system of state level accountability for student achievement results. The NAEP results, along with the state results, would be used by the Secretary of Education in making monetary awards to or in exacting penalties upon states. The Senate bill required state participation in NAEP only; the House bill permitted states to choose NAEP or an alternate test, which would be used in place of NAEP for decisions about awards and penalties to the states.

House and Senate conferees were named in July 2001 and, after protracted negotiations, produced a conference agreement in December that was quickly passed by large majorities in both chambers. The Act was signed into law by President Bush on January 8, 2002.

The Act contains many provisions that affect the National Assessment. Under Title I, in addition to their own tests, states are required to participate once every two years in the National Assessment in reading and mathematics in grades 4 and 8 beginning in school year 2002-2003. School districts that receive Title I funding are required to participate if selected for the NAEP sample in these subjects and grades. The requirement to participate in NAEP biennially replaces the original proposal for annual NAEP assessments in these subjects and grades and the House provision that would have allowed states to select an alternate test. No awards or penalties are attached to NAEP performance in any way, but states that do not meet their "adequate yearly progress" targets for two consecutive years will receive technical assistance from the Department of Education.

The NAEP legislation is amended to conform with the Title I requirements: biennial assessments in reading and mathematics at grades 4 and 8 are made NAEP's first priority; the cost of in-state coordination and test administration is shifted from the states to the federal government; the security of test questions is strengthened; and public access and review of NAEP instruments are codified.

Although the legislation is clear in requiring states and districts to participate in the biennial NAEP in reading and mathematics in grades 4 and 8, the legislation and the conference committee report are silent about the use of NAEP to confirm state test results as proposed in President Bush's January 2001 blueprint.

An indication of congressional intent regarding the use of NAEP to confirm state test results is found in Conference Committee Chairman John Boehner's Fact Sheet on H.R. 1. Chairman Boehner states that the National Assessment will serve "as a single independent 'confirming test'…[in order to] verify the results of the statewide assessments…"[2] The Department of Education fact sheet on the Act, under the heading "Confirming Progress," provides further guidance, making it clear that the "confirmation" or "verification" of state test results using NAEP is intended to be performed under the auspices of the Department.

[2] Fact Sheet on H.R. 1, Chairman John Boehner, January 2002, http://edworkforce.house.gov/issues/107th/education/nclb/accountfact.htm

> "Under H.R. 1 a small sample of students in each state will participate in the fourth- and eighth-grade National Assessment of Educational Progress (NAEP) in reading and math every other year in order to help the U.S. Department of Education verify the results of statewide assessments required under Title I to demonstrate student performance and progress."[3]

Based on the legislation and official statements contained in the respective fact sheets, the following conclusions are drawn about the role of the National Assessment under the No Child Left Behind Act:

- the National Assessment must conduct biennial assessments in reading and mathematics in grades 4 and 8 as its first priority
- participation of states is required in biennial assessments in reading and mathematics in grades 4 and 8 conducted by the National Assessment, beginning in school year 2002-2003
- participation in such assessments is required of school districts that receive Title I funding and are selected for the NAEP sample
- NAEP state results in grades 4 and 8 reading and mathematics may be used to "confirm" state test results under the purview of the U.S. Department of Education
- NAEP state results in grades 4 and 8 reading and mathematics will not be used to make awards to or exact penalties against states

Governing Board Action

In translating the President's blueprint into legislation over the winter and spring of 2001, Congress set 2003 as the base year for NAEP data. Decisions about NAEP had to be made soon to ensure that the National Assessment was ready to perform its new role in time for the 2003 assessments. This involved not only decisions about assessments to be performed beginning in February 2003, but preparatory activities to be carried out in 2002 as well. As a result, the Governing Board conducted a special, additional meeting on June 28, 2001. At this meeting, the Board changed the design of NAEP to make it more flexible and to provide results sooner—within six months after the completion of testing. The Board also augmented the assessment schedule to provide for annual testing in reading and mathematics at grades 4 and 8, but made this action contingent upon passage of final legislation and sufficient appropriations.

In July 2001, the Governing Board established an Ad Hoc Committee on Confirming Test Results to help plan for the use of NAEP in this new role. The charge to the Ad Hoc Committee was "…to prepare a report to the Governing Board, to be presented at the March 2002 meeting, discussing issues related to using NAEP to confirm state test results, identifying

[3] Fact Sheet: The No Child left Behind Act, U.S. Department of Education, January 2002, http://ed.gov/offices/OESE/esea/factsheet.html

strengths and potential gaps in the ability of NAEP to perform this role, and recommending appropriate action to be taken by the Governing Board." The Committee would receive support from the Governing Board staff and from a Planning Work Group (PWG) comprised of technical and policy experts (Appendix A).

The Ad Hoc Committee met initially in August, followed by meetings or teleconferences in November, January, and February. At the August meeting, the Committee approved a project timeline and set out initial guidance for the PWG. Subsequently, the Committee monitored the progress of the PWG and prepared this report.

The Work

At its August 2001 meeting, the Ad Hoc Committee discussed a host of issues that would need to be studied in examining how the National Assessment might be used in confirming state test results. Examples of these issues include: What does "confirm" mean? How should the term "gap" be defined? How much improvement is good enough? What are NAEP's limitations as confirmatory evidence? What measures should be used—standards, means, quartiles, percentiles, etc.?   What are the subgroups of interest and how are they defined, by states and by NAEP? What organization or individuals will be responsible for performing the confirmation of state test results?

None of these (or many other) questions addressed by the Committee had immediate or obvious answers. The Committee decided to proceed by having the Planning Work Group perform "confirmations" using extant data from selected states. The state test data would be presented in the form of "arguments" that states might make in reporting their results to the Department of Education. Extant NAEP data for the selected states would be used as confirmatory evidence in light of the state test data. The Committee expected that through these exploratory studies, the strengths and challenges inherent in using NAEP data in this new role would surface.

The Planning Work Group convened on September 10, 2001 for a daylong meeting. The PWG focused on several key issues and discussed them in depth: the rationale for employing "informed judgment" and "a reasonable person standard" as the basis for reviewing NAEP confirmatory evidence; definitions and options for measuring achievement gains and gaps; the sensitivity of NAEP in detecting change; how comparability between the respective test instruments and procedures affects NAEP's power in providing confirmatory evidence about state test results; and plans for preparing the model state "arguments," including the use of NAEP results.   The meeting concluded with a set of assignments: to identify a sample of states from which to obtain test data for preparing the "arguments," to write short papers on measuring achievement gains and gaps and on determining what size gain should be considered substantive. The assignments were to be completed in time for presentation to the Committee at the November 2001 Board meeting and for data collection from the sample of states to begin in October.

Sidebar: A Breakthrough on Representing "Gains" and "Gaps"

Following the September 10 meeting, Planning Work Group member Paul Holland wrote the paper on measuring gains and gaps (Appendix B). It proposes an important innovation in portraying achievement results—a way to visualize and comprehend at a glance how all tested students performed, how achievement is changing over time, and how subgroups compare. This is done elegantly and simply without losing fine details that may be masked when only average scores or only standards-based reporting is employed. It is being highlighted in this report because it deserves special consideration in using NAEP to confirm state test results.

To set the stage for explaining the innovation being proposed, a brief discourse follows on some of the shortcomings of using average scores and standards for reporting test results.

The deficiency in using average scores is that some important changes in performance may go undetected. For example, if the average score increases from one year to the next, there is an assumption that this gain represents improvement for all students. However, the average score can increase because more able students are doing better while less able students show no change or experience a decline in achievement that is less than the gain for the higher performing students. Another example of the average score masking important changes could happen if the scores for more able students decline while scores for less able students improve. These examples may not always be true, but there is no way of knowing if only average scores are considered. Knowing whether there are discrepant results within parts of the tested population is important information because it can give policy makers and school staff indications about whether programs aimed at particular subpopulations are working or are needed.

Although standards-based reporting is essential for knowing the degree to which students are performing at a sufficient level of proficiency (another factor on which average scores fall short), standards-based reporting alone also may miss some important aspects of student achievement. For example, while the percent of students who meet or exceed a standard may be unchanged from one year to the next, the scores of students below the standard may have increased substantially. This would be important to note, because it permits the progress that is being made to be recognized and acknowledged, even if it is not yet enough for more students to reach the standard.

Paul Holland suggested applying charts used in engineering and medical statistics. These charts permit the display of achievement data at all levels of performance. The idea is to chart, for all scores, the

percent of students below each NAEP score. The curve represents the "achievement distribution" for all the students; the chart will be referred to as the achievement distribution chart. Vertical lines are added at the cut scores that separate the Basic, Proficient, and Advanced achievement levels. The curve crosses the Basic cut score at 70 percent. This means that 70 percent of the tested students score below Basic and thus (100 – 70) 30 percent score at or above Basic.

1 at every point. The whole "achievement distribution" has moved to the right. Moving to the right on the horizontal axis means that scores are increasing across all students: the lowest scoring student in the second year scored higher than the lowest scoring student the first year; the highest scoring student the second year scored higher than the highest scoring student the first year; and every student in between in

the second year scored higher than those in similarly ranked positions in the first year. One can also see in Chart B that the percent at each standard has increased the second year and the percent below Basic has decreased.

Chart C uses a similar approach to display the gap in achievement between two subgroups (8th grade boys and girls) in math in a single year. The curve for boys is to the right of the curve for girls at almost every point in the achievement distribution. The relative position of these two curves illustrates the gap in achievement between boys and girls. The gap is defined as the difference in performance between all boys and all girls across the achievement distribution. The gap is represented by the space between the two curves. If there were no space between the two curves, it would mean that boys and girls performed equally well across all boys as compared to all girls. However, there is a space between the two curves, f

NAEP scale score points. Similar calculations were performed across the achievement distribution, subtracting girls' scores from boys' scores at each percentile. The line illustrates the size of the difference for all students that were tested. You can see at a glance that the difference between boys and girls begins with a slight advantage for girls below the 5 percentile. From the 5 to about the 55 percentiles, the score gap increases from 0 to 8 points in favor of boys, at which it remains about an 8 point difference through the high end of the achievement distribution. Being able to see the comparative results for all students simultaneously is what is remarkable, and exciting, about this approach.

Chart E compares two subgroups over time: students eligible for free lunch and students ineligible for free lunch in 1996 and 2000. Chart E shows that both groups improve achievement from 1996 to 2000 and that there is a gap in achievement, in which students ineligible for free lunch outperform those who are eligible for free lunch. What is not clear in Chart E is whether the size of the gap in achievement is changing.

the information from Chart E, Chart F shows the difference in achievement between eligible and ineligible students in 1996 and 2000. The line for year 2 is below the line for year 1 at almost every point. This means that the gap between the two subgroups is getting smaller and that this is true for almost all students across the achievement distribution.

It is important to know if the gap closing involves all students making improvements with the lower scoring students "catching up." In some cases, apparent gap closings can result from higher achieving students losing ground while lower achieving students make no improvement. In such a case, the size of the score difference has decreased, but in an unsatisfactory manner. Chart E shows that the scores of both subgroups have gone up from 1996 to 2000. In other words, both groups have improved, but the students eligible for free lunch have "gained ground." The gap closing in Chart F, therefore, should be considered "real."

Achievement distribution charts can be powerful tools for examining test results. Once their meaning is understood, they are easy to use and interpret. However, they are not yet in a form that would be easily accessible to the general public. Work needs to be done to improve their design, simplify the labeling, and educate potential users about their import and value.

Chart A
Achievement Distribution Example
NAEP Mathematics Grade 8
With Achievement Levels Indicated

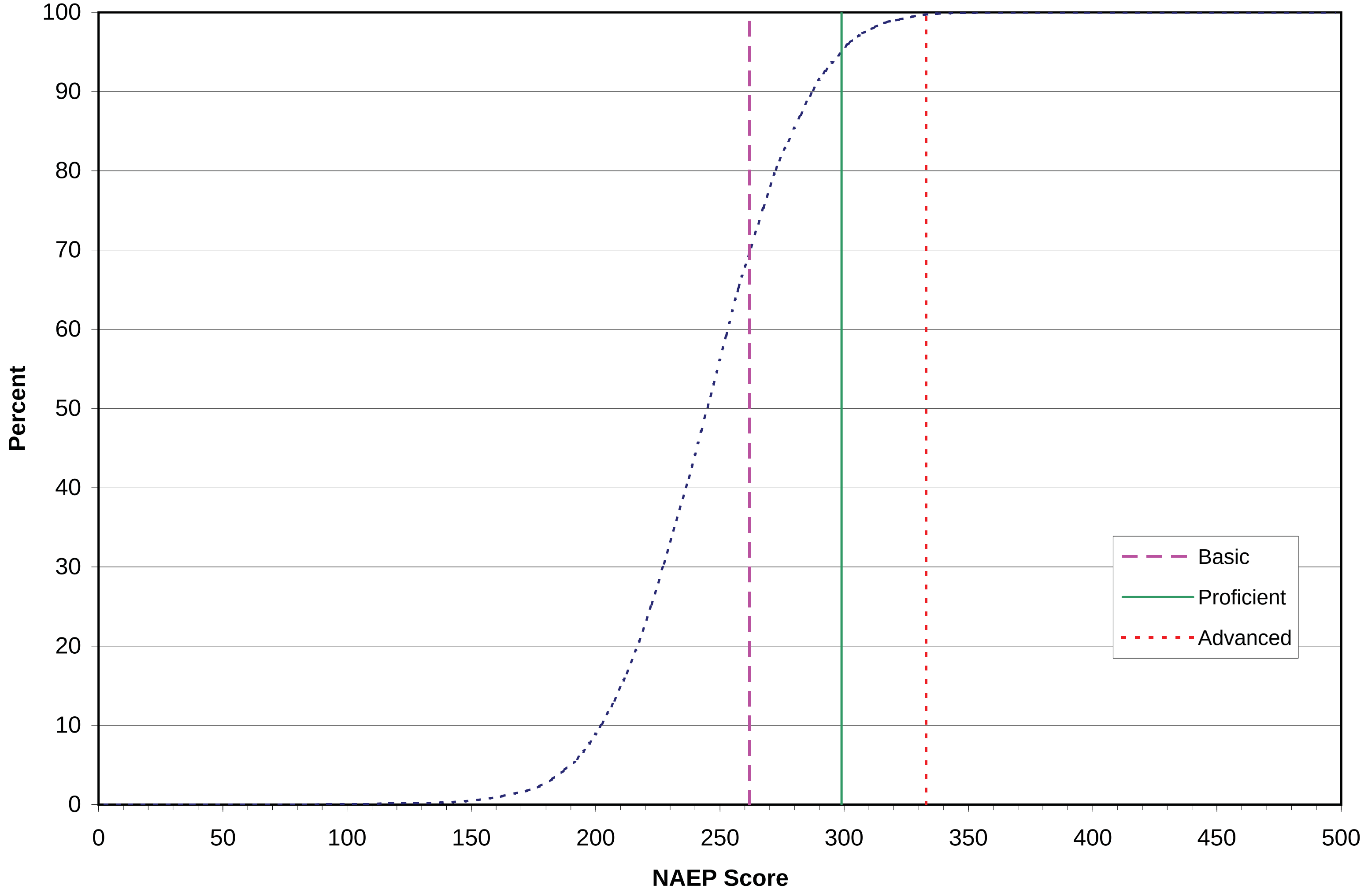


Chart B
Achievement Distribution: Change Over Time Example
NAEP Mathematics Grade 8: Year 1, Year 2
With Achievement Levels Indicated

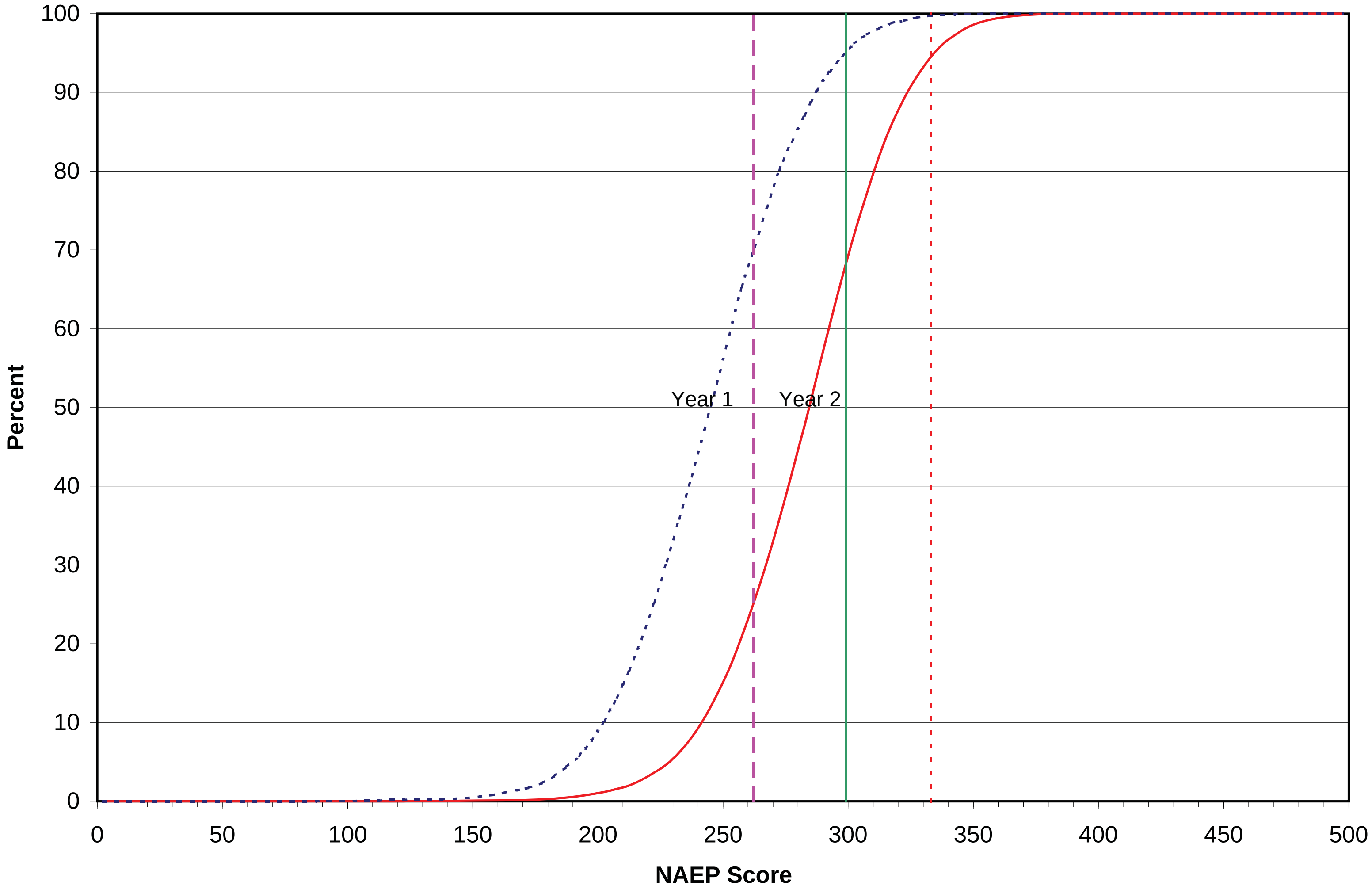


Chart C
Achievement Distribution: Gap Between Subgroups Example - Boys and Girls
NAEP Mathematics Grade 8

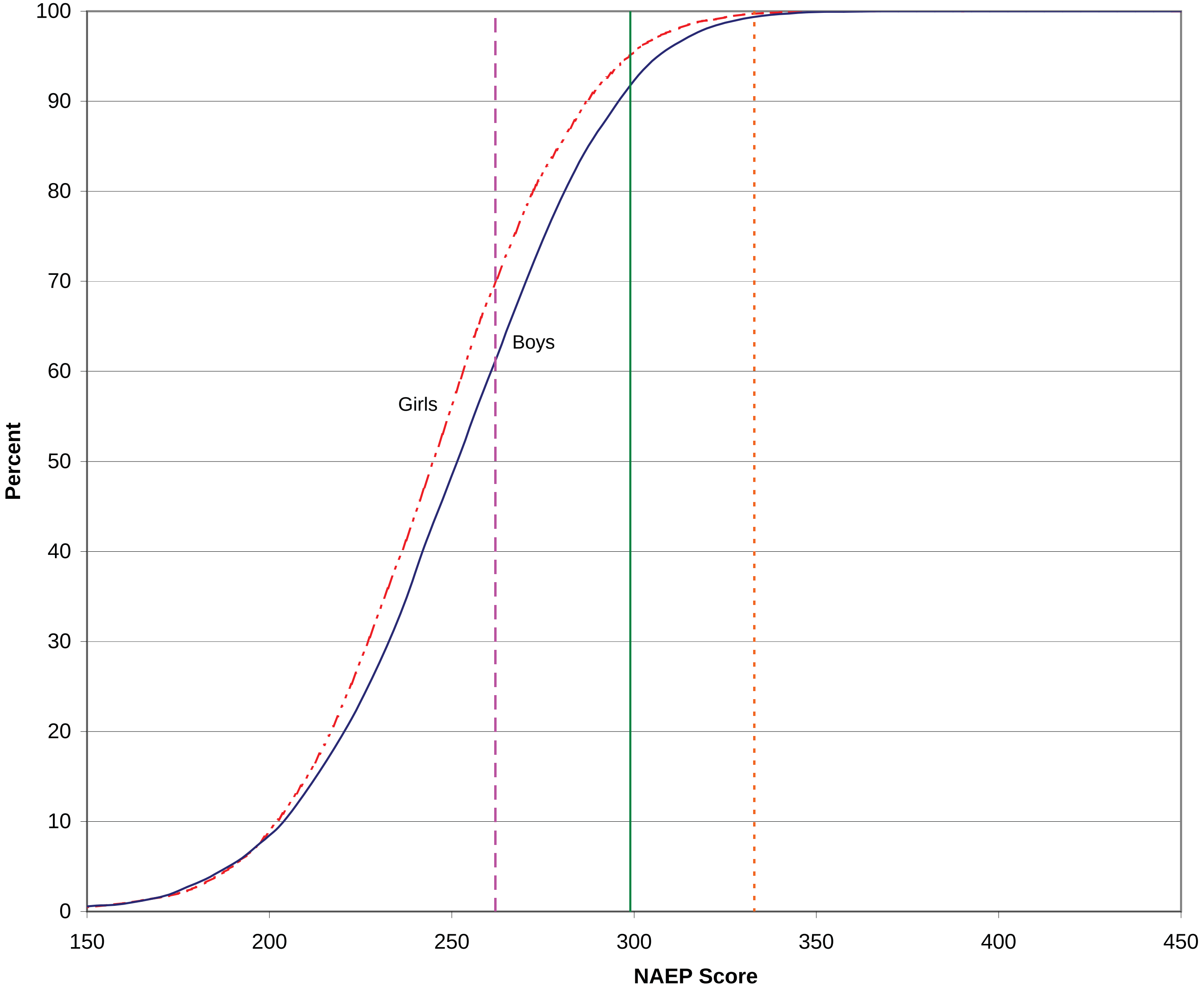


Chart D
Achievement Distribution: Size of Gap Example
Gap Between Boys and Girls in a Single Test Year
Score Differences in Achievement By Percentile
NAEP Mathematics Grade 8

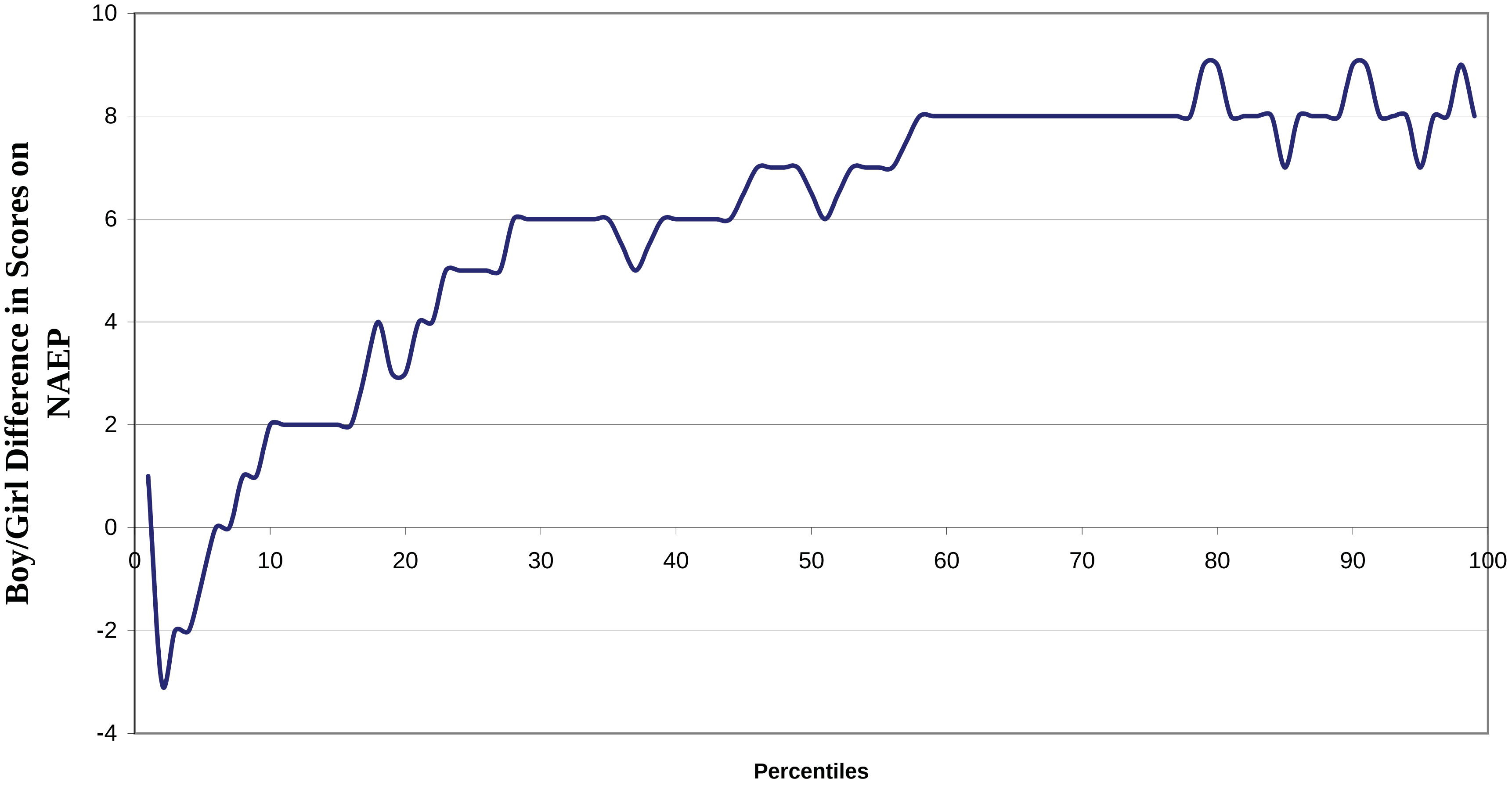


Chart E
Achievement Gap: Comparing Two Subgroups Over Two Years Example
NAEP Mathematics Free Lunch Eligibility: 1996 and 2000

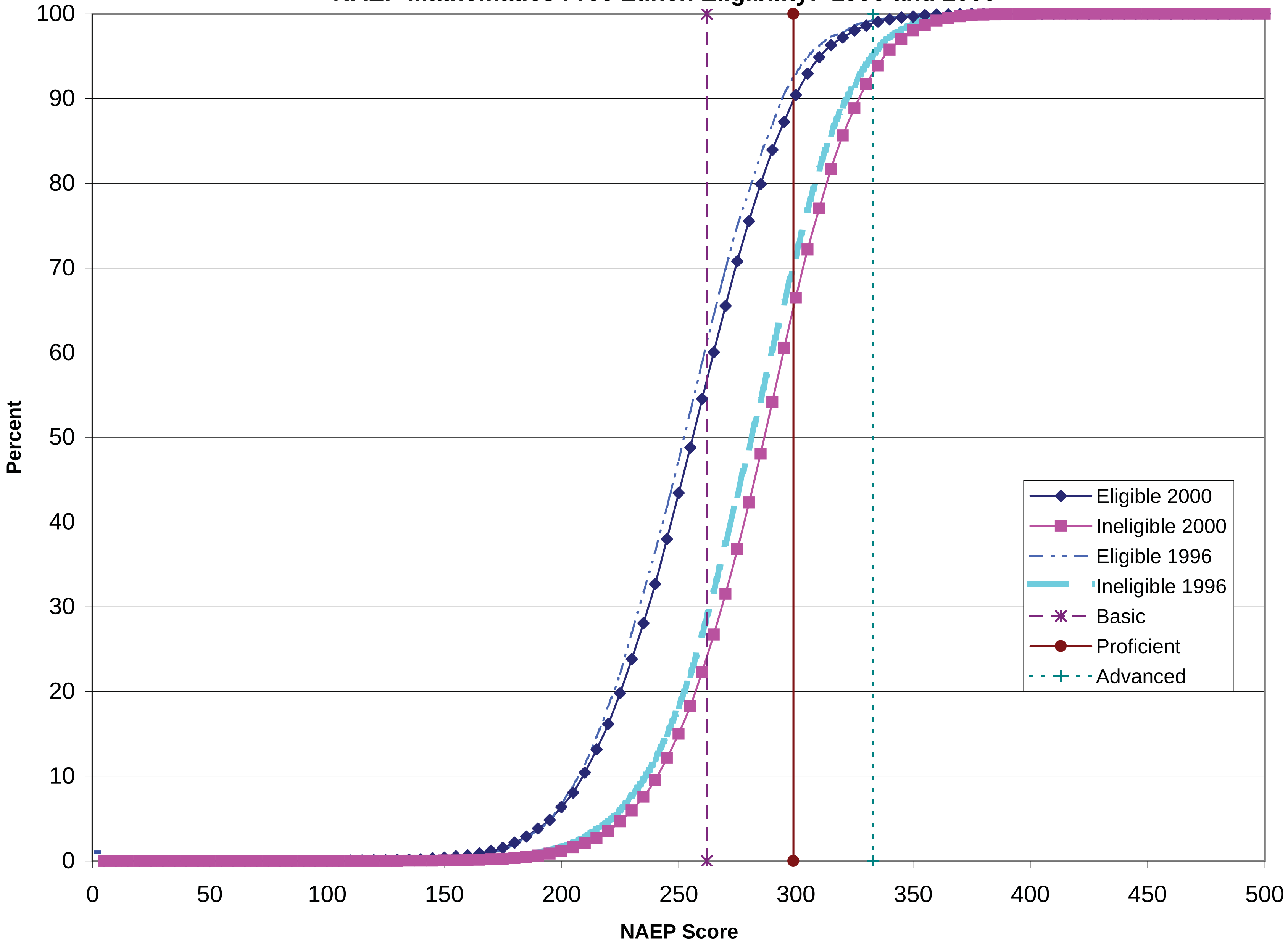
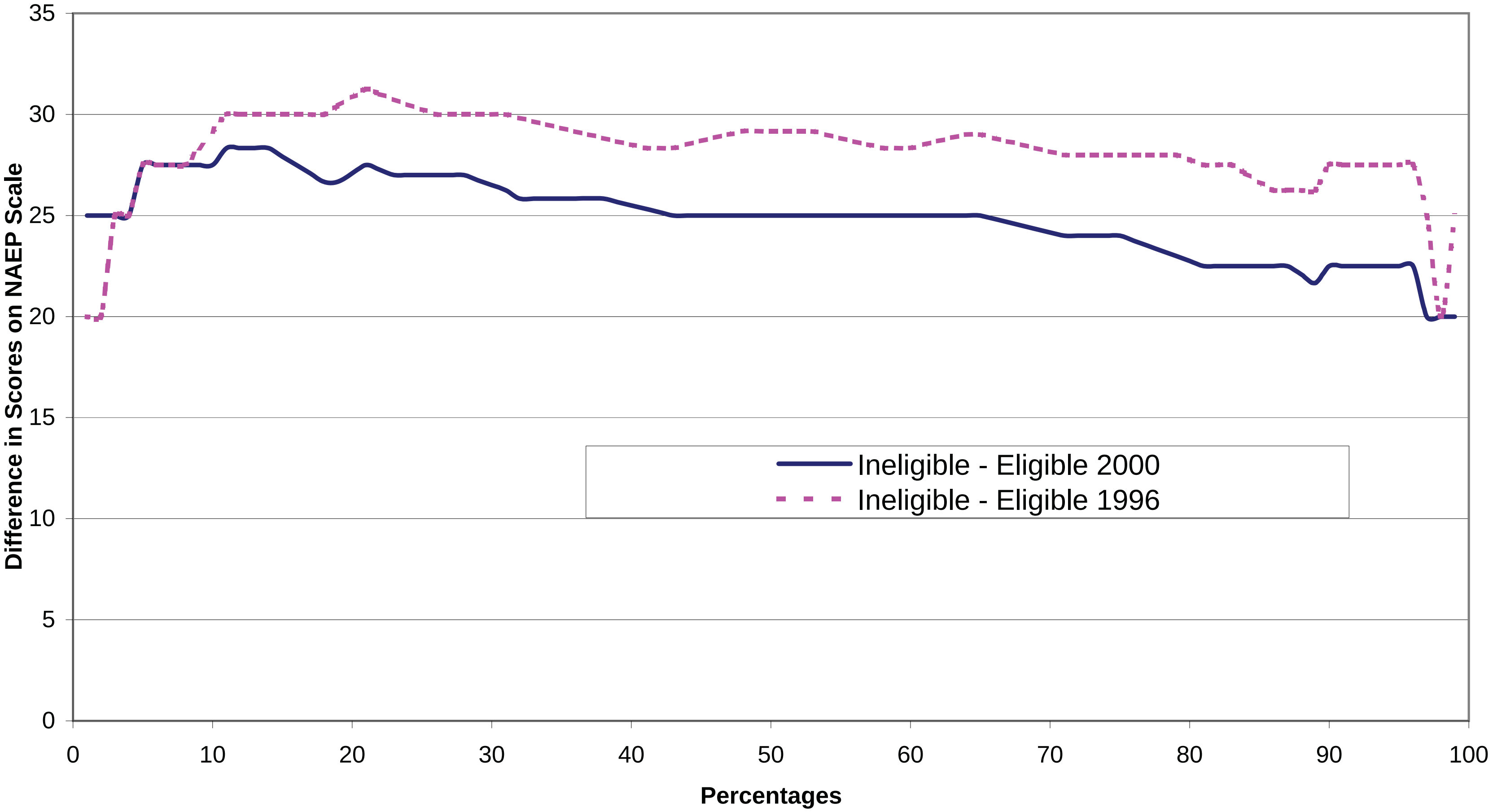


Chart F
Achievement Distribution: Change in Size of Gap Example
Ineligible- Eligible: Free Lunch
NAEP Mathematics Grade 8: 1996 and 2000



Paul Holland's paper on measuring gains and gaps and another he wrote on the size of gains (Appendix B), the status of data collection from the nine state sample, and the plans for preparing the state "arguments" were presented to the Ad Hoc Committee at the November 2001 Board meeting. The Ad Hoc Committee discussed the Holland papers and approved the plans to go forward with developing the state arguments.

On November 26, a three-member subgroup of the PWG—Mark Reckase, Wendy Yen, and Paul Holland, the state argument authors—met with Ad Hoc Committee Chair Michael Nettles, Committee member Nancy Kopp, and NAGB and NCES staff. The purpose of the meeting was to review the state data that were collected and discuss issues and procedures related to the development of the state arguments. It was decided that the eight states for which data were available would be divided among the three authors. Each author would review the respective sets of state data and select one state that held the most promise in illustrating issues and problems relevant to using NAEP as confirmatory evidence. An outcome of this meeting was the distillation of the discussion into a set of preliminary "principles" about using NAEP as confirmatory evidence.

The PWG met on January 14, 2002 to review the three state arguments that had been prepared (Appendix C). On the basis of the day's discussion, the preliminary principles were further revised and recommended to the Ad Hoc Committee for consideration.

The Ad Hoc Committee augmented and refined the recommended principles and organized the principles along categories that match the components of the National Assessment: test development, sampling, data collection, and reporting. The final principles form the substantive core of the project report.

Principles for Using the National Assessment to Confirm State Test Results

1. The National Assessment of Educational Progress (NAEP) can be used as evidence to confirm the general trend of state test results in grades 4 and 8 reading and mathematics.

The No Child Left Behind Act (the Act) is clear in making state curricula, content standards, performance standards, tests and definitions of "adequate yearly progress" the foundation for educational accountability. Beginning in school year 2002-2003, states are required to participate once every two years in the National Assessment in reading and mathematics in grades 4 and 8. School districts that receive Title I funding are required to participate if selected for the NAEP sample. The data that result from this state and district participation in the National Assessment will be available "to help the U.S. Department of Education verify the results of statewide assessments required under Title I to demonstrate student performance and progress." [4]

[4] Fact Sheet on H.R. 1, U.S. Department of Education website; http://ed.gov/offices/OESE/esea/summary.html

Based on a review of test results from eight states and three demonstration "arguments" designed to illustrate how states might report test results to meet the requirements of the Act, the Ad Hoc Committee found state level National Assessment results useful as confirmatory evidence for state progress. Generally, the National Assessment results were in the same direction as state assessment results.

The National Assessment is a good choice as a source of confirmatory evidence for state achievement results. First, the National Assessment occupies a unique position among tests. The National Assessment is a cooperative project with the states. It is conducted at the same time of year across all participating states, following the same standardized procedures and maintaining the highest standards for test security. It is the only dependable source of regularly provided, comparable state achievement data.

The National Assessment also is attractive as a source of confirmatory evidence because its test frameworks—the content and format of each test—reflect states' views on what is important to assess. The test frameworks are developed taking into account state curriculum documents. States are represented on the framework development committees and participate in the review of the frameworks in draft form. In addition, NAEP's sampling and test design permit very broad coverage of a subject in a minimum of student test taking time. The National Assessment sampling procedures require no student to take the complete assessment, which would be the equivalent of four to six hours of testing time for an individual student. Instead, each student is tested for about one hour and is presented a sample of the total set of test questions in the assessment.

There has been consistently high voluntary participation by states over the twelve years in which NAEP has conducted state assessments. This would not be the case if states viewed NAEP as lacking sufficient correspondence with the content of state tests or believed that NAEP results are invalid depictions of state performance. In fact, states view the information provided by the National Assessment as important. About twenty states have adopted legislative or regulatory requirements for district participation in NAEP and many states already have used their NAEP results on an informal basis as a means of confirming progress on their own tests.

2. "Informed judgment" and a "reasonable person" standard should be applied in using National Assessment data as confirmatory evidence for state results. Confirmation should not be conducted on a "point by point" basis or construed as a strict "validation" of the state's test results.

The purpose of confirmation or verification is to provide a point of comparison or second "snapshot" of state results. Through this additional look, changes in state scores can have a common frame of reference that can add to the security of inferences made about the results. But in weighing the confirmatory evidence, a strictly statistical approach should not be used, such as one that might be applied if a test were being validated. A number of factors exist that potentially limit the degree of convergence between NAEP and state test results. These factors, discussed below in the remainder of this report, affect the degree of convergence to differing extents and in differing ways among the states. These factors limit the use of NAEP as a strict point-by-point validation tool for state results. It also is noteworthy that high stakes decisions on the basis of the confirmatory evidence are not intended under the Act. No rewards or penalties are being attached to the NAEP results. Accordingly, the degree of precision expected of NAEP as confirmatory evidence should be considerably relaxed.

The confirmatory evidence should not have to reach such a high level standard as "beyond a reasonable doubt." Instead, states should be given the benefit of the doubt about whether their results are confirmed. Any amount of growth on the National Assessment should be sufficient to "confirm" growth on state tests. The amount of growth on state tests should be the primary focus in considering state progress. Although states are required to set targets for "adequate yearly progress" on state tests, targets for increases on NAEP should **NOT** be set. Obviously discrepant results between the state test and the National Assessment should be explained, or hypotheses about the differences should be provided. The state results may be questioned when there is consistent, compelling contrary evidence from the National Assessment that cannot be explained simply by the differences between the two tests or other relevant factors.

3. Limitations in using NAEP to confirm the general trend of state test results should be acknowledged explicitly.

Potential differences between NAEP and state testing programs include: content coverage in the subjects, definitions of subgroups, changes in the demography within a state over time, sampling procedures, standard-setting approaches, reporting metrics, student motivation in taking the state test versus taking NAEP, mix of item formats, test difficulty, etc. Such differences may be minimal or great in number and in size and cannot reasonably be expected to operate in all states in equal fashion. The greater the differences between the respective state tests and NAEP, the greater the complexity in using NAEP as confirmatory evidence for state test results and the greater the cautions in interpretation that should accompany the weighing of the confirmatory evidence.

4. Test frameworks for the National Assessment should continue to be developed with the active participation of states. Content coverage in a subject should be broad, inclusive of content valued by states as important to assess, and reflect high aspirations for student achievement. The focus should continue to be on measuring what students know and can do, not on advancing a particular instructional approach or pedagogy. In order for NAEP's value as a consistent, stable measure to be optimal, changes in NAEP test frameworks should be made infrequently and only when the goal of stability in measurement is significantly at odds with the goal of appropriate content coverage.

The National Assessment can be relied upon as a solid source of confirmatory evidence because there is sufficient correspondence in content coverage between the National Assessment and state tests. States independently develop their own content standards, curricula, performance standards, and tests. Although it is unrealistic and infeasible for any second test to be perfectly aligned with fifty different state tests (per subject, per grade), a high degree of overlap is possible.   Because of the active participation of states in framework development, the broad content coverage in each subject area, and the very large number of test questions and their range in format, the National Assessment is particularly well situated to optimize the overlap with state tests. Now that confirmation of state test results is an explicit purpose of the National Assessment, attention to state curricular goals, content standards, and performance standards should be given heightened attention in determining what NAEP should measure as frameworks are revised or created anew.

Likewise, as the National Assessment carries out this role, its attention should continue to be on measuring what students know and can do, not on prescribing how they should be taught. The fear that the National Assessment could lead to a national curriculum should be kept in mind and addressed through broad, active, participation in framework development by external groups, experts, interested members of the public and educators.

The Governing Board's policy is to keep test frameworks in a subject stable for at least ten years, at which time it evaluates whether a revision is in order. In general, this policy has served the public and the National Assessment well. Keeping frameworks stable ensures that the National Assessment can report change in achievement effectively. As psychometrician Al Beaton has been quoted: "If you want to measure change, don't change the measure." The corollary related to using NAEP for confirmation would be: "If you want a stable continuing source of confirmatory evidence, don't change the measure."

Over time, however, curricula do change. At some point, curricula change enough so that new test frameworks and tests are needed. The Governing Board should continue its current policy with respect to maintaining the stability of frameworks. However, in those subjects and grades for which the National Assessment would be providing confirmatory evidence about progress in achievement on state tests, the Governing Board should revise frameworks only when the rationale for doing so is compelling.

5. Sampling procedures for the National Assessment should ensure that results for major subgroups within a state are reliable and that the size of standard errors is minimized. To the extent practicable, exclusion rules in states and in NAEP should be similar.

The Act identifies several target groups: economically disadvantaged students, students from major racial and ethnic groups, students with limited English proficiency, and students with disabilities. Although states are required to disaggregate results for other subgroups, it is the aforementioned target groups that are the basis of state level accountability. It follows, then, that the National Assessment should provide information on these target subgroups for use in confirming state results.

In confirming state results, comparability in subgroup definitions and identification procedures between the state data and the National Assessment is very important. Discrepant results between the state and NAEP may be explained, in part, by different subgroup definitions and identification procedures. To help ensure that results are comparable, it would be ideal if students were categorized identically on the state test and on NAEP. However, it is also important to affirm that NAEP's policies and procedures should remain uniform across states.

Change over time in the demography of the state also adds to the complexity in interpreting state test results and NAEP confirmatory evidence. The National Assessment should be used to confirm subgroup results only when it has reasonable data on the subgroups in question. If the NAEP sample does not have sufficient subgroup data, the conclusion should be that subgroup results can neither be confirmed nor disconfirmed.

Comparability of the NAEP state sample and the state tested population also is important. Participation rates in the NAEP sample should be high. To the extent practicable, exclusion rates for limited English proficient students and for students with disabilities in NAEP should be similar to such rates in the state tested population. Where necessary and feasible, over-sampling of subgroups should be expanded to improve the accuracy of estimates of achievement for these subgroups. However, sampling plans will need to be tailored to individual states due to variations in the subgroups present in the respective states and due to variations in the size of subgroups within states. The Act provides, for state tests, that results for targeted subgroups are not required where

> "the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student." [5]

This is likely to be an issue in states with very small populations of certain subgroups. The National Assessment should conduct research to determine the extent to which over-sampling is needed, the impact on test burden, precision of achievement estimates, and the cost.

In using NAEP to confirm the general direction of state test results, exclusion rates for students with disabilities and for limited English proficient students in the respective assessments should be compared and a statement should be made about the degree to which the results should be interpreted with caution because of differences in exclusion rates.

The National Assessment should conduct research to develop test administration procedures that will ensure that students are categorized identically on the respective state tests and on NAEP and that exclusion rates on state tests and on NAEP are similar.

6. The release of NAEP data should include comprehensive means for presenting and analyzing state NAEP results.

The release of NAEP results, particularly in reading and mathematics in grades 4 and 8, should be conceived, in part, as a service to provide useful, relevant NAEP information that states can use for their own purposes and as a service to those who will be using NAEP results for

[5] ESEA Title I, section 1111(b)(2)(C)(v)(II), P.L. 107-110.

confirmation of state test results. This service approach should result in a “customer” orientation for the program. It should involve seeking on-going feedback from states and participants in the confirmation process so that NAEP’s ability to provide relevant information is continually enhanced.

It is likely that states will report their own test results in different ways and use differing measures for reporting progress. Variety rather than sameness is likely to be found among the state reports. The National Assessment should provide the widest possible range of information to ensure the completeness, relevance and usefulness of the confirmatory evidence it provides.

Beginning with the release of results for base year 2003 assessments, the National Assessment should provide at least the following for each state, for the total sample and for appropriate subgroups: means, percent at or above each NAEP achievement level, percentiles, quintiles, achievement distribution charts, and achievement distribution gap charts.

Achievement distribution charts, achievement distribution gap charts, or suitable alternative means for displaying scores for all of the students tested and for all of the students in a subgroup, represents an innovation that should become a standard part of NAEP reporting. Being able to visualize the achievement of all students simultaneously in a simple format makes it easy to see whether achievement is improving for all students or just some, whether gaps are closing for all members of a subgroup or just some, whether gaps are closing within a subgroup, and whether students at particular ability levels within a subgroup are making differential gains in achievement.

Using charts such as these can overcome the weaknesses inherent in summary statistics, such as average scores, which communicate efficiently but may permit important changes in achievement experienced by certain parts of the tested population to remain undetected, a result that is antithetical to the policy of leaving no child behind. Such charts also make clear whether gap closing is real or an artifact of the higher performing group stagnating or declining in achievement. All of these considerations are important factors in examining state test results and in weighing confirmatory evidence.

7. In using NAEP achievement levels as confirmatory evidence, the percent at or above basic, proficient, and advanced, and the percent below basic, should always be presented and considered in light of the full range of state standards. Information about the achievement distribution should be used to augment the standards based interpretations of the results.

The Act requires states to designate their performance standards as “basic,” “proficient,” and “advanced,” the same designations that are used by the National Assessment. States will vary in how they set their standards. In some cases, the state designations may closely match those for NAEP; in other cases they may not. One example might be that the percent of students at or above “basic” on the state test may be equivalent to the percent at or above “proficient” on NAEP. Although it might be useful for individuals familiar with the vagaries of standard-setting to examine changes over time at “basic” on the state test in light of changes at “proficient” on NAEP, it also may be confusing to others, who would wonder how “basic” on one test can be equivalent to “proficient” on another. Individuals likely to be involved in reviewing state achievement data and NAEP confirmatory evidence undoubtedly will have the experience and knowledge to handle such dissonance. However, reports to the general public should be cautious about displaying such direct comparisons and should do so only if accompanied by clear explanations.

Because of the differences between state standards and NAEP achievement levels, and the likelihood that confirmation results will be public information, wider understanding of the nature of the NAEP achievement levels and definitions becomes important. Dissemination of information explaining the achievement levels to states and other audiences should be intensified. Test questions and samples of student responses should be used extensively in illustrating each achievement level.

Although standards based reporting is essential for knowing the extent to which students’ achievement is “good enough,” it is important to augment the analysis to ensure that changes that may have occurred are not missed. For example, it is possible that students scoring below the basic level may have improved scores from year 1 to year 2 of testing, but not enough to reach or pass the basic standard. The percent at or above basic would not change in such a case and “no improvement” might be the erroneous conclusion. Therefore, examination of the percent at or above the NAEP achievement levels should be accompanied by a close analysis of the achievement distribution. This will make changes in achievement apparent wherever they are occurring and ensure that complete information is available.

Under the Act, the target for improvement is based on the goal of all students reaching the proficient level on the state test within 12 years. State targets for “adequate yearly progress” will be based on this goal. The AD Hoc Committee reiterates its recommendation that NAEP results be used to confirm overall improvement and that annual improvement targets on NAEP not be established for states nor used in confirming state results.

Some states may set their standards on subscale results in a subject. Thus, within a subject, two or more standards may be set, depending on the number of subscales employed. States that report results solely by subscales add complexity to using NAEP as confirmatory evidence. However, NAEP subscale results should not be used to confirm state subscale results.

Conclusion

The Ad Hoc Committee, assisted by the Planning Work Group, studied NAEP’s capacity to serve as a source of confirmatory evidence for state test results. Through the examination of state test results in eight states, the preparation of “arguments” about performance in three of those states, and the use of relevant NAEP data, the Ad Hoc Committee concludes that the National Assessment of Educational Progress can serve this role effectively. The Committee has identified factors that may limit this role and made recommendations to address these factors. The Committee also recommends new ways of representing achievement gains and achievement gaps and encourages further work to provide such information in formats accessible to the general public.

The Committee believes that a significant portion of the NAEP budget should be dedicated to continual strengthening of the program, particularly with respect to sampling procedures, to increase the accuracy of subgroup results, and with respect to reporting, to emphasize the service aspect of data releases and to ensure that comprehensive information about NAEP state results is made available.

While the focus of this report is on the National Assessment, the Committee believes that the lessons learned in preparing this report, the papers that were written, the demonstration state “arguments,” and the new approaches in representing achievement gains and gaps all have wider potential relevance to states and to those involved in the confirmation. The Committee recommends that the Governing Board disseminate this report widely and make it available on the Governing Board’s website.

Table of Appendices

Members of the
Planning Work Group

Mark Reckase, Planning Work Group Chair
Michigan State University

Peter Behuniak
Criterion Consulting, LLC
On Sabbatical, Connecticut State Department of Education

David Francis
University of Houston

Paul Holland
Educational Testing Service

Scott Jenkins
Office of the Governor, State of Michigan

Mary Jean Letendre
Former Director, Title I Program

Gerry Shelton
California State Department of Education

Wendy Yen
ETS K-12 Works

Appendix B

Measuring Progress in Student Achievement:
Changes in Scores and Score-Gaps over Time.

Paul W. Holland
11/9/01

1. Introduction.

As educational reforms are implemented across the US, measures of student progress become of even more interest than usual. Measures of progress can take several forms, and here I discuss only two of them. These are (1) changes in scores over time and (2) changes in gaps between the scores of groups of students over time. My goal here is to suggest some displays of data that help convey the size of these changes in ways that are both clear and complete. By "clear" I mean clear enough that an interested lay person can comprehend the basic ideas and data. By "complete" I mean that the information presented does not suppress the important contextual issues that arise when "changes in" and "gaps between" the scores of test takers are concerned. One of the outcomes of my discussion is that different ways of looking at changes in test scores and of gaps over time can give apparently contradictory interpretations of the same data. These contradictions arise easily in real data and are best resolved by comparing two different types of data displays.

From the point of view taken here, comparing scores (i) at two or more *different points* in time or (ii) for two or more *different groups* at the same point in time require exactly the same types of data displays. A gap over time and a gap between groups are both just "gaps" and what works for one will work for the other. So the two topics that I will really address here are how to *display gaps* and then how to *display changes in gaps* over time.

The "important contextual issues" alluded to above (regarding data displays of gaps) arise from an inherent ambiguity in gap-measures stemming from our interest in *groups* of test scores rather than *single* scores. By analogy, the "gap" between two runners at the end of a marathon is easy to understand because we only have two persons to compare. We can measure their "gap" by the difference in their times of finishing the race or by differences in their average speed. But, what should we mean by the "gap" between two *groups* of runners, say between men and women? It will rarely be the case that all the men will finish before all the women, there will be overlap. The fastest runner will probably be a man but the fastest women will beat many of the men. The same is true when we deal with groups of test takers, rather than runners. In testing we never have just two test takers, we have many, and the scores for two groups will usually overlap in just the same way that the finishing times for men and women runners in a marathon will overlap. In such a situation, what does it mean for the scores of two groups of examinees to have a "gap" between them? How do we measure this gap? And how do we know if it is closing or not as time goes on?

Ignoring the details of these issues can lead to bad policy decisions about whether or not progress is being made or that a gap is closing sufficiently fast to impress us. The key idea is that in some way or another we must compare whole groups of scores and this leads to more complex ideas than the simple difference or "gap" between two scores.

I will first introduce some ideas that have been developed over the years to make these questions easy to frame precisely and illustrate them with NAEP data. I then go on to apply them to measuring gaps and their changes over time. I will not discuss issues of statistical accuracy or "statistical significance" in this paper but, instead, will concentrate on descriptive statistical issues. Of course, issues of statistical accuracy are relevant here but they will merely add another layer of complexity to the points I wish to make, which are already complex enough.

2. Displaying Gaps: Comparing CDFs.

As mentioned above, from my point of view, the same tools can be used to display score-progress over time and score-gaps between groups. For this reason I will emphasize "gaps" in this discussion, but, in section 4, I will show how the same data displays can show changes over time.

The problem we face in trying to describe a "gap" between the scores of two groups of examinees is that we are dealing with whole groups of scores. The way that we have learned to compare such groups of scores is by describing their "distributions." There are several ways of doing this, but a fundamental one, that is easily used to describe the most important ways of defining, seeing, and measuring "gaps," is the *cumulative distribution function*, or CDF curve of the scores for a given group of test takers. Every group of test scores can be described by its CDF curve.

For each possible score, the CDF is the percent of people in the group with a test score less than or equal to that value. (For the experts, I have replaced proportions by percents in order to focus on the simple connection between percents and percentiles. This makes no difference in the ideas, it just multiplies the usual proportion scales by 100, and removes the need for using decimal values.) I will denote the CDF curve by F(x), etc. The "x" denotes a possible test score, and F(x) is the percent of test takers with scores less than or equal to x. A CDF curve is a function of x because we let x range over all of the possible score values. So, with CDFs we are not talking about a single score but about all possible scores for a given assessment.

When I need to introduce another group of examinees I will use G(x) for their CDF curve to keep it separate from F(x).

Figure 1 shows a typical CDF curve. It is for the distribution of NAEP Math scores for all White eighth graders in 2000. In Figure 1, the horizontal axis gives the values of the NAEP scaled scores, x, while the vertical axis gives the value of F(x). The curve is the CDF, F(x).

(Figure 1 about here)

Some useful facts about CDFs. First of all: all CDF curves are roughly "S-shaped", starting at 0 on the left and rising up to 100 on the right. This is due to the definition of a CDF as the *percent* of test takers with scores less than or equal to x. If x is small enough, *no one* has a score less than x (i.e., F(x) = 0%), and if x is large enough, *everyone* has a score less than or equal to x (i.e., F(x) = 100%).

Second, CDFs display *percentiles* in a "backward sort of way". A score, x, is the percentile determined by the percent, p, if x satisfies the equation:

$$F(x) = p.$$

As examples, the score that is the 50th percentile, x_{50}, satisfies the relationship

$$F(x_{50}) = 50,$$

while the score that is the 90th percentile, x_{90}, satisfies the relationship $F(x_{90}) = 90$.

In Figure 1, I have illustrated this equation with the 70th percentile, x_{70}. The arrows in that graph demonstrate how we go from a percent to a percentile by reading the graph "backwards". Thus, in Figure 1 we find the percentile by first choosing a percent, p, on the vertical axis and then looking along the horizontal line at p over to where it cuts F, and then looking down below this cut point to the corresponding percentile value, on the horizontal axis.

Third, the usual aspects of distributions, central tendency or mean values and spread or standard deviations, have their influence on CDFs. Changes in mean values change the "location" of the CDF along the score scale. Shifts to the right indicate higher score distributions, shifts to the left indicate lower score distributions. Changes in the tilt or slope of the CDF are the result of changes in measure of spread or standard deviation. If the tilt gets steeper then the spread gets smaller and if the tilt gets less steep then the spread gets larger. In the former case the curve of the CDF is "spread over" *less* of the score scale and in the latter case it is “spread over” *more* of the score scale. Both location and spread are important aspects of score distributions, but changes in location to the right are usually associated with “progress and improvement”, while changes in spread are less clearly associated with progress or improvement. This point will come up again in section 4 when I discuss change.

Finally, in these notes CDFs are shown as smooth curves, even though, in reality, for test scores they ought to increase by little jumps as we move from left to right across the graph because we never really have a test score for every possible fractional value. When the tests being considered have a substantial number of test questions in them, this approximation is usually sufficient for most purposes. Because the NAEP scale has many score points this will not be very important for this paper except in three graphs, Figures 8, 11 and 13. The curves in these graphs are rougher and wiggle more than they need to due to the discreteness of score distributions. In more sophisticated versions of these graphs the wiggles can be smoothed out.

There are other ways to describe the distribution of a set of test scores, but the virtue of using CDFs in discussing gaps is that they make it easy to *see* the gap between two distributions in a graph. The other main method of describing score distributions is the frequency bar graph or “histogram” and these displays *tend to hide* (rather than reveal) gaps. Comparing Figures 2 and 3 illustrates this point very well. Figure 2 shows two bar graphs for the Math NAEP scores for White and Black eighth graders in 2000. Figure 3 shows the same two distributions as CDFs.

(Figures 2 and 3 about here)

Using a frequency bar graph, as in Figure 2, shows that the two score distributions *overlap* to a great degree, but the view of the “gap” presented in Figure 2 is less easy to describe. Surely it has something to do with the location of the two humps of frequencies, but exactly what this means is easier for an expert to comprehend than for anyone else. In contradistinction, the gap between the two groups is quite evident in Figure 3. It is the space between the two curves. The curve to the left, F(x) for the Black eighth graders, indicates it is the *lower scoring* group, while the one to the right, G(x) for White eighth graders, shows it is the *higher scoring* group. The space in between the two curves in Figure 3 is the “gap” between the two distributions. Bar graphs and histograms tend to emphasize *the amount of overlap* between score distributions, while CDFs emphasize *gaps between* the distributions. This is why I prefer CDFs for discussing gaps in scores. However, as we shall see, the story does not end with CDFs.

In Figure 3, there *really is* a gap between F(x) and G(x) for each test score, x. This is because F(x) is greater than G(x), for every x. When F is always above G this means that the curve of F is *to the left* of G and, therefore, the scores for F(x) *tend to be lower* than the scores for G(x). Not all scores are lower, of course, there are some Black eighth graders who score higher than some other White eighth graders (that is the "overlap" we see in Figure 2). However, in Figure 3, the scores for Black eighth graders are generally lower than the scores for White eighth graders, and hence, there is a "gap" between the curves F and G.

Technically, the relationship that allows this clear-cut notion of a gap between two CDFs is that of "stochastic ordering." F is said to be *stochastically less than* G if F(x) is always greater than or equal to G(x) for any choice of score, x. It is too bad that *stochastically less than* means that F is actually *greater in size than* G at each x, but that is just the way the mathematics works out. If we look at the percentiles, then we do get the ordering in the "right" direction. It may be shown that F is stochastically less than G if every percentile of F is less than the corresponding percentile of G. This is a very natural sense of "less than" for distributions, and it is why "stochastic order" is such an important idea.

In more complex situations, there is no clear-cut "higher scoring group" like there is in Figure 3, and the notion of a "gap" is less intuitively clear in such cases. An example of a more complex case occurs when F(x) and G(x) *cross each other*. Figure 4 illustrates two CDFs that cross and is for eighth grade Boys and Girls on Math in the NAEP 2000 assessment. These two CDFs are actually very close to each other and I have just shown a small portion of them in Figure 4 to illustrate crossing CDFs. When CDF-crossing occurs it is not clear what the gap is. Most importantly, as in Figure 4, one group can not be said to have higher scores than the other, it depends what range of scores we are talking about. These more complex cases can arise in real situations, so we may have to face the fact that the notion of simple "gap" might not be a careful way to describe the difference between two score distributions in some cases. Fortunately, for the type of distributions that arise in NAEP data, crossing CDFs are not all that common, but they can occur.

(Figure 4 about here)

So, if there is a neat gap between the distributions of test scores for two groups of examinees, as in Figure 3, the CDF will clearly display it, and the question then turns to how to measure it?

3. <u>Measuring Gaps.</u>

There are two basic ways to measure the gaps in Figures 3 and 4. We can either measure the *vertical* distance between the two CDFs or we can measure the *horizontal* distance between them. In this section I will discuss both ways of measuring gaps and try to show the strengths and weaknesses of each.

Vertical Differences and "Percent Above Cut": This is a common-sense way to measure gaps and it is what we do when we compare percents above specific Achievement Level cuts (AL's) in NAEP. In Figure 5, I have added to Figure 3 a vertical line at each AL cut-point for the eighth-grade NAEP Math scale.

(Figure 5 about here)

For a specific cut point, x, the vertical distance between the curve, F(x), and 100 percent is the percent of the examinees who score *above* the cut, x. If we compare this with the distance between G(x) and 100% we are measuring the gap between F and G using the "Percent Above Cut x". The CDF curve that is *furthest* from 100% at the score x is the one with the *most* test takers scoring above x. So we see that computing the gap by the *vertical* distance between F and G at the point x bases the gap measure on the difference between the percents above the cut point, x, in each group.

Figure 6 is a special display that is designed to *emphasize the gap* as measured by the vertical distance between the two CDFs in Figure 5. The smooth curve in Figure 6 is the "gap" or *vertical difference* between F and G from Figure 3. It is plotted against the score scale indicated on the horizontal axis form Figure 5. In addition, it includes vertical lines at the three AL cuts for Basic, Proficient and Advanced.

(Figure 6 about here)

As you move from left to right in Figure 6, the gap between F and G starts out at 0, then it increases to a maximum value and then it decreases to 0 again on the right side of the graph. This "from zero-to-a-max-back-to-zero again" pattern always occurs when we measure the gap by comparing percents above cuts and there is a well-defined gap between the two CDFs (as there is in Figure 5). Figure 6 shows that measuring a gap by the *vertical distance* between two CDFs has the potential problem that this gap-measure is sensitive to which choice of cut-points that we use. This is quite dramatic in Figure 6. The gaps at the three AL cuts are all very different. The gap at the Advanced cut-point is about 5% but the one at the Basic cut is about 45% and is as large as such a gap-measure can get in this example. It is important to be aware of the sensitivity of "gaps in percents above cut" to the place where the cut is made. This has interesting consequences for measuring *changes* in gaps that we will see later on.

The value of a graph like Figure 6 is that it focuses our attention on the gap alone and does not confuse the issue with other factors. The CDFs are less satisfactory in this regard because they show much more information, most of which has nothing to do with the gap between the two distributions. There is no standard name for a graph like Figure 6 so I will call it the “gap in percents display” for comparing two score-distributions.

Horizontal Differences and Percentiles: Instead of the *vertical* distance between F and G as a “gap”-measure, we could also consider the *horizontal* distance between them. If we do this, we must choose a specific percent, say, p, and find the two percentiles, x_{pF} and x_{pG}, such that:

$$p = F(x_{pF}) = G(x_{pG}).$$

The distance between the percentiles, x_{pF} and x_{pG}, is the “gap” between F and G measured by the *horizontal* distance between them. Of course, it uses the value of p to determine where we make the gap-measurement (just as we used the score value, x, in computing vertical-distance gap-measures in Figure 6. In Figure 7, I have added horizontal lines for p = 20% and p = 70% to the CDFs in Figure 3 to illustrate this way of measuring the “gap” between F and G.

(Figure 7 about here)

In Figure 7 the gap difference between the 20th percentiles is about 39 points on the NAEP scale, and the same is true for the difference between the 70th percentiles, in this example.

Figure 8 is also a special display, like Figure 6, in that it *focuses solely on the gap* between the two CDFs rather than on the CDFs themselves. In Figure 8, the curve is the *differences in percentiles* for percent p for the data in Figure 3 plotted against each possible percent, p, on the horizontal axis. This curve shows that the horizontal gap varies between 35 and 40 points on the NAEP scale with the slightly smaller differences holding at the percentiles above the p = 80. Thus, while the horizontal gap-measure varies somewhat with the choice of percent, there is much less variability in the gap-measure using horizontal distance than there is using vertical distance as in Figure 6. We would interpret this by saying that the score distributions for Black eighth graders in Mathematics in 2000 are shifted down by about 39 points relative to White eighth graders all along the score scale with slightly smaller shifts among the highest scoring students. This simple “shift” description is almost impossible to see in Figure 6.

Figure 8, like Figure 6, shows the gap-measure over the whole range of possible choices of the percent, p. Again, there is no standard name for a graph like Figure 8 so I will call it a “gap in percentiles display”.

(Figure 8 about here)

When there is a clear-cut gap between the two CDFs, with no substantial crossing of the curves, there can be quite different measurement of the gap depending on whether one uses *percents* (Figure 6) or *percentiles* (Figure 8). The major differences between gap-measures based on percents and percentiles is that percentile differences are often rather similar across different values of p whereas differences in percents above cut must vary from zero to a maximum value and back to zero as we move along the score scale. It is true that at the extremes of 0 or 100 percent the curve in Figure 8 must return to 0, but these extreme percentiles are usually ignored when looking at real score distributions. In short, measuring gaps in terms of differences in percentiles over the range of 1 to 99 percent is often quite insensitive to where the measurement is made on the percent scale when there is a clear gap between the distributions as shown in Figure 5.

Now I don’t want to be accused of pushing this point too far. Gaps measured by percentiles can vary with p, of course, (and they do slightly in Figure 8) but generally they *do not vary as much* as those measured by percents above cut points. I would even go so far as to say, that when they do vary substantially, as they do in Figure 11, they signal easily understood differences between the shapes of the two score-distributions.

(Figures 9, 10, and 11 about here)

To examine Figure 11, let us first consider Figure 9 which displays the CDFs for Black and Hispanic eighth graders on Math in 2000. Except for a small region below scores of 200, there is a clear gap between the two CDFs, with Hispanics being the higher scoring group for most of the score range. Figure 10 shows the corresponding “gap in percents display” for the CDFs in Figure 9. Figure 10 shows the general pattern in Figure 6, but it is more jagged and also indicates the small crossing of F and G for the lower scores by the dip below 0 between 150 and 200 NAEP points. Using Figure 10, it is hard to see how the two CDFs differ in any other important respect due to the over-arching dominance of the “zero-to-a-max-back-to-zero” pattern that “gaps in percents display” always show when the CDFs display a gap, as Figure 9 does. Figure 11 is the corresponding “gap in percentiles display” and it shows that the gap between the two CDFs steadily increases as the score level increases. At about the 60th percentile, the horizontal distance between the two CDFs stops growing and the difference between the upper percentiles is about 8 points on the NAEP scale. The gap is largest for the higher scoring examinees in these groups. This observation is almost imperceptible in Figure 10 where the gap is displayed in terms of percents above various cut points.

Abruptly changing the subject, I need to mention one small but often overlooked point here. Another method of measuring the gap between two distributions is to compare their means or average values. A little mathematics shows that this is exactly the same as forming the *average gap* as measured by percentiles (where we average with equal weight over all possible values of p). Thus, there is a direct connection between gaps as measured by differences in mean scores and the average gap measured by differences in percentiles, i.e, Figures 8 and 11. These two approaches give the same result—they will only give different results when we average over a *small number of percentile differences*.

I think it is the sort of reasoning given in this section that predisposes most psychometricians and statisticians to prefer percentile-based gap measures to gap measures based on percents above one or more cut points. Put another way, those who prefer to measure gaps based on percents above cut-points need to realize the sensitivity of their measures to the cuts they choose. A display such as Figure 6 shows the full picture across all possible cuts, and de-emphasizes the pattern at a few cut-points. However, as we will see in section 4, there are still other ways in which gap-measures based on percentile differences are less sensitive to choices of where to make the gap measurement than are those based on differences in percents above cut-points.

Other gap measures? I have only discussed measures based on slicing the graphs of the CDFs either vertically or horizontally, and I did this because these are the types of simple gap measures that people use when trying to compare two score distributions. There are two other possibilities that I considered.

The *area* between the two curves, F and G, turns out to be exactly the same as the difference between their means, or equivalently, the average of their differences in percentiles. So, that approach did not lead me a new measure, even though it looked like it might.

Another possibility that has been mentioned is to ignore identified groups entirely, and to measure the *dispersion* or spread within a given score-distribution as a measure of inequality or overall “gap” between examinees. Such considerations often arise in economics where income distributions are compared over time. While there are several indices that economists have devised to measure changes in *income* inequality, I did a little investigation of the well-known “Gini inequality index” but I felt that its application to test-score distributions was somewhat forced and unnatural. For this reason I will not pursue it further here.

4. Measuring Change.

Up to now, I have only indicated how we might display and measure gaps between two score-distributions. As indicated in the introduction, there are really three cases of comparisons that ought to be differentiated—gaps, changes, and changes in gaps. I reserve the term “gap” to refer to a comparison of the score-distributions of two different examinee-groups in the same assessment year. I use “changes” to refer to comparisons of the score-distributions for the *same* group in two *different* years. Finally, by “changes in gaps” I mean a comparison of the *gap between two groups* in two different years.

Both “gaps” and “changes” are simple comparisons between two distributions and the same methods, i.e., those discussed in sections 2 and 3, can be used to do either. For example, Figure 12 shows the change in NAEP Math scores between the 1996 and 2000 assessments for all eighth graders. It is analogous to Figure 5, and it shows that there is a clear gap between the two CDFs for the two assessment years, indicating an increase in test scores between the two years. In my opinion, gaps and changes are both usefully displayed by the use of CDFs or, when more precision is needed, the gap or change can be shown in more detail by the “gap displays” illustrated in Figure 6 (for percents), and in Figure 8 (for percentiles).

(Figure 12 about here)

“Changes in gaps” may be measured by combining the tools that we have already discussed so far. My preference is for exploiting either the “gap in percents” or the “gap in percentile” displays because these focus on the gaps alone and let us see how the gaps have changed rather than all the other changes that can accompany them. I will illustrate this using the change in the gap between students eligible for free or reduced lunch and all other students over the years 1996 and 2000 on the NAEP eighth-grade math assessment.

Figure 13 shows the “gap in percentiles” displays between students eligible for free or reduced lunch and all others students for the two assessment years, 1996 and 2000. The two curves correspond to the two years. We see that the percentile-based gap has become smaller over time (by 3 to 4 NAEP points) all across the percentile scale (except for the very lowest and highest percentiles). Thus, Figure 13 is a graph showing *clear progress* in closing a gap. The gap will become completely closed when the curve has slipped down to zero instead of the 22 to 27 point range that it is in 2000.

(Figure 13 about here)

Figure 14 shows the same change in gaps as Figure 13 does, but as measured by percents instead of by percentiles. Again the two curves correspond to the two years. We see that the percent-based gap has changed by *increasing* for the higher scores (above 260 points) and remaining about the same for the lower scores. At each of the three NAEP Achievement Level cuts the gap has clearly *increased* over the two assessments. This hardly seems like progress! What happened? Why do percents and percentiles give us a different story? Is one wrong and the other right?

(Figure 14 about here)

Before addressing this question, I point out that in NAEP it is not unusual for gaps between groups (when measured as differences in percents at or above achievement levels) to appear to change in difficult-to-understand ways that are not reproduced by measures of change based on means or percentiles. This is a general phenomenon, and not a special occurrence.

Figure 15 shows the four CDFs that lie behind Figures 13 and 14, and includes the AL cut points as vertical lines. The four CDFs show that, for the most part, the distributions of scores for the two groups of examinees (i.e., those eligible and those ineligible for Free and Reduced Lunch programs) have all shifted slightly to the right (towards higher scores) over the two assessment years. The gap between the two CDFs for the two years is quite similar except for this shift. Figure 13, giving a more detailed focus on the gap in percentiles, shows that in fact the gap between the two groups has gotten a bit smaller over the two years. It is too small a change to be easily seen in Figure 15, but that is why I suggest using a graph like Figure 13 as a kind of “microscope” to sharpen our view of the gap. We also see that the vertical distance between the two CDFs in the two years is very sensitive to what score level we choose to look at. This the same story as before except that this time it is *changes in differences* between the percents above cut-points that interest us. In Figure 15, it is almost impossible to see how these have changed, but when we look at Figure 14 we see an apparent paradox, the small *progress* seen in Figure 13 has become *regress* in Figure 14.

(Figure 15 about here)

Neither Figure 13 or 14 is *wrong*, but gap-closing progress that is easy to understand is the type that is revealed by scores generally getting higher, that is by the CDFs shifting to the right over time. Differences in percentiles focus on such shifts. It is true that the differences in the percents above each of the three AL cuts have *all increased* over the two assessment years. However, this trend has more to do with the location of the ALs along the score scale than with the small amount of positive movement to the right that these two score distributions have made over the two assessment years. For showing improvement in score-distributions, it is *changes in their location along the score scale* that is easy to understand. Did scores go up or down? That is a question of location. Percentiles focus on such questions directly while percents above cut points do so only indirectly and can give quite unintuitive results, as Figure 14 illustrates.

This discussion gives another reason why statisticians and psychometricians often prefer to compare distributions in terms of their percentiles rather than their percents above cut points. They might put it as follows. Percentile differences are robust to small changes in the score distributions, while differences in percents above cuts are less robust to such choices.

5. Final Comments.

First of all, I think that using graphs, like those in Figures 6 and 8, to *display* gaps and to *supplement* numerical gap-measures is an important move regardless of whether one uses percentile-based or “percent above cut”-based gap-measures. These displays show the full range of possibilities and give a context for interpreting a single number used to report a gap in scores.

Secondly, however, I think I have to come down on the side of measures based on percentiles when it comes to measuring *changes* in gaps over time. The example of section 4 was not cooked up in order to illustrate an arcane statistical problem. Rather it was the first one I looked at and I was surprised at the apparent contradiction that it immediately flushed out. Using changes in percents above cuts as a way of studying how fast a gap is closing or opening can be a very delicate matter and one that needs to be supplemented with full distributional data (as in Figure 15) before the change it represents can be understood. At the very least, a graphical display like Figure 15 or Figure 13 can help resolve the confusing message that changes in percents above cut points will sometimes relay when compared to changes in mean or percentile differences between groups.

Figure 1
Example of Cumulative Distribution Function
NAEP Mathematics 2000 Grade 8
Whites

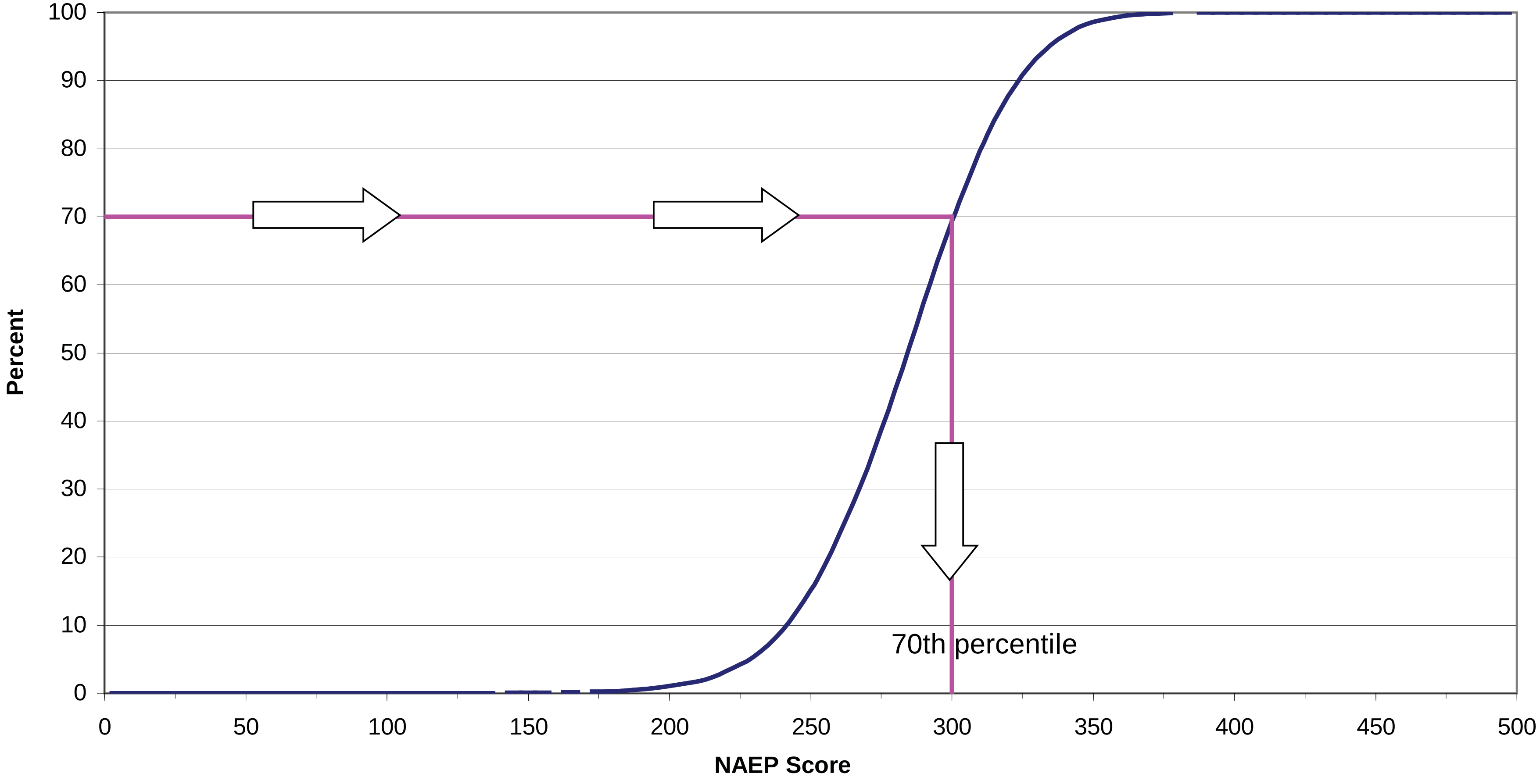


Figure 2
Frequency Bar Graph
NAEP Mathematics 2000 Grade 8

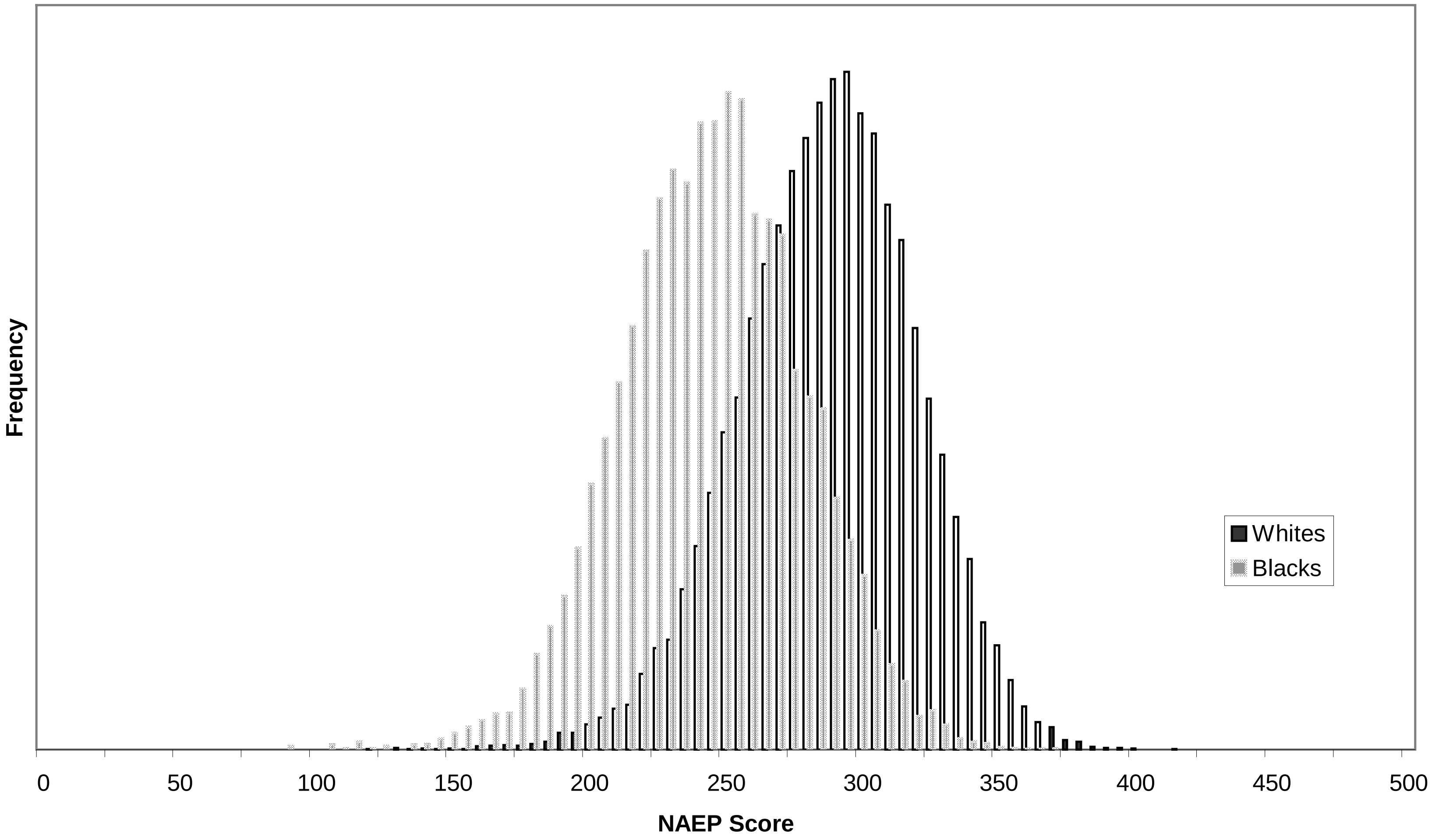


Figure 3
Black/White Gap
NAEP Mathematics 2000 Grade 8

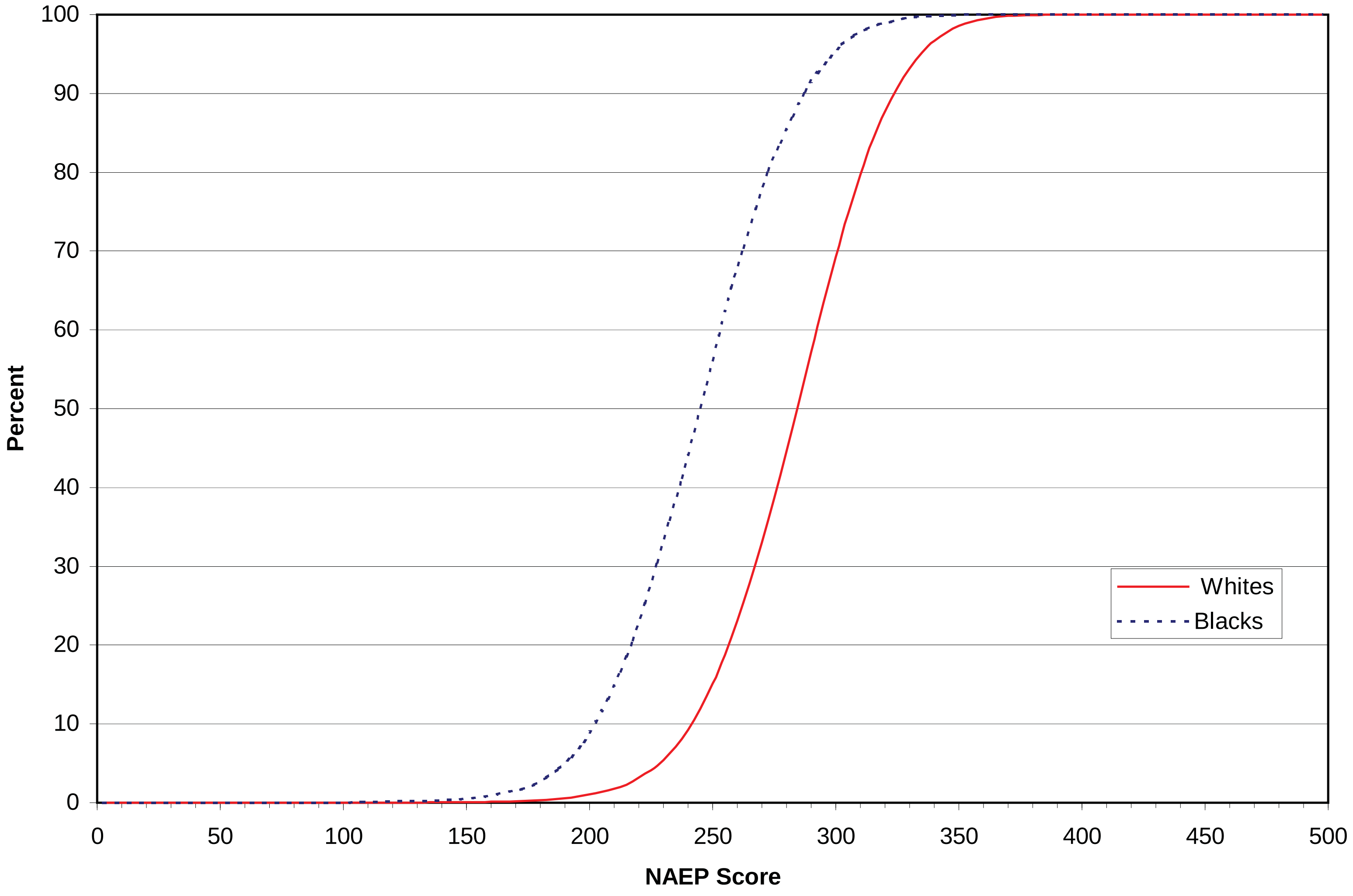


Figure 4
Complex case - Cumulative distributions
NAEP Mathematics 2000 Grade 8

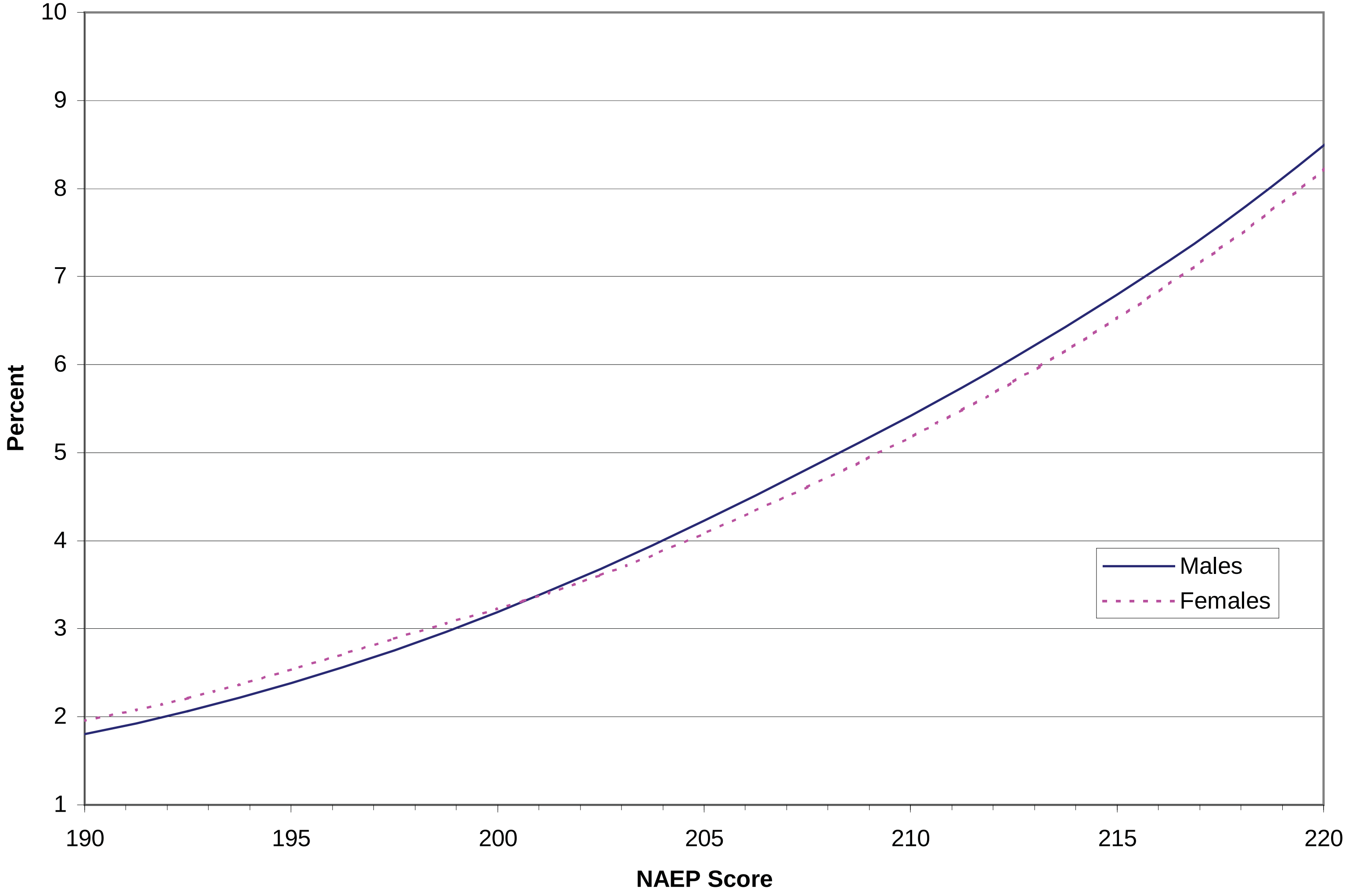


Figure 5
Distributions with Achievement Levels Marked
NAEP Mathematics 2000 Grade 8

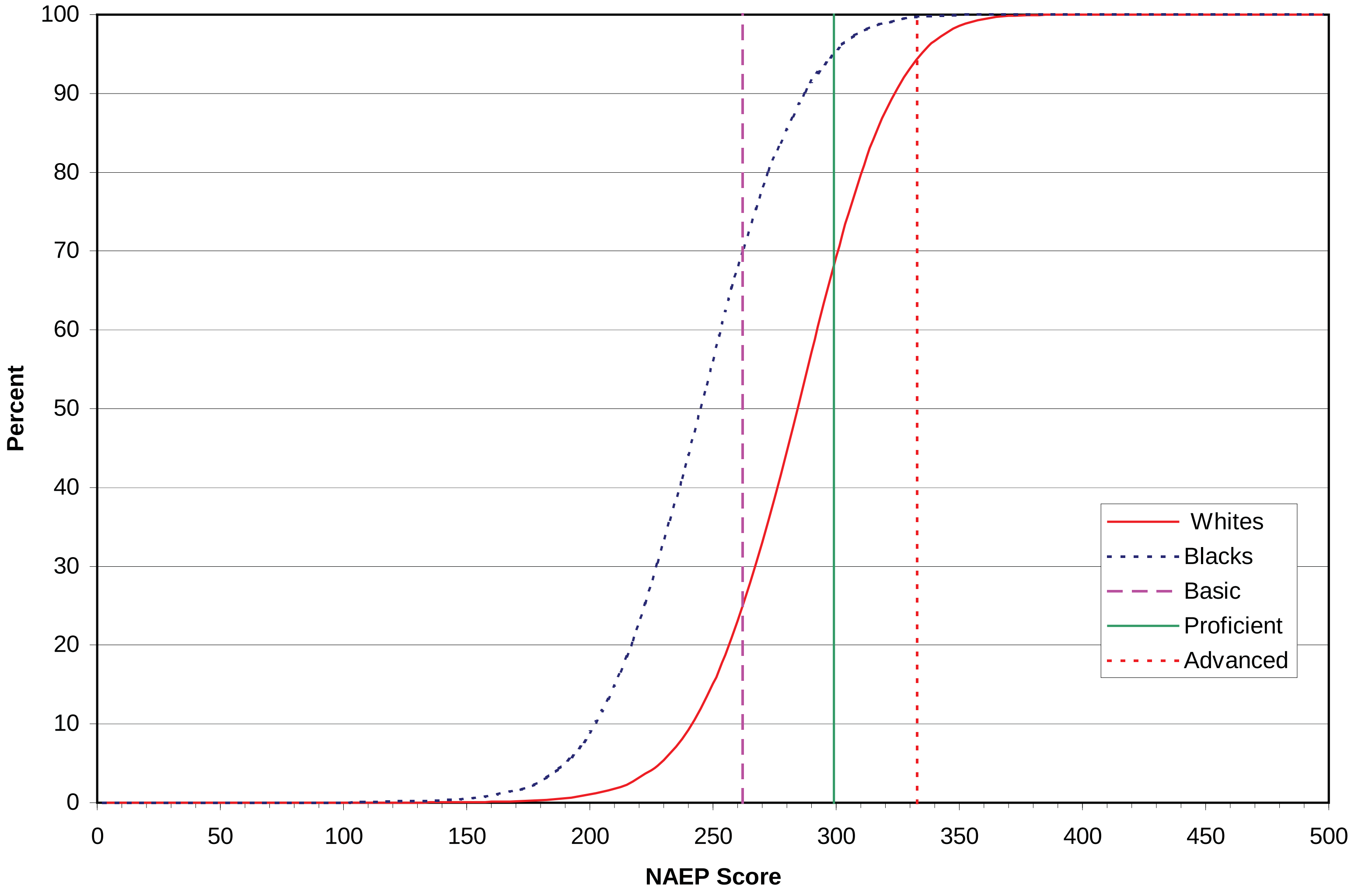


Figure 6
Gap in Percents between Blacks and Whites
NAEP Mathematics 2000 Grade 8

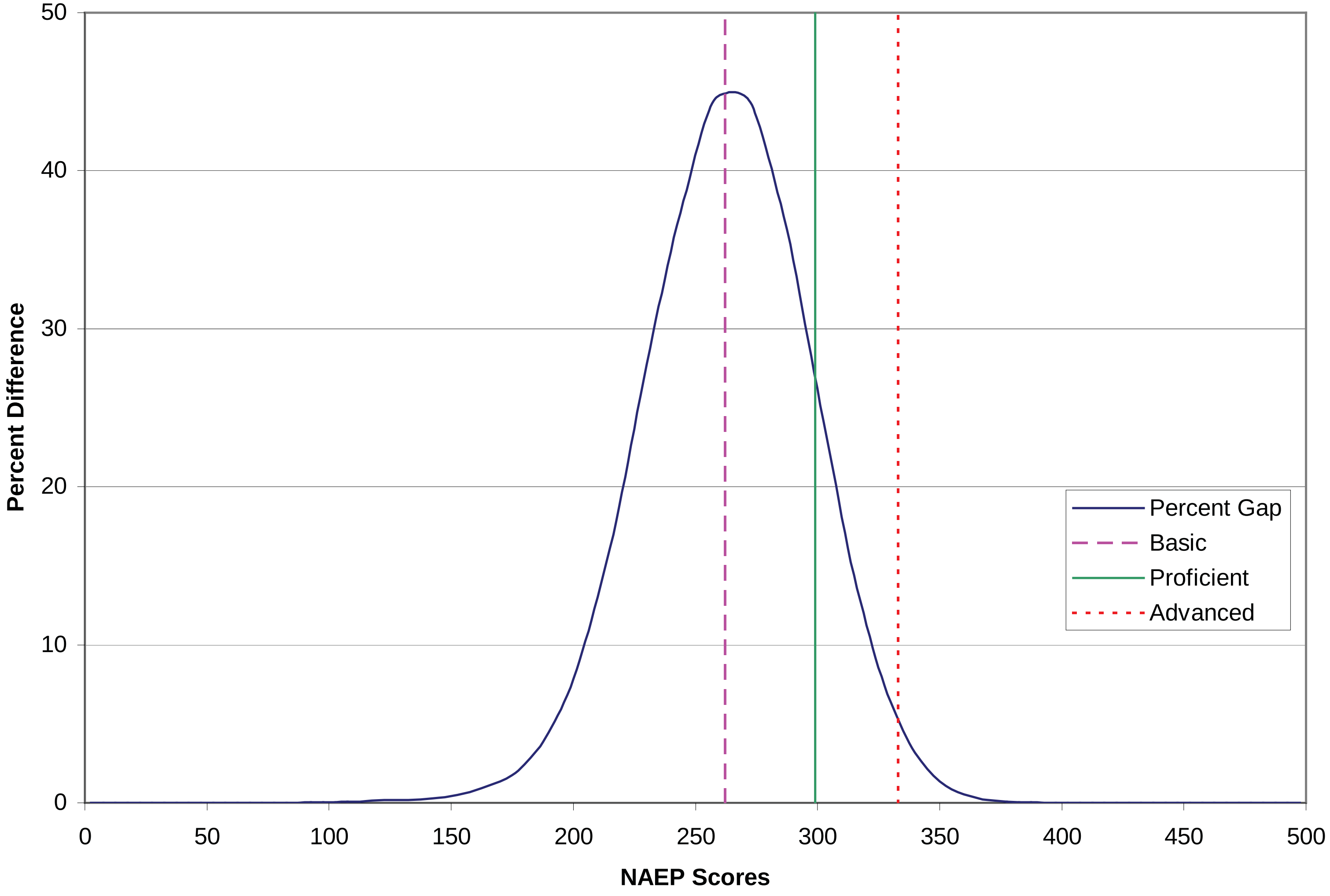


Figure 7
Black/White Horizontal Gap
NAEP Mathematics 2000 Grade 8

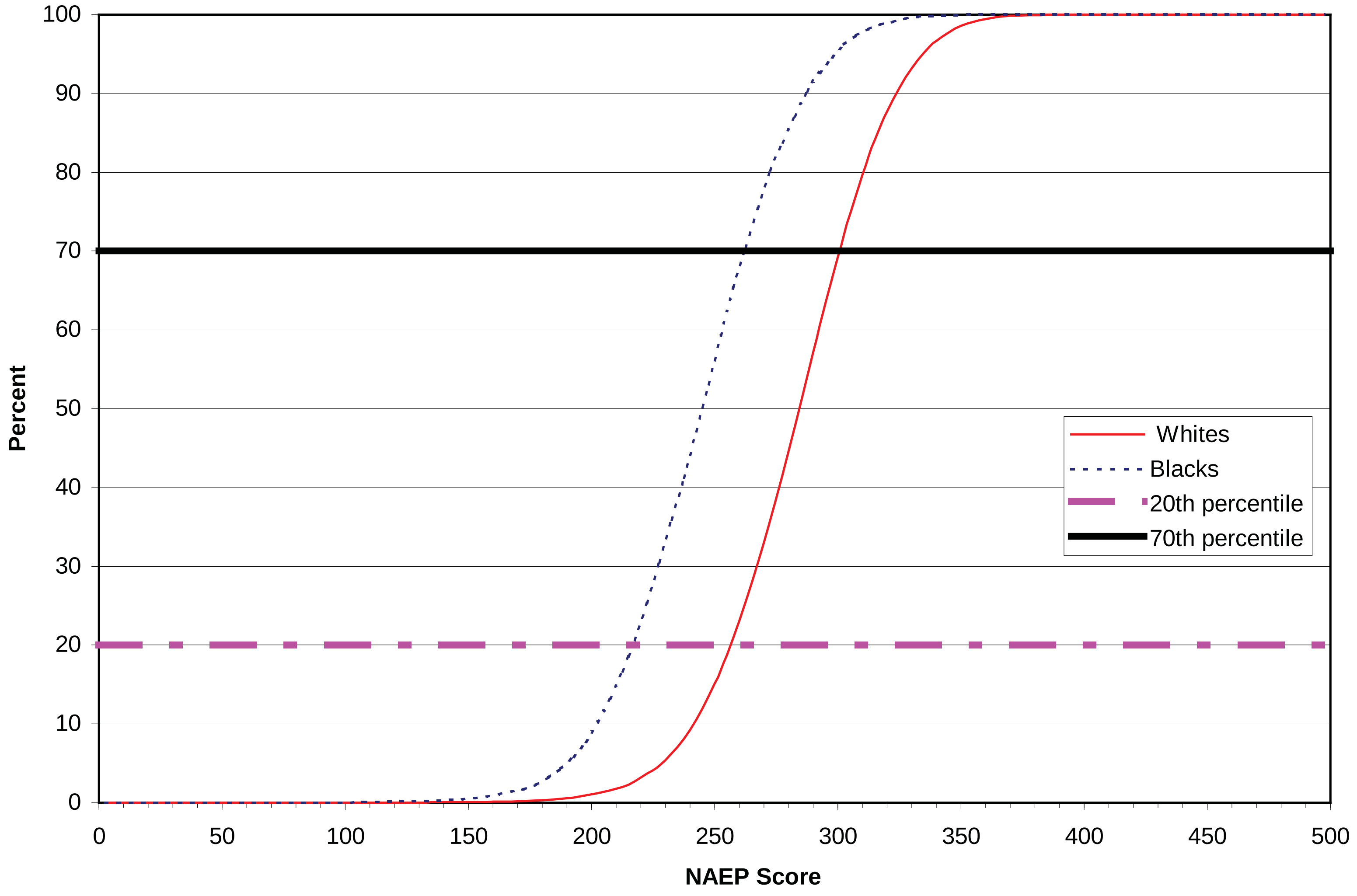


Figure 8
Percentile Differences vs. Percentages
NAEP Mathematics 2000 Grade 8
Whites - Blacks

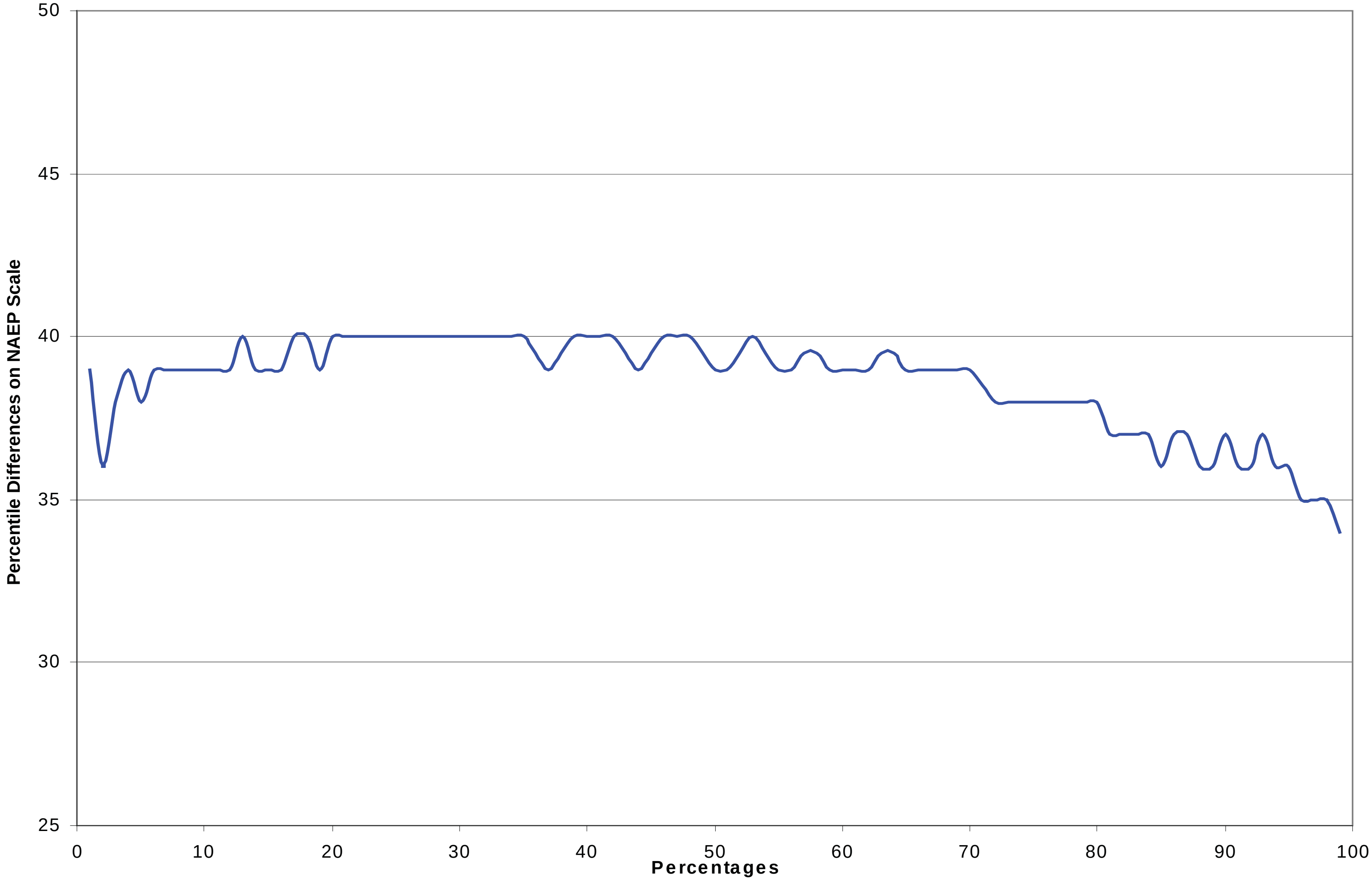


Figure 9
Distributions with Achievement Levels - Blacks and Hispanics
NAEP Mathematics 2000 Grade 8

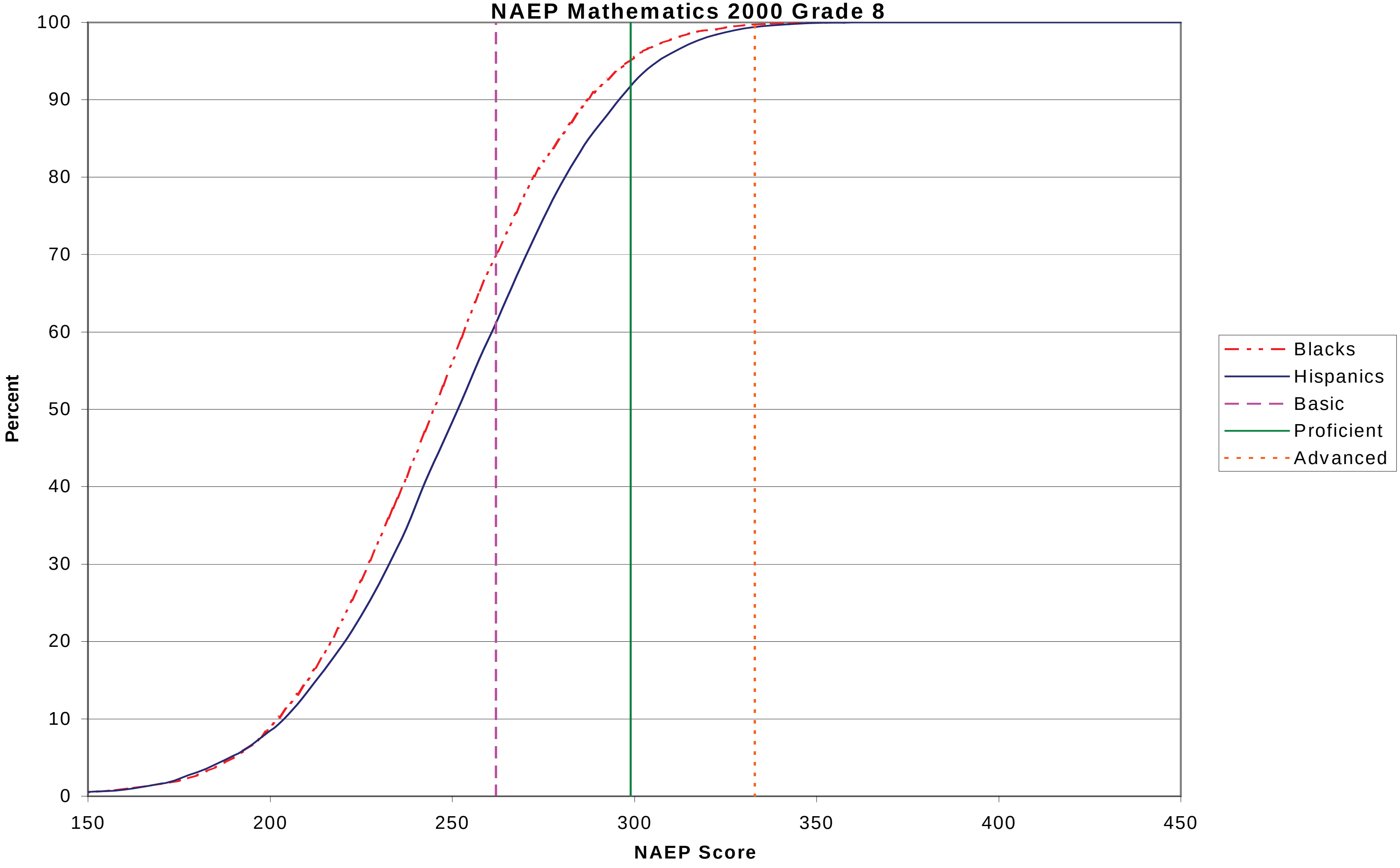
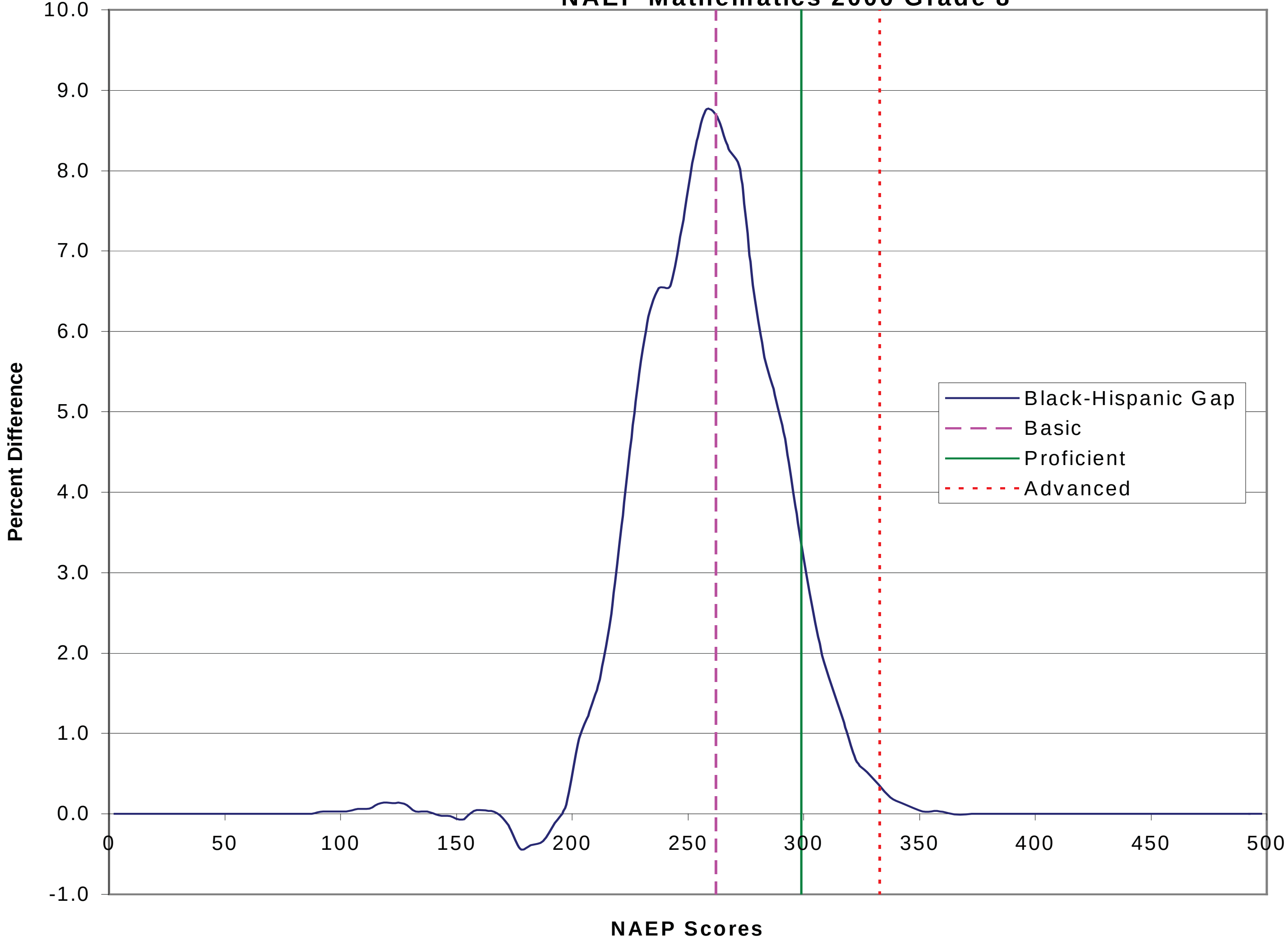


Figure 10
Gap in Percents between Blacks and Hispanics
NAEP Mathematics 2000 Grade 8



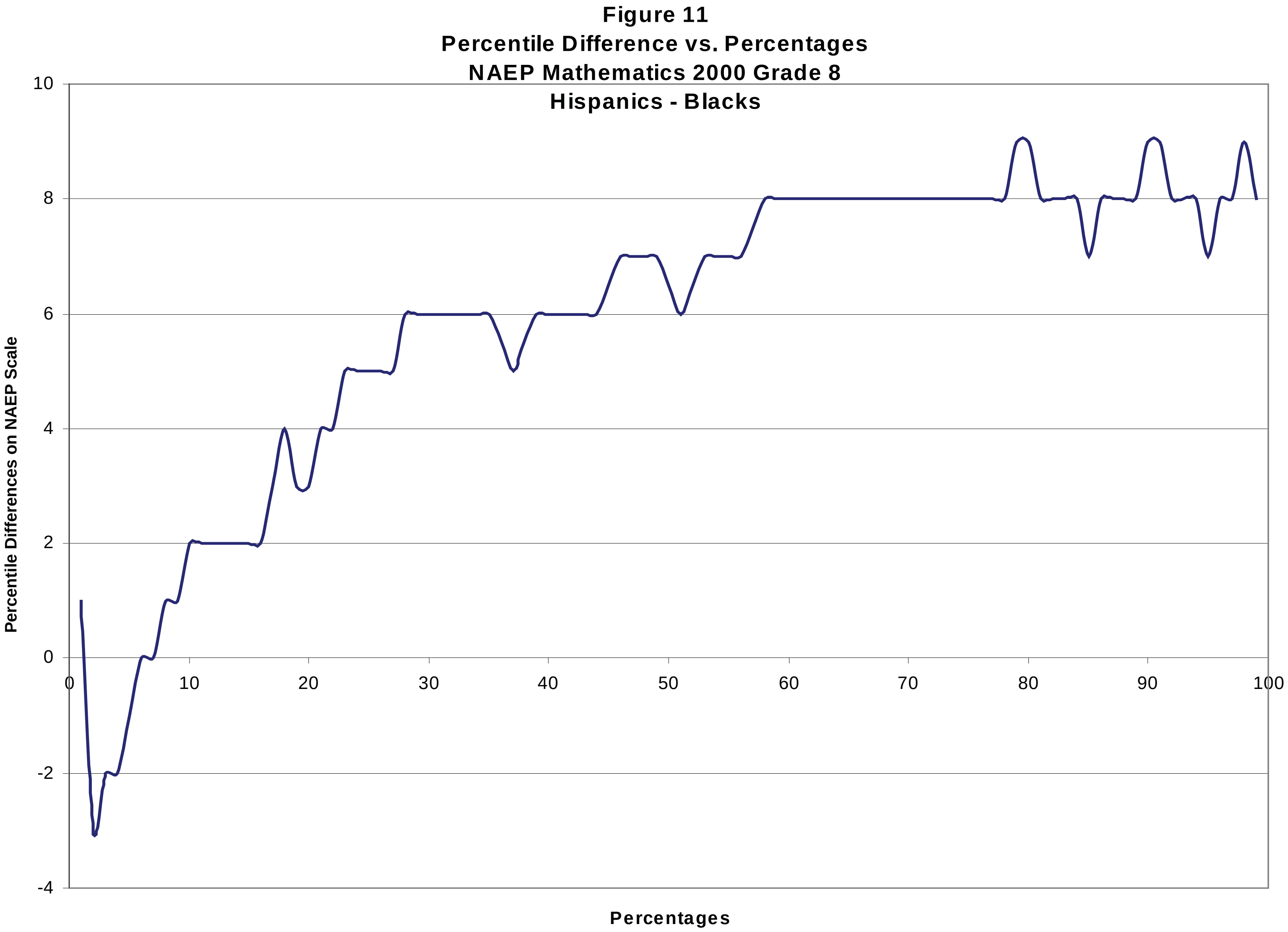


Figure 11
Percentile Difference vs. Percentages
NAEP Mathematics 2000 Grade 8
Hispanics - Blacks

FIGURE 12
Distributions with Achievement Levels Marked
NAEP Mathematics Score Grade 8: 1996 and 2000

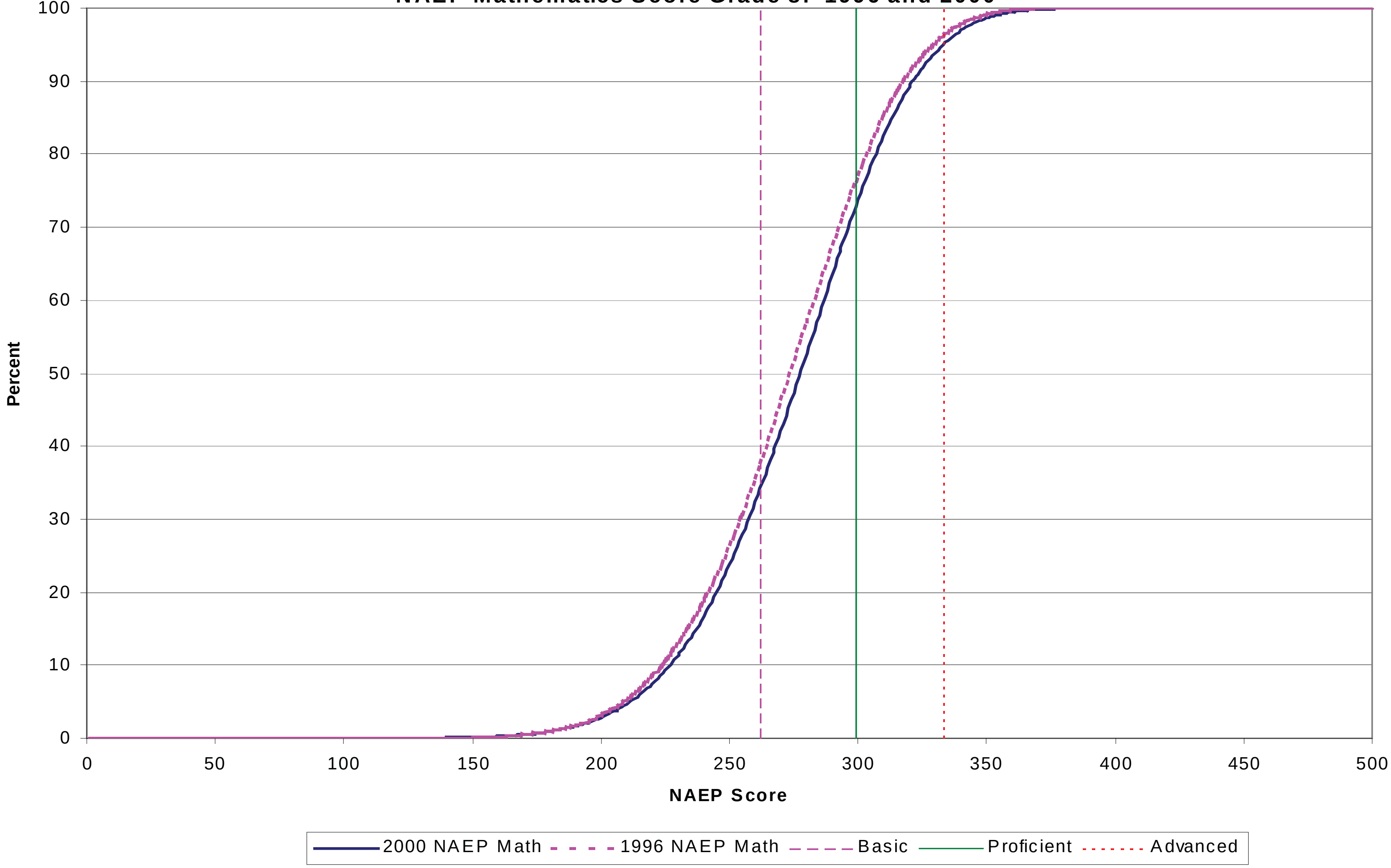


Figure 13
Gap In Percentiles
Ineligible- Eligible: Free Lunch
NAEP Mathematics Grade 8: 1996 and 2000

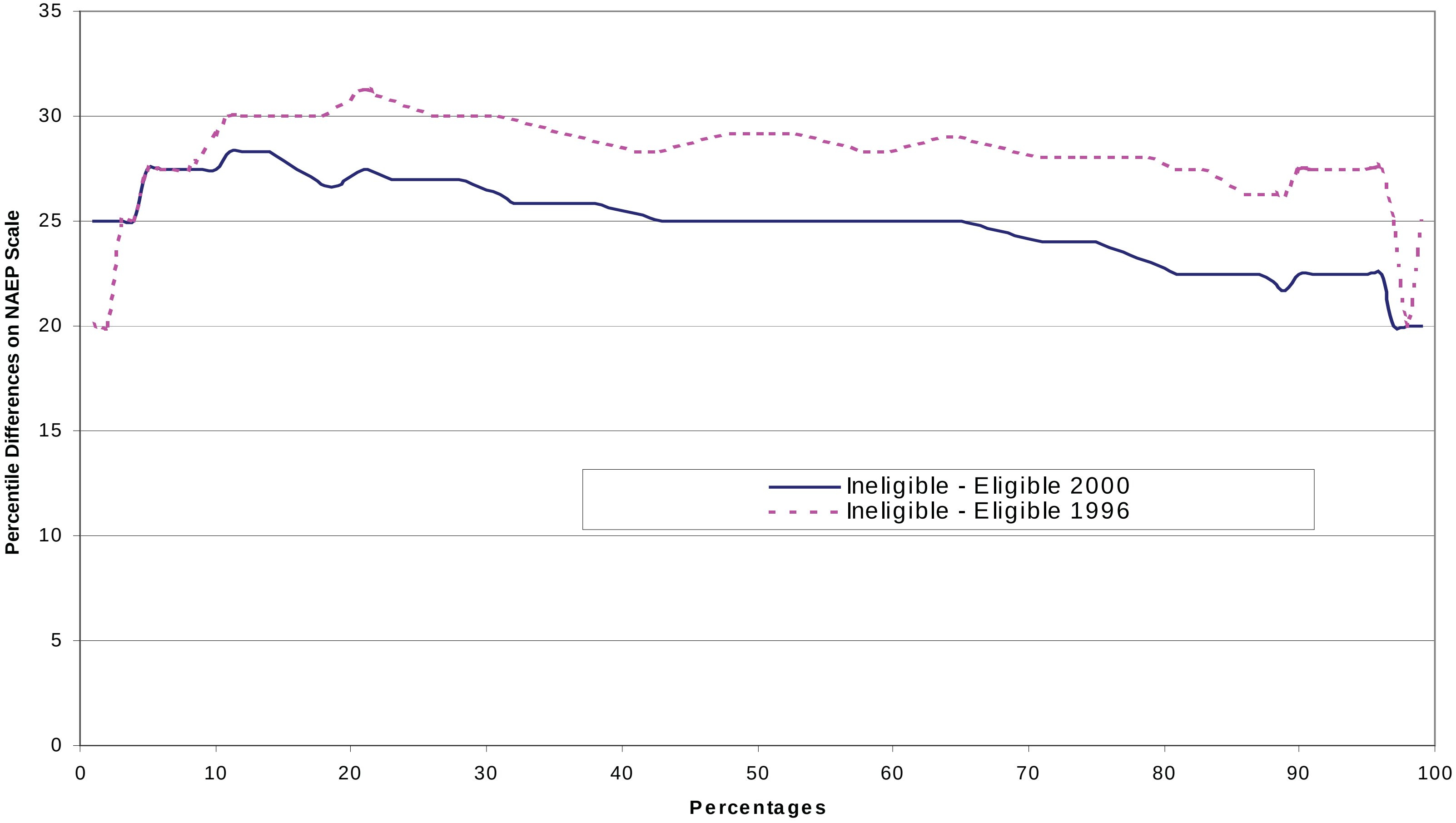


Figure 14
Gap in Percents
Ineligible - Eligible Free Lunch
NAEP Mathematics Grade 8: 1996 and 2000

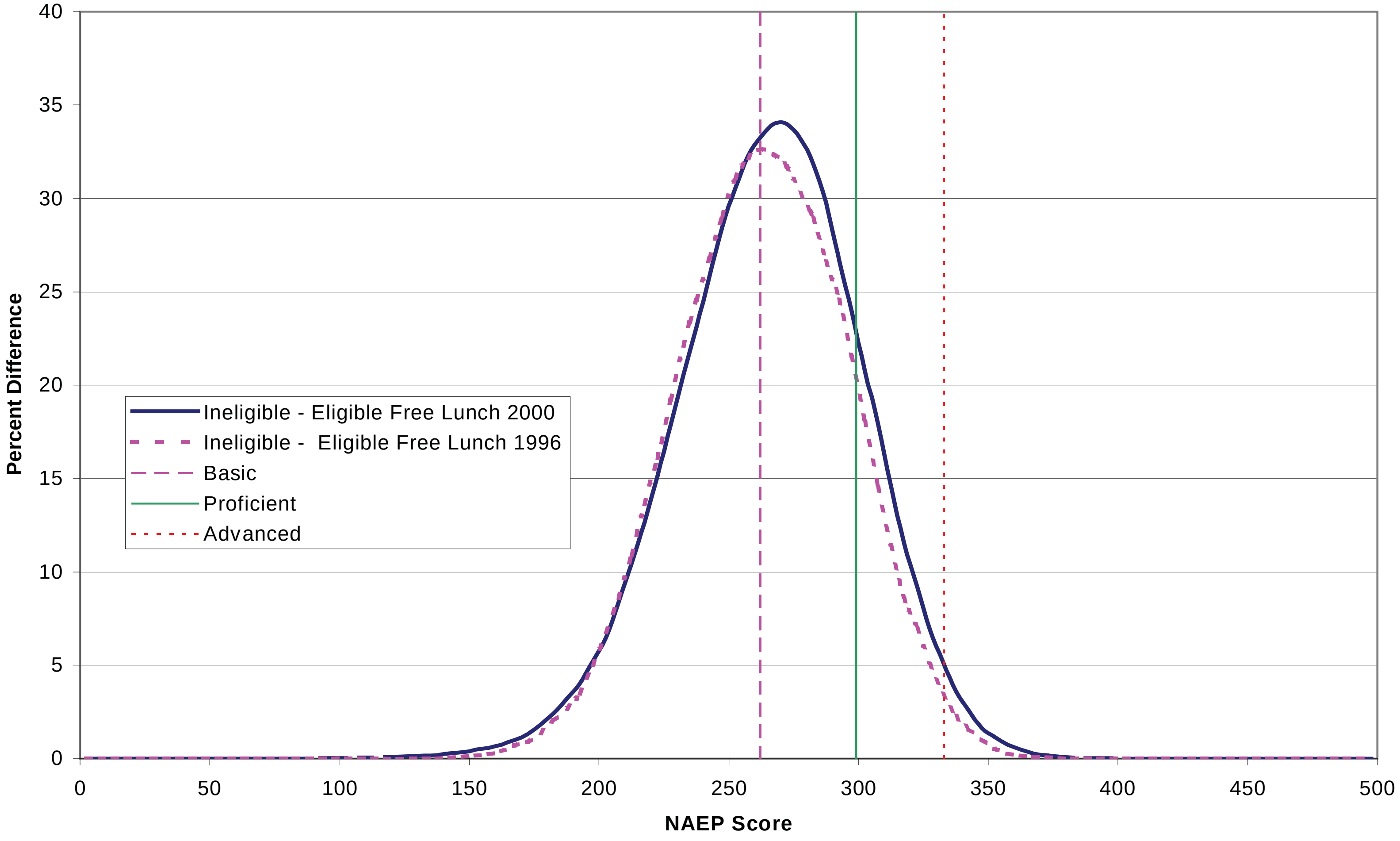
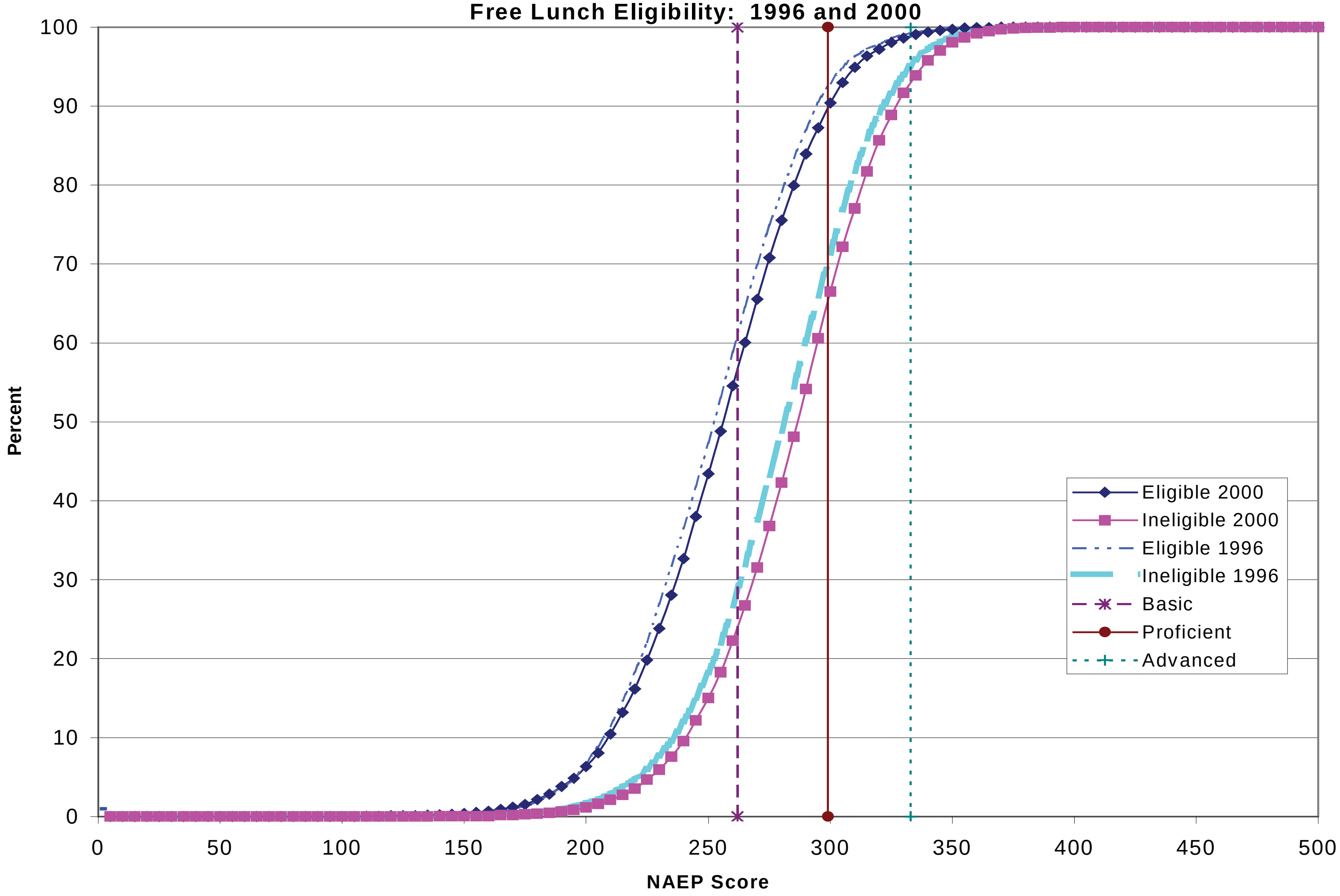


Figure 15
NAEP Mathematics
Free Lunch Eligibility: 1996 and 2000



How Big Is Big When It Comes To Gaps In Scores?

P W. Holland
11/08/01

1. Introduction

Here I begin to address the question of when is a gap between the scores of different groups of examinees “large” or “small”? My approach is not to use statistical significance as the touchstone, (though I mention it) but rather to give illustrations of how large gaps really are in different circumstances. The idea is to suggest what an “inch”, a “foot” and a “yard” might be in the ways that we discuss gaps. My companion piece, called “Measuring Progress in Student Achievement: Changes in Scores and Score-Gaps over Time” is a more detailed look at measuring gaps. This one is about interpreting their size.

By an “inch” I mean a very small gap, a “foot” is one of modest size and a “yard” is really big. These are intuitive notions and I only use physical measurement as a useful analogy rather than something more precise. I believe that by having real examples of gaps on the table we can begin to develop notions of when scores are really different and when changes in scores are substantial. I use the idea of inch, foot and yard because of their simplicity and familiarity and nothing more. While it is true that in some situations “an inch is as good as a mile” in many others these rough distinctions between distances are useful size-guidelines.

2. Getting a rough and ready start.

As a start, consider just how different on average would we expect the NAEP Math scores of two randomly selected public-school fourth-graders to be in 2000? One answer to this involves the standard deviation of the distribution of public-school fourth-grade NAEP Math scores in 2000, which is 31.4 points on the NAEP Math Scale. A useful rule of thumb, which works for distributions like the Normal, is to multiply the standard deviation by 1.13. A distribution with much shorter tails than the Normal, such as the Uniform suggests multiplying the standard deviation by 1.15. Because the tails of score distributions are somewhere in between these extremes, I will compromise and use 1.14 (it hardly makes a difference for a rough start at sizing the problem). From this we would expect an average difference of about 36 points between the scores on the NAEP scale of two randomly chosen public-school fourth-graders. For 1996, the standard deviation for public-school fourth-graders is the same so again the average difference would be about 36 points. Table 1 summarizes these values in its third column for both fourth- and eighth-grade public-school students for Math and Reading, using the two most recent assessment years for these two Subjects.

(Table 1 about here)

I give these values as a first step toward assessing what we might mean by a very large difference, i. e. bigger than a "yard". Students vary in many ways across the country and the information in Table 1 shows one way of indicating how much they do. While there are some differences in subject, grade, and year, these values range from 36 to 47 points on the two NAEP scales.

3. What have social scientists done in this area?

Now, social scientists have wrestled with the problem of deciding when an effect is small, medium or large for some time. The most well-known articulation of this idea is that in J. Cohen's work, Cohen (1988). Cohen proposes that effects less than 20% of a standard deviation be called "small", those near 50% of a standard deviation be called "medium" and those in excess of 80% be called "large". While obviously subjective, these rules of thumb have served a useful purpose in a variety of settings, including the evaluation of educational programs. For this reason, I added these values to Table 1. Note that the "Random Difference" column uses 114% of a standard deviation and its values are bigger than all of the values from Cohen's rule of thumb.

Using Cohen's criteria, "small" ranges from 6 to 8 points, "medium" from 16 to 20 points and "large" from 25 to 30 points. Again, I emphasize, these figures are just to get us started so that we have some place to orient our thinking. It is a bit arbitrary, but not completely so. I think it is clear that if someone advocated moving scores by 60 points on the NAEP scale, the data in Table 1 suggest that this would be an enormous change in several senses of the word.

4. Using Achievement Level differences to define a "Yard".

Another easy way to get a feel for differences in NAEP points is to examine the differences in the Achievement Level cuts for the two subjects and grades given in Table 1. These are given in the last column of Table 1. The Achievement Levels indicate substantial differences in student performance, not small differences, and we see that this interpretation fits well with the other values in the table. The AL differences are generally bigger than Cohen's "large" effects, and smaller than the "Expected random differences". I interpret this as support for using the AL differences, as a rough *measure of large changes*, i.e., the "yard". The average of the AL differences in Table 1 is 35 NAEP points. If we carry this through to feet and inches as suggested, then a single NAEP point is about an inch and 12 of them is about a foot. This is pretty easy to remember, but we need to check to see if it comports with other differences. In particular, a single NAEP point is quite a bit smaller than what Cohen's rule of thumb would be for a "small effect" which was from 6 to 8 points on the NAEP scale. A small effect is about a "half-foot" in these terms.

Another way to look at what an "inch" means is to consider statistical accuracy. NAEP results give standard errors for various quantities, the easiest to understand being the mean scores for various groups. The standard errors for the means for all US public school students for the two grades and assessment years considered earlier are given in the last column of Table 2. All these standard errors are about 1 NAEP point. Because the public-school sample is the largest in NAEP, the accuracy of these mean estimates is the most statistical accuracy that NAEP achieves. It just means that had a different sample been drawn, the results would differ from what this sample gave us by one or two NAEP points in either direction. I take this to mean, that a NAEP point is really small, and is at the noise level of the data. Changes of an inch, in this sense, are generally imperceptible, in the sense of being detectable in a statistically reliable way.

(Table 2 about here)

5. What about other types of score differences?

We can also look at changes over time in average scores. Table 2 gives the mean scores for all public school students for the two Subjects and grades for the two assessment years that appear in Table 1. In addition, I have shown the differences across assessment years for each subject and grade-level, and then divided them by 4 years to get an annual rate of change in average NAEP scores. We see that a rate of one NAEP point per year is the simple message of Table 2. This is just an "inch" in terms that we have developed so far, and suggests that annual change at the national level is quite slow, as is found in many other domains of American life. It is so slow, that year-to-year change is hardly detectable at the accuracy level that NAEP is able to achieve with its very sophisticated sampling design.

Scores vary by many factors. One that is of general interest is by region of the country. Using the two Subjects and assessment years in Table 1, the difference between the highest and lowest averaging Region of the

Country ranges from 9 to 18 NAEP points, and averages about 12 points, that is, a “foot”. I used this comparison of regions to avoid comparing specific States, but this type of comparison is closely related to state differences. It suggests that overall, state differences are of modest size more in the “foot” range than the “yard”.

While it is a bit more problematic, it is worth seeing how large the gaps are between the grade levels in a given year. Using the data in Table 2 we see that this ranges from 45 to 48 points, and if we divide this by four years we get 11 to 12 points per year as a rough estimate of “normal” yearly growth from fourth to eighth grade. Again, a rate of about a “foot” a year.

6.What about other scales?

The numbers in Table 1 depend on the scale of measurement used. If for some reason, NAEP changed its scale all these numbers would change to reflect that. We might also ask about score differences in terms of the *percentiles* of two randomly selected US fourth graders in the NAEP 2000 Math assessment. This measure of the average gap between student scores does not depend on the scale of measurement used. It is always the same, for any grade, scale, or subject. It is always about 33 percentage points. This gives us another version of what a “yard” might be. However, from what I have shown for the NAEP scale, I would suspect that it is probably a bit *too large* for a “yard”.

A final comment about percents above achievement levels. The meaningfulness of a difference in a percent depends on where one starts. Changing 10 percentage points when starting at 50% is one thing, but the same amount of change when starting at 95% is impossible. You just can not get more than 100%. For this reason, I think my attempt to give a rough meaning to small, medium and large score changes has the best chance of working when we stick with the NAEP scale and use its very well-developed point system. The Achievement Levels give meaning to places along this scale and the kind of examination that I have given here can add further information to the question: “Is that a big change or what?”

Reference

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences. Second Edition. Hillsdale NJ: Erlbaum

Table 1: Some measures of size of gaps based on standard deviations of test scores from some recent NAEP Assessments

Subject and Assessment Year	Grade	Standard Deviation	Expected random differences	Cohen's Effect sizes S, M, L	Differences between AL's P-B, A-P
Mathematics					
1996	4	31.4	36	6, 16, 25	35, 33
	8	36.6	42	7, 18, 29	39, 34
2000	4	31.4	36	6, 16, 25	
	8	37.4	43	7, 19, 30	
Reading					
1994	4	40.9	47	8, 20, 33	30, 30
	8	36.8	42	7, 18, 29	38, 42
1998	4	37.6	43	8, 19, 30	
	8	34.8	40	7, 17, 29	

Table 2: Mean NAEP Scores and Their Differences For Two Assessment Years And Two Subjects For Fourth And Eighth Grade Public School Students.

Subject and Grade	Year	Mean score	Difference between Assessment Years	Per year change	Standard Errors of the Mean Scores
Mathematics					
4	1996	222.2			1.0
	2000	226.2	4	1	1.0
8	1996	270.5			1.2
	2000	274.4	4	1	0.8
Reading					
4	1994	212.3			1.1
	1998	215.4	3	.75	0.8
8	1994	257.3			0.8
	1998	261.4	4	1	0.8

Appendix C

Wendy Yen
ETS K-12 Works
January 30, 2002

The Use of NAEP Results in a Confirming Role for ESEA: A Thought Experiment with Historical Data from State A

Introduction

This report explores how NAEP data might play a role in confirming state results under new ESEA legislation. Using historical state and NAEP data from State A, alternatives for data presentation are examined. Several recommendations are made about appropriate and inappropriate ways of using NAEP data in a confirming role.

As summarized by Reckase (2002), the No Child Left Behind ESEA legislation requires of states a variety of evidence that the state has met ESEA requirements:

1. The state must have challenging academic content standards.
2. The state must have challenging student academic achievement standards.
3. The standards apply to public elementary and secondary school children.
4. The standards exist for mathematics, reading or language arts, and science (beginning in 2005-2006).
5. The achievement standards should be aligned to the content standards.
6. The achievement standards should have two levels of high achievement, proficient and advanced. The proficient standard is the critical one that appears later in the legislation.
7. A basic level is included to help track the performance of lower performing students.
8. The state should have an accountability system that provides a definition for adequate yearly progress.
    a. The definition should be statistically valid and reliable.
    b. The definition should result in continuous and substantial academic improvements for all students.
9. States should work toward narrowing achievement gaps.
10. There should be separate measurable annual objectives for the following groups:
    a. Economically disadvantaged.
    b. Major racial and ethnic groups.
    c. Disabled.
    d. Limited English proficient.
11. The starting point for the process is the 2001-2002 school year. For that year the base level is the percentage of students

meeting or exceeding the state's proficient level of achievement. This point is defined as the higher of the follow two options:

a. The lowest achieving group
b. The school at the 20th percentile ranked by the percentage of students at the proficient level.

12. The state will have 12 years to have all students above the proficient level. This seems to imply 12 years of education to reach the proficient 12th grade standard.
13. There should be annual measurable objectives for mathematics and reading.
14. Each target group should have its own minimum.
15. There should be intermediate goals that include equal increments over the 12 years.
16. There is a 10% cushion for meeting the standard. A state may miss the target by 10% in a given year.
17. At least 95% of each group must take the assessment.
18. The state can average up to three years of data.
19. Before 2005, students must be assessed once in 3 – 5, 6 – 9, and 10 – 12.
20. Starting in 2005, students must be assessed every grade 3 – 8.

State A was selected as a useful example of how a state might present its case of meeting these requirements. This state has had a continuous standards-referenced state testing program in Reading and Mathematics in grades 3 to 8 since 1993, which provides an unusually rich database. State A data for grades 4 and 8 are selected as the focus of this report. State A has also participated in NAEP in years that substantially overlap the state testing years. NAEP data used in this report come from 1992, 1994, and 1998 (Reading grade 4), and 1992, 1996, and 2000 (Mathematics grades 4 and 8). 1998 data for NAEP Reading in grade 8 are also referenced, but because only one year's data are available, they are not useful in trend analyses.

This report conducts a thought experiment that imagines that 1993 is the base year for the implementation of the new ESEA legislation, and it displays the state results relative to ESEA performance requirements. The historical NAEP data are then evaluated in a confirming role. It must be acknowledged that these NAEP data are more sparse than the biennial NAEP data to be collected under the new legislation. This example is further limited because State A could rightly choose to analyze and present its data in a manner that differs from the procedure chosen here. Also State A has, and would be expected to present, far more information about the quality of its testing and educational programs than is presented here. Despite these limitations, this experiment can provide some insight about NAEP's usefulness in a confirming role.

An Example of How State A Might Present Its Case for Meeting ESEA Requirements

Assumptions

1) State A would provide information in defense of the quality of its standards (Requirements 1, 2, 4, 5) and the applicability of the standards to all children in public schools (Requirement 3).
2) The state would select what it calls "performing at or above grade level" (equivalent to performing at Performance Levels III or IV on the state exam) as the "proficient" level required by ESEA. In this report the students performing at or above grade level will be described as reaching the State A standard. State A would also present data for an Advanced level and for a Basic level (Requirements 6 and 7).
3) The state would present results for subgroups identified in the legislation (Requirement 10). This report presents results for three major ethnic groups (White, Black, and Hispanic) for Mathematics for three years (1998, 1999, 2000) for the state exam and two years (1996, 2000) for NAEP. As an example of analysis of performance for economically disadvantaged students, some analyses of NAEP data are presented for students eligible for free school lunches.
4) To satisfy requirements 8, 11, 12, 13, 14, 15, and 16, within each content area, grade, and category of student, the state would set two minimum targets for Adequate Yearly Progress. The Primary Target is geared toward having all students reach the state standard in 12 years. The "starting point" for this target is the percent of students reaching the standard in the base year, P_B. The Primary Target for performance for the *i*-th year beyond the base year is the linear interpolation, over 12 years, between P_B and 100:

   $$\text{Primary Target}_i = P_B + i(100-P_B)/12.$$

   It can be noted that this target definition necessarily reduces gaps between groups (Requirement 9).

   The Alternate Target provides a 10 percent reduction, as compared with the previous year, in the percent of students not reaching the standard:

   $$\text{Alternate Target}_i = 100- .9(100-P_{i-1}).$$

   The state could be counted as having satisfied the ESEA requirements if either the Primary or Alternate Target is met.

5) Satisfying Requirements 19 and 20, students are assessed in every grade 3 to 8.
6) For Requirement 17, the state would provide evidence that at least 95 percent of each group of students is tested.
7) No averaging of results over years will be done.

NAEP data

NAEP results could be analyzed for three proficiency levels: Basic, Proficient, and Advanced. For the purposes of this report, the NAEP Basic level was chosen as the primary focus of attention because it is the proficiency level most similar to the State A standard in terms of “difficulty” (i.e., percents of students reaching the standard) in the base years. However, as will be noted, the NAEP Basic level is somewhat more challenging than the State A standard for all grades and content areas examined, and this difference in difficulty has implications for results.

While the NAEP data have substantial overlap with the State A test data in terms of years of testing, the two tests do not match exactly in terms of years of assessment. Also, there are substantially fewer years of data for NAEP than the state test. This mismatch in years tested would normally limit the appropriateness of using the NAEP data in a confirming role. However, for the purposes of this illustration, all the NAEP data are used and compared with all the State A data.

The primary method of examining NAEP data in this study is to review trends in the percents of students reaching the Basic level and compare these trends with the percents of students reaching the Proficient level for the state test. Because the state data are compared to targets, it was of interest to examine what would be found if NAEP data were compared to such targets. So, for purposes of illustration, the Primary Target is calculated for the NAEP data that are available. It can be noted that because the NAEP Basic level is more challenging than the State A standard, more growth will required each year on NAEP than the State A exam to reach the 100 percent standard in 12 years. Because NAEP data are not available in consecutive years, it is not possible to calculate the Alternate Target for NAEP.

The ESEA legislation requires measurement of achievement with respect to students reaching performance levels. Such measurement does not reflect changes in the performance of students who are not near the cut-point for the performance level being examined. Cumulative distributions of scale scores, and changes in those distributions, reflect changes in the performance of students throughout the achievement continuum. Examples of such analyses based on NAEP data (which have been proposed by Paul Holland, “Measuring Progress in Student Achievement: Changes in Scores and Score-Gaps Over Time,” 2001) are presented to show their power and usefulness.

Gaps between the performance of students in different subgroups can also be readily understood via appropriate graphs. These graphs (also proposed by Paul Holland) display for each percentile the difference in NAEP scale scores for the two subgroups of interest. As will be seen, use of these graphs makes it straightforward to determine if gaps are decreasing over time for students at all levels of achievement, not just at selected performance levels.

Results

Overall State Performance Trends. Figure 1 displays, on the left side, the percents of students by year reaching the State A standard for Reading in grade 4. Overlaid on the results for the State A exam are that state's NAEP results, in particular the percent of students reaching the Basic NAEP level. In general, State A shows steady progress in student performance on both the State exam and NAEP. The trend lines for the two assessments have similar slopes.

On the right side of Figure 1is the percent of State A students reaching the highest proficiency level on the State A exam (Level IV) and the state's performance at or above the NAEP Proficient level. While it is reasonable to make such comparisons, for this figure and similar figures, there will not be further discussion or analysis of these results for the higher of the two proficiency levels.

The left-hand plot in Figure 2 shows the trend lines for grade 4 Mathematics. It is striking that the NAEP growth trend line is so similar to the trend line for the state test. In Figure 3 there are insufficient data to calculate a trend for Grade 8 NAEP Reading. In Figure 4 are the trend lines for grade 8 Mathematics. Again the similarity of the growth trends for NAEP and the state test are remarkable.

In terms of trends in performance over time, it is clear that the NAEP data confirm the growth seen with the state test.

Subgroup Performance Trends. Figures 5 to 8 display trend lines for White, African-American, and Hispanic students, for the state test and NAEP. In each figure the performance of White students is displayed on the left and the performance of the minority students, Black/African American or Hispanic, is displayed on the right. As was done with overall state performance, NAEP performance is included in each graph. Results are presented only for Mathematics, because NAEP data were not available for measuring trends for Reading for the subgroups in the years for which state test data were available for subgroups. The differences in the years covered by the two assessments and the small number of data points limit the suitability of comparisons between the two assessments.

Regardless of these limitations, growth trends seen with the state test for subgroups are also seen with NAEP.

Overall State Performance Relative to Targets. Table 1 contains for grade 4 Reading comparisons of the percents of students reaching the State A standard with the percents expected by the Primary and Alternate Targets. With respect to the Primary Target, the difference between the target and the state results increases as the years progress. Another way to see this is to imagine that on Figure 1a straight line is drawn from the base year performance (62 percent in 1993) to the 12-year target (100 percent in 2005); that Primary Target line would have a steeper slope than the line representing actual state performance, and the state results would be seen to farther away from the Primary Target in later years.

As part of the exploration of using ESEA-like targets for the NAEP data in Table 1, the Primary Target is calculated for NAEP. The NAEP results are farther away from the NAEP target than the state results are from the state targets.

Table 2 shows the results for grade 4 Mathematics. State A results are very close to the targets in all years, and the NAEP results are farther from their targets. Table 3 displays the results for the state test for grade 8 Reading, but there are insufficient data to calculate targets for NAEP. In Table 4 the grade 8 Mathematics NAEP results are not as close to their targets as the state results are to their targets.

For all grades and content areas examined, NAEP results are farther away from the NAEP targets than the state results are from the state targets. Because the NAEP performance in the base years is at a lower percent reaching the standard than the state test is in the base years, the target growth rate for NAEP is necessarily greater than the target growth rate for the state test. In using NAEP in a confirming role, it appears unreasonable to apply the Primary Target to NAEP results because it is substantially more stringent than the targets applied to the state results.

Subgroup Performance Relative to Targets. In Table 5 the subgroups' performance is compared with Primary and Alternate Targets calculated for each group. Note that the Primary Targets are calculated with respect to different base years in these tables than those used in Tables 1 to 4. This occurs because results by ethnic group were not available for this report for earlier years. It is worth noting that because the base year for the subgroup calculations (1998) is a later year than the base year for the total state analyses (1993), and performance was higher in later years, the amount of growth per year needed to meet the 12-year Primary Target is lower for the subgroup analyses than it was for the total state analyses.

For grade 4 Mathematics (Table 5), the students in all the subgroups (except Black students for the Alternate Target in 2000) met the Primary and Alternate Targets for the state test in every year. The NAEP Primary Target was not met for Hispanic students in 2000, despite substantial improvement in that subgroup's performance between 1996 and 2000.

At the bottom of Table 5 gaps between the performance of minority students and White students are examined. Gaps are measured by differences in the percents of students meeting the standard. The gaps closed for both minority groups in both 1999 and 2000 for the state test and for 2000 for NAEP.

Gaps can also be examined by examining Figures 5 to 8: If the slope for the growth trend line for the minority students is greater than the slope for the White students, then a gap is closing.

Table 6 displays the subgroups' performance for grade 8 Mathematics. For the state test, none of the subgroups met either target in 1999, but all three subgroups met at least one of the targets in 2000. With respect to NAEP performance, only the White students were measured as meeting the Primary Target.

In terms of gaps, the White-Black gap decreases on the state test in both 1999 and 2000 but it increases for NAEP between 1996 and 2000. The White-Hispanic gap decreases on the state test only for 2000, and NAEP shows a decrease from 1996 to 2000.

In summary, in terms of meeting targets for subgroups, there could be substantial inconsistency between results for the state test and results for NAEP. When there were inconsistencies in meeting targets, the NAEP results tended to be substantially less positive than the state test results. Gap closure tended to be somewhat more consistent between the state results and NAEP.

NAEP Distributional Information. Figure 9 displays the cumulative distribution of NAEP grade 4 Reading scale scores for 1992, 1994, and 1998. Also displayed on that figure are the cut-points for the NAEP Basic and Proficient performance levels.

A curve that is to the right of another curve shows higher performance (i.e., a curve to the right shows higher NAEP scale scores at a given percentile). Note that the 1998 and 1994 curves intersect. Among lower-achieving students (those below the middle—60th percentile-- of the State A distribution), students were doing better in 1998 than 1994. However, among higher achieving students, students were doing worse in 1998 than 1994. If one were evaluating performance based solely on a particular performance level, that evaluation could misrepresent how students were doing at other performance levels (i.e., looking at the Basic level, one would conclude that State A was improving performance, but such improvement would not be seen at the Proficient and Advanced levels).

It can be very daunting to review tables of numbers that show percents of students reaching various standards in various years and to attempt to draw overall conclusions. On the other hand, a quick review of the cumulative distributions presents a clear picture.

Gaps between the performance of students in different subgroups can also be readily understood via an appropriate graph. Figure 10 displays cumulative distributions of NAEP grade 4 Mathematics performance for White and Black students for 1992, 1994, and 1998. The gaps between groups are reflected in the differences between the curves for a given year. It is difficult to evaluate gaps by looking at cumulative distributions, and to make evaluation easier, the differences are calculated and displayed in Figure 11. Each curve shows the difference between the NAEP scale scores (the gap) for students at a given percentile. For example, in 1994 at the 50th percentile, White students had scale scores that were (about) 33 points greater than those for Black students. In 1998, the gap at the 50th percentile was about 28 scale score points. The fact that the line for 1998 is lower than the line for 1994 (except among students below the 20th percentile) means that the gap was closing between those years.

Figures 9 to 27 display distribution and gap information for grade 4 Reading (1992, 1994, 1998) and Mathematics (1992, 1996, 2000), and grade 8 Mathematics (1990, 1996, 2000). (Data needed for distribution comparisons between years were not available for grade 8 Reading.)

The cumulative distribution graphs were reviewed to determine if NAEP performance was improving in general for the state as a whole, major ethnic subgroups, and those students eligible/ineligible for free school lunch. These comparisons were conducted only for Mathematics because the Reading results were not available for year spans that overlapped the years for which state test subgroup data were available. Changes in the more recent pair of years (1996 and 2000) were classified as being predominantly positive (+), predominantly negative (-), or mixed or insubstantial (0).

These evaluations are summarized in Table 7 and compared with results for the State A test. For example, from 1996 to 2000, for the state as a whole grade 4 NAEP Mathematics performance increased throughout the distribution (+, Figure 14). For subgroups, the performance of White students increased (+, Figures 15 or 17), as did the performance of Black students (+, Figure 15) and Hispanic students (+, Figure 17). The gap in White-Black student performance predominantly decreased (+, Figure 16), while changes in the gap in White-Hispanic student performance were mixed (+/0), with the gap narrowing among lower achieving students and staying fairly constant for higher achieving students (Figure 18).

Also appearing in Table 7 are results based on the State A exam. NAEP conclusions about growth in performance agree with the results for the State A test. However, NAEP does not confirm the state gap conclusions in about half the cases. For grade 4 Mathematics the White-Hispanic gap improves for the state test, but gap results are mixed for NAEP (Figure 18). For grade 8 Mathematics, the White-Black gap improves for the state test but get worse for NAEP (Figure 23).

Discussion

This report describes a thought experiment in which historical state and NAEP achievement results are analyzed in light of new ESEA legislation. In particular, possible alternative ways of using NAEP results in a confirming role are examined. The following conclusions are reached:

1) In terms of general trends in improvements in performance for matching years, NAEP results were found to confirm results for the State A test. Increases in the percents of students reaching state standards were matched, often to a striking degree, by similar increases in the percent of students reaching the NAEP Basic level. This match occurred for the state as a whole and for the major ethnic groups.

2) The state NAEP results never met adequate yearly progress goals implied by the ESEA Primary Target of all students reaching the standard in 12 years. The NAEP level that was most similar in difficulty to the state standard in the base year was the Basic level. However, in the base year a lower percent of students met the NAEP standard than the state standard. This meant that to meet the Primary Target, growth on NAEP had to occur at a greater rate than growth on the state test. Given that the state test is much more likely than NAEP to match the state content standards and curriculum, it would not be rational or fair to expect growth on NAEP to be greater than, or even equal to, growth on the state test. Thus, unless growth targets for NAEP were moderated to be lower than growth targets on the state test, it would not be reasonable to use NAEP in a confirming role by examining whether NAEP growth met ESEA Adequate Yearly Progress growth targets. A case could be made that any amount of growth on NAEP should be sufficient to confirm state growth that meets ESEA targets.

3) In terms of the closure of ethnic group gaps, there were few data points available to evaluate the consistency of NAEP and the state test, and data for matching years were not available for measuring these gaps. Further, it was not verified for this study that equivalent definitions of student ethnicity were used for the state test and NAEP. Conclusions for NAEP could differ depending on whether entire distributions of scores, or only one selected performance level, were examined. Given the measurement error inherent in the comparison of changes in differences, it is not clear how NAEP results should be used in a confirming role with respect to gap closure. The examination of entire distributions, rather than selected performance levels, does appear to be a more valid means of examining gap closure with NAEP data. Further exploration of appropriate methods of using NAEP to confirm gap closure is warranted.

Table 1
State A Grade 4 Reading
Percents of Students Reaching Standard

	Year									
	92	93	94	95	96	97	98	99	00	01
State Test										
Actual		62*	66	64	69	68	71	71	72	75
Primary Target			65.2	68.3	71.5	74.7	77.8	81.0	84.2	87.3
Actual- Pri. Target			**0.8**	-4.3	-2.5	-6.7	-6.8	-10.0	-12.2	-12.3
Alternate Target			65.8	69.4	67.6	72.1	71.2	73.9	73.9	74.8
Actual-Alt. Target			**0.2**	-5.4	**1.4**	-4.1	-0.2	-2.9	-1.9	**0.2**
NAEP										
Actual	56*		59				62			
Primary Target			63.3				78.0			
Actual- Pri. Target			-4.3				-16.0			

Note. State standard is "Reaching Grade Level." NAEP standard is reaching the Basic level.

Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

Table 2
State A Grade 4 Mathematics
Percents of Students Reaching Standard

	Year									
	92	93	94	95	96	97	98	99	00	01
State Test										
Actual		64*	67	69	72	75	79	83	84	87
Primary Target			67	70	73	76	79	82	85	88
Actual- Pri. Target			**0.0**	-1.0	-1.0	-1.0	**0.0**	**1.0**	-1.0	-1.0
Alternate Target			67.6	70.3	72.1	74.8	77.5	81.1	84.7	85.6
Actual-Alt. Target			-0.6	-1.3	-0.1	**0.2**	**1.5**	**1.9**	-0.7	**1.4**
NAEP										
Actual	50*				64				76	
Primary Target					66.7				83.3	
Actual- Pri. Target					-2.7				-7.3	

Note. State standard is “Reaching Grade Level.” NAEP standard is reaching the Basic level.

Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

Table 3
State A Grade 8 Reading
Percents of Students Reaching Standard

	Year								
	93	94	95	96	97	98	99	00	01
State Test									
Actual	67*	71	73	73	75	80	80	83	83
Primary Target		69.8	72.5	75.3	78.0	80.8	83.5	86.3	89.0
Actual- Pri. Target		**1.2**	**0.5**	-2.3	-3.0	-0.8	-3.5	-3.3	-6.0
Alternate Target		70.3	73.9	75.7	75.7	77.5	82.0	82.0	84.7
Actual-Alt. Target		**0.7**	-0.9	-2.7	-0.7	**2.5**	-2.0	**1.0**	-1.7
NAEP									
Actual						76			
Primary Target									
Actual- Pri. Target									

Note. State standard is “Reaching Grade Level.” NAEP standard is reaching the Basic level.

Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

Table 4
State A Grade 8 Mathematics
Percents of Students Reaching Standard

	Year									
	92	93	94	95	96	97	98	99	00	01
State Test										
Actual		62*	62	68	68	69	76	78	81	80
Primary Target			65.2	68.3	71.5	74.7	77.8	81.0	84.2	87.3
Actual- Pri. Target			-3.2	-0.3	-3.5	-5.7	-1.8	-3.0	-3.2	-7.3
Alternate Target			65.8	65.8	71.2	71.2	72.1	78.4	80.2	82.9
Actual-Alt. Target			-3.8	**2.2**	-3.2	-2.2	**3.9**	-0.4	**0.8**	-2.9
NAEP										
Actual	47*				56				70	
Primary Target					64.7				82.3	
Actual- Pri. Target					-8.7				-12.3	

Note. State standard is “Reaching Grade Level.” NAEP standard is reaching the Basic level.

Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

Table 5
State A Grade 4 Mathematics

	State			NAEP	
Group	Year			Year	
	98	99	00	96	00
Percent Reaching Standard					
White					
Actual	87*	90	91	77*	86
Primary Target		88.1	89.2		84.7
Actual- Pri. Target		**1.9**	**1.8**		**1.3**
Alternate Target		88.3	91		
Actual-Alt. Target		**1.7**	**0.0**		
Black					
Actual	63*	68	71	37*	58
Primary Target		66.1	69.2		58.0
Actual-Pri. Target		**1.9**	**1.8**		**0.0**
Alternate Target		66.7	71.2		
Actual-Alt. Target		**1.3**	-0.2		
Hispanic					
Actual	70*	77	80	43*	56
Primary Target		72.5	75		62.0
Actual-Pri. Target		**4.5**	**5.0**		-6.0
Alternate Target		73.0	79.3		
Actual-Alt. Target		**4.0**	**0.7**		
Gaps**					
White-Black	24	22	20	40	28
Gap closure***		+	+		+
White-Hispanic	17	13	11	34	30
Gap closure***		+	+		+

Note. Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

**Gaps are differences between percents of students reaching the standard.

*** "+" indicates that a gap was smaller as compared with the previous observation; "-" indicates that a gap was larger.

Table 6
State A Grade 8 Mathematics

	State				NAEP	
Group	Year				Year	
	98	99	00		96	00
Percent Reaching Standard						
White						
Actual	85*	86	88		69*	83
Primary Target		86.3	87.5			79.3
Actual- Pri. Target		-0.3	**0.5**			**3.7**
Alternate Target		86.5	87.4			
Actual-Alt. Target		-0.5	**0.6**			
Black						
Actual	57*	59	64		31*	42
Primary Target		60.6	64.2			54.0
Actual-Pri. Target		-1.6	-0.2			-12.0
Alternate Target		61.3	63.1			
Actual-Alt. Target		-2.3	**0.9**			
Hispanic						
Actual	66*	66	70		41*	57
Primary Target		68.8	71.7			60.7
Actual-Pri. Target		-2.8	-1.7			-3.7
Alternate Target		69.4	69.4			
Actual-Alt. Target		-3.4	**0.6**			
Gaps**						
White-Black	28	27	24		38	41
Gap closure***		+	+			-
White-Hispanic	19	20	18		28	26
Gap closure***		-	+			+

Note. Primary target is the percent of students needing to reach the standard so that all students are projected to meet the standard in 12 years, given linear growth from the base year.

Alternate target is a 10 percent reduction, relative to the previous year, in the percent of students not reaching the standard.

*Base year performance.

**Gaps are differences between percents of students reaching the standard.

*** "+" indicates that a gap was smaller as compared with the previous observation; "-" indicates that a gap was larger.

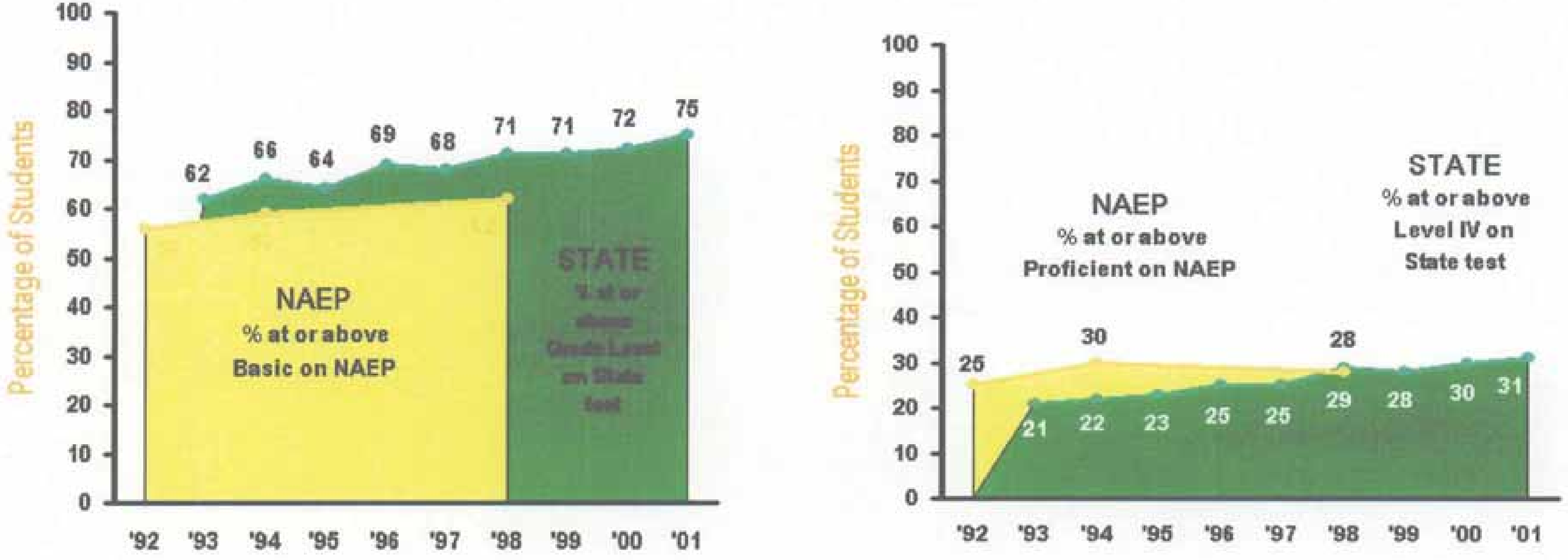
Table 7
State A:
Evidence of Progress in Student Mathematics Achievement

Group	Grade 4		Grade 8	
	State Test	NAEP	State Test	NAEP
Performance*				
State	+	+	+	+
White	+	+	+	+
Black	+	+	+	+
Hispanic	+	+	+	+
Eligible Lunch		+		+
Ineligible Lunch		+		+
Gaps**				
White-Black	+	+	+	-
White-Hispanic	+	+/0	+	+/0
Inelig-Elig. Lunch		+		-/0

Note. Changes evaluated for 1996 to 2000 for NAEP and 1998 to 2000 for the state test.
"+" means predominant increase in performance or closing of gap.
"0" means mixed or insubstantial changes.
"-" means predominant decrease in performance or widening of gap
*Performance changes measured by cumulative distributions for NAEP and by percent reaching standard for state test.
**Gaps measured by graphs of Gaps in Percentiles for NAEP and by changes in percents reaching the state standard for the state test.

Figure 1. State A – Reading, Grade 4

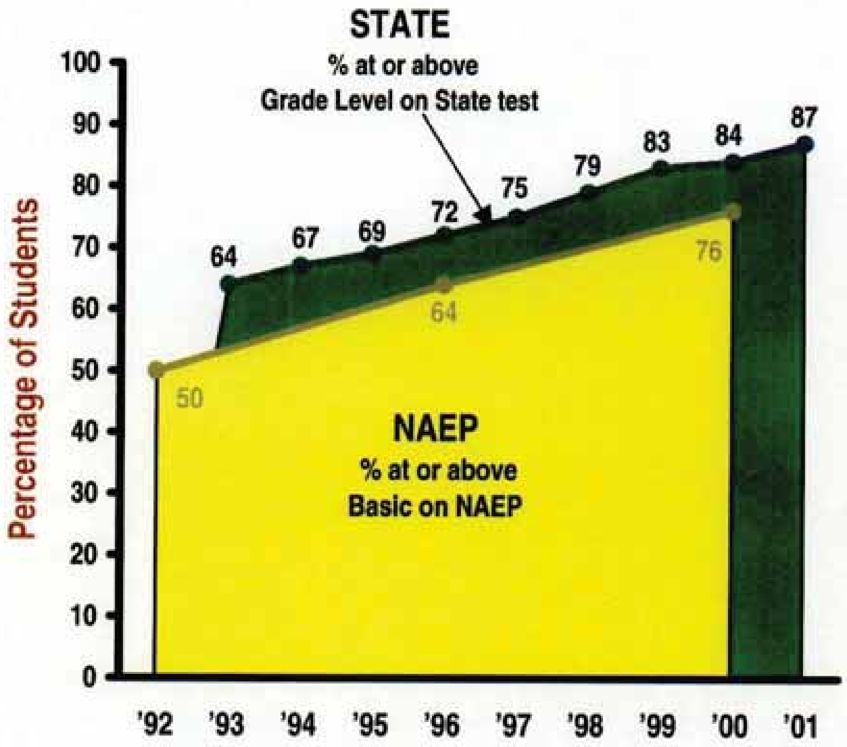
All students

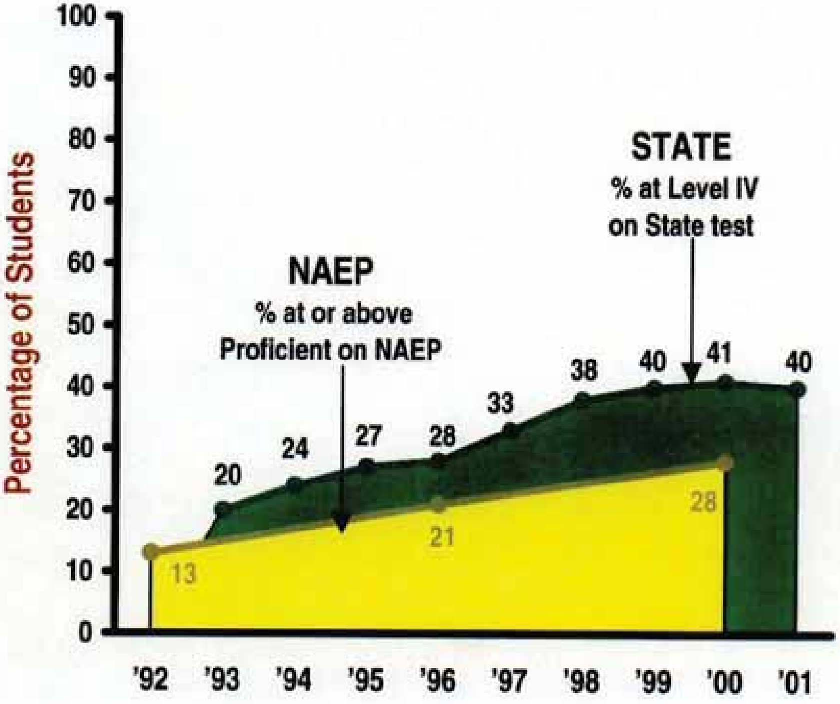


The term "Grade Level" is equivalent to the aggregation of Levels III and IV.

Figure 2. State A -- Mathematics, Grade 4

All students

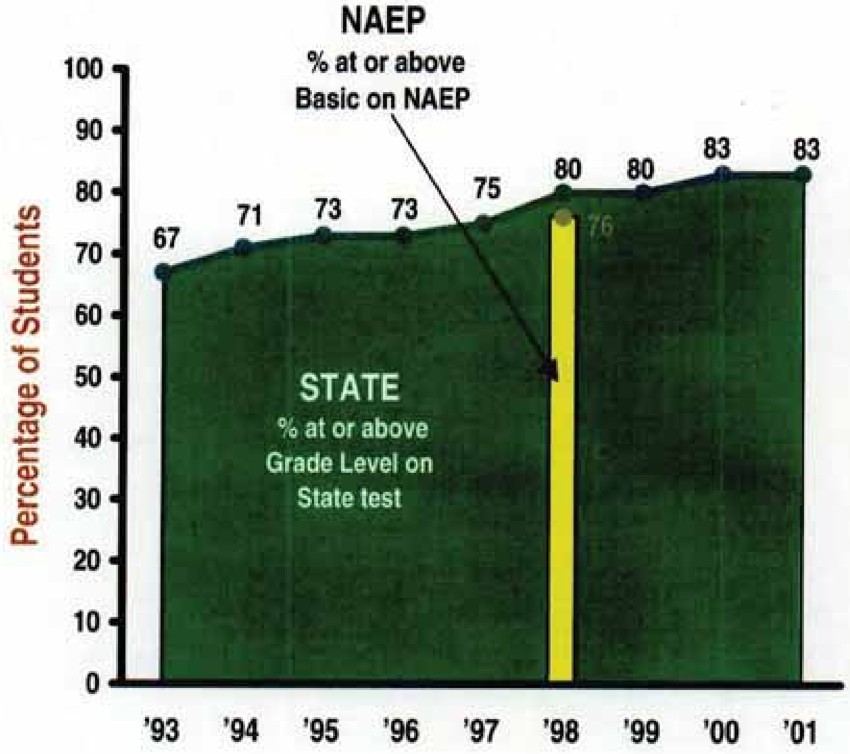


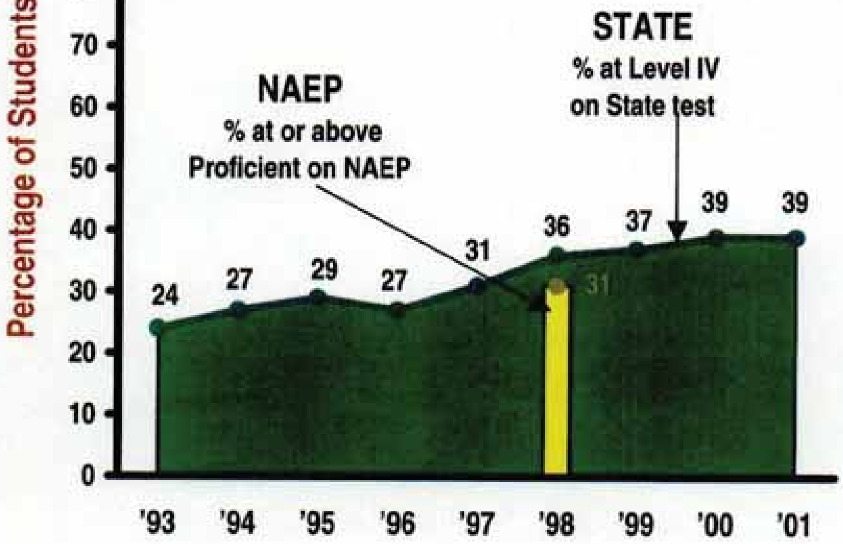


The term "Grade Level" is equivalent to the aggregation of Levels III and IV.

Figure 3. State A -- Reading, Grade 8

All students

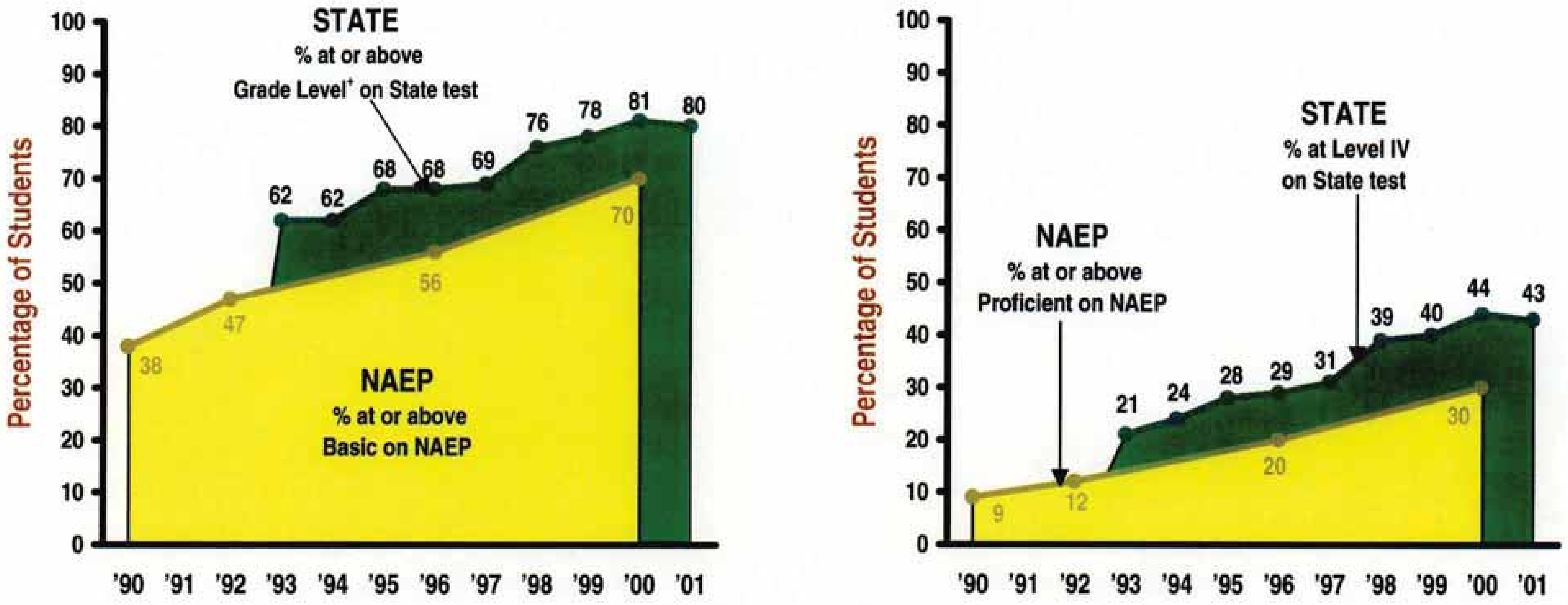




The term "Grade Level" is equivalent to the aggregation of Levels III and IV.

Figure 4. State A -- Mathematics, Grade 8

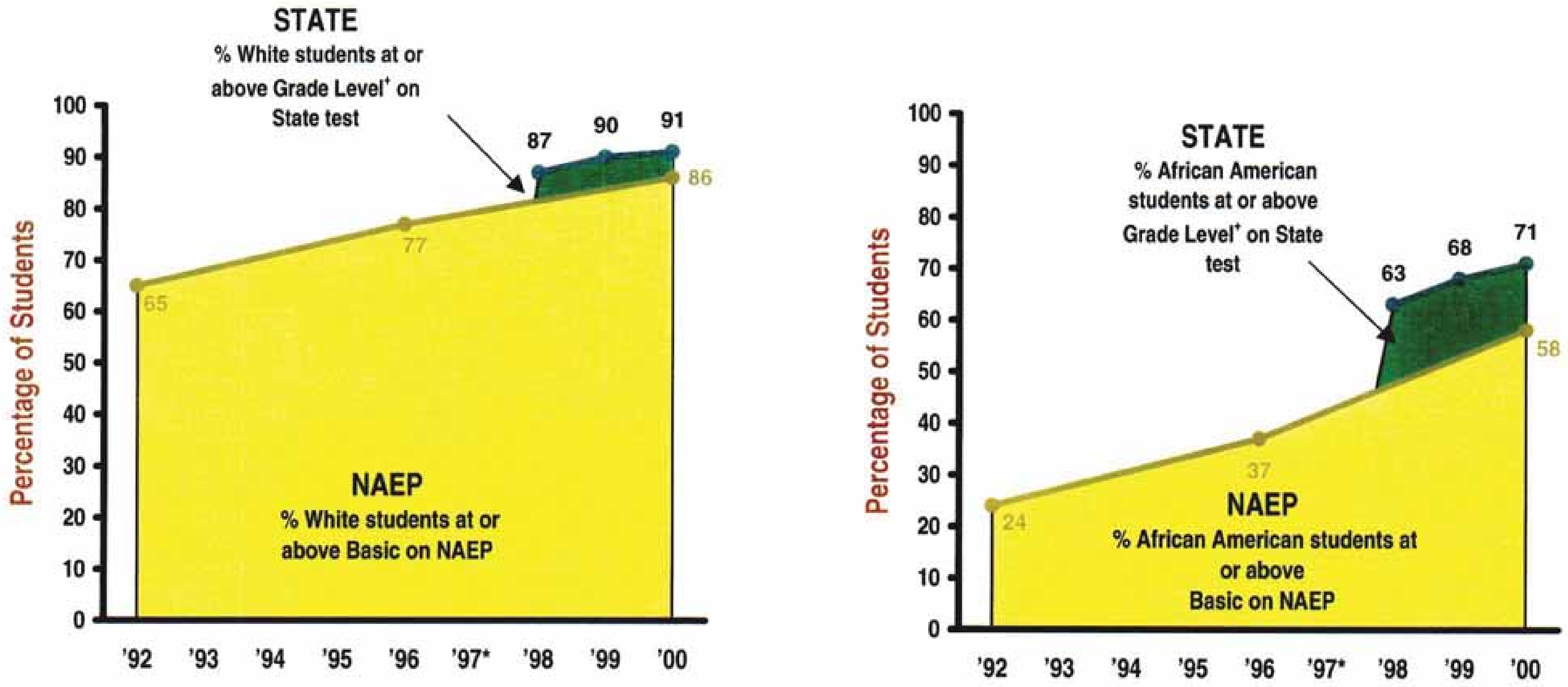
All students



+The term "Grade Level" is equivalent to the aggregation of Levels III and IV.

Figure 5. State A -- Mathematics, Grade 4

White and African American students



STATE
% White students at or above Grade Level⁺ on State test

NAEP
% White students at or above Basic on NAEP

STATE
% African American students at or above Grade Level⁺ on State test

NAEP
% African American students at or above Basic on NAEP

Figure 6. State A -- Mathematics, Grade 4

White and Hispanic students

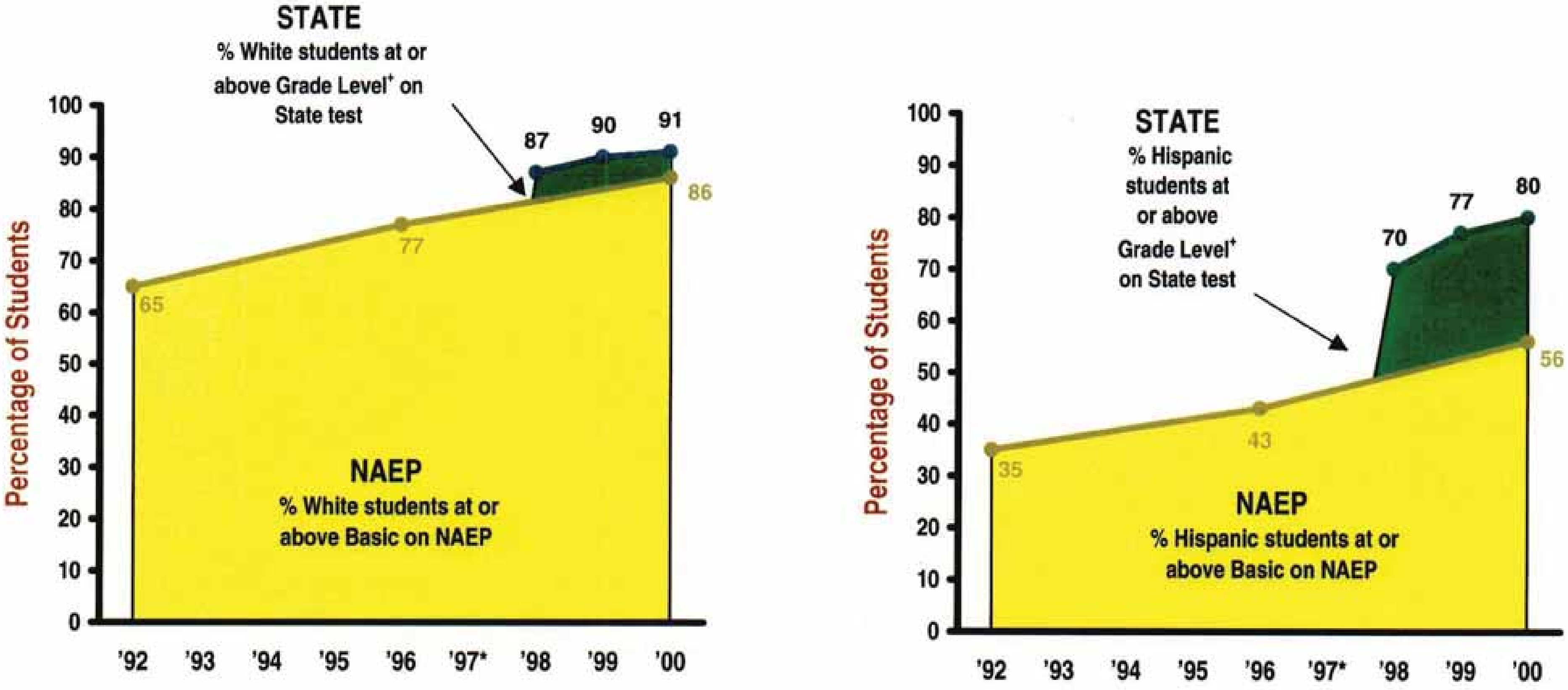
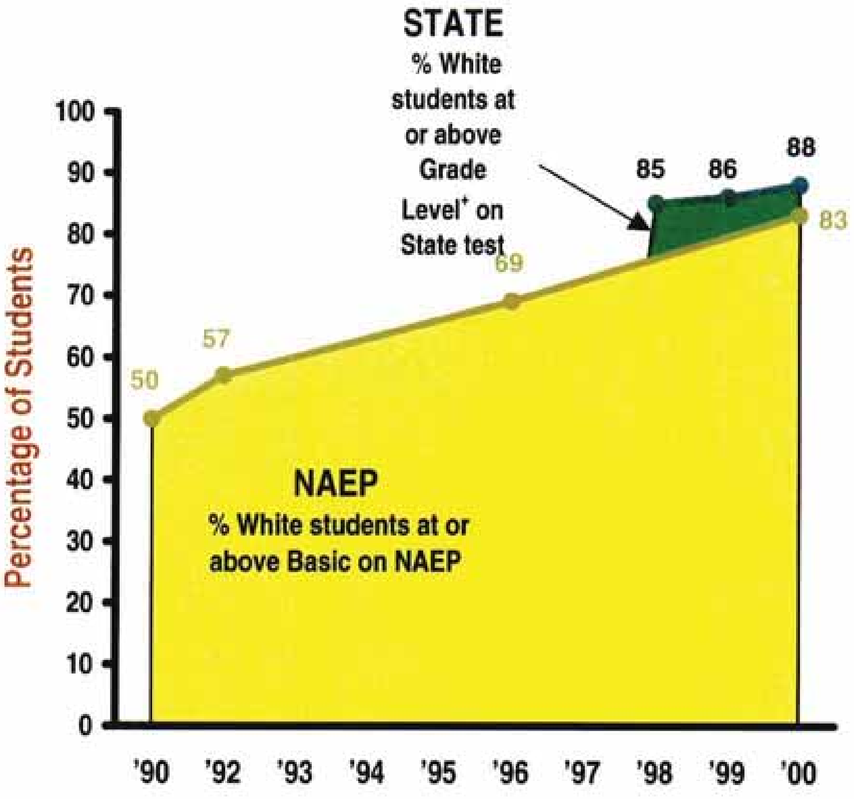


Figure 7. State A -- Mathematics, Grade 8

White and African American students



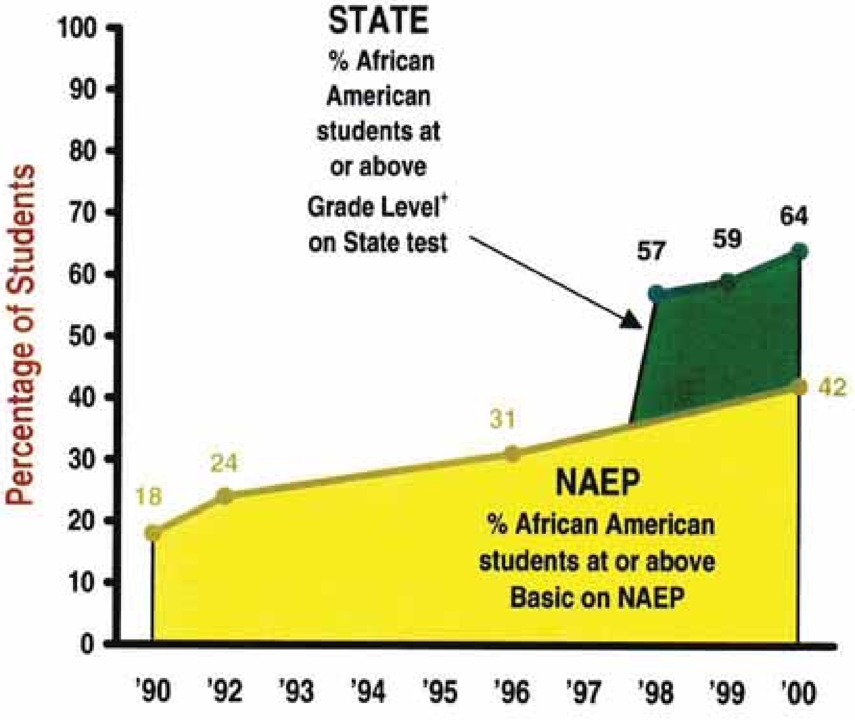
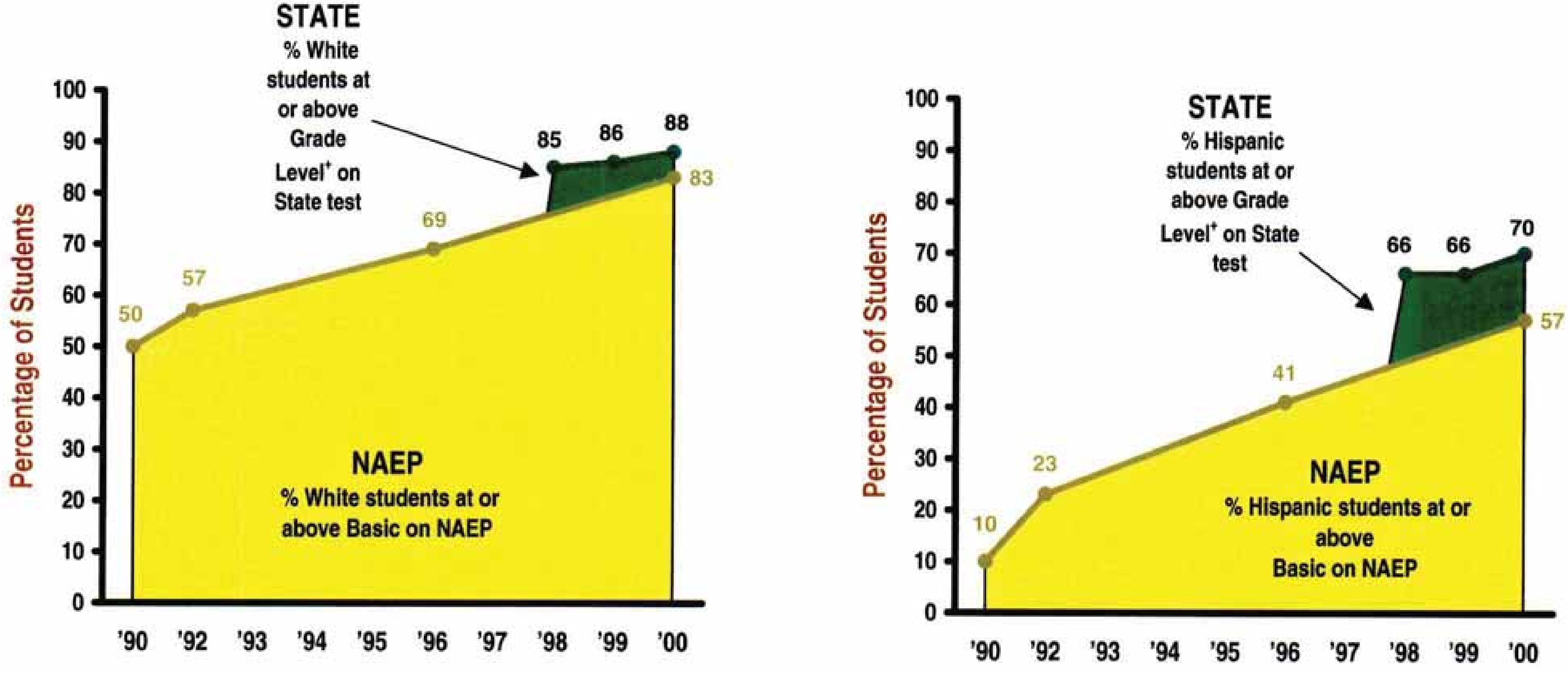


Figure 8. State A -- Mathematics, Grade 8

White and Hispanic students



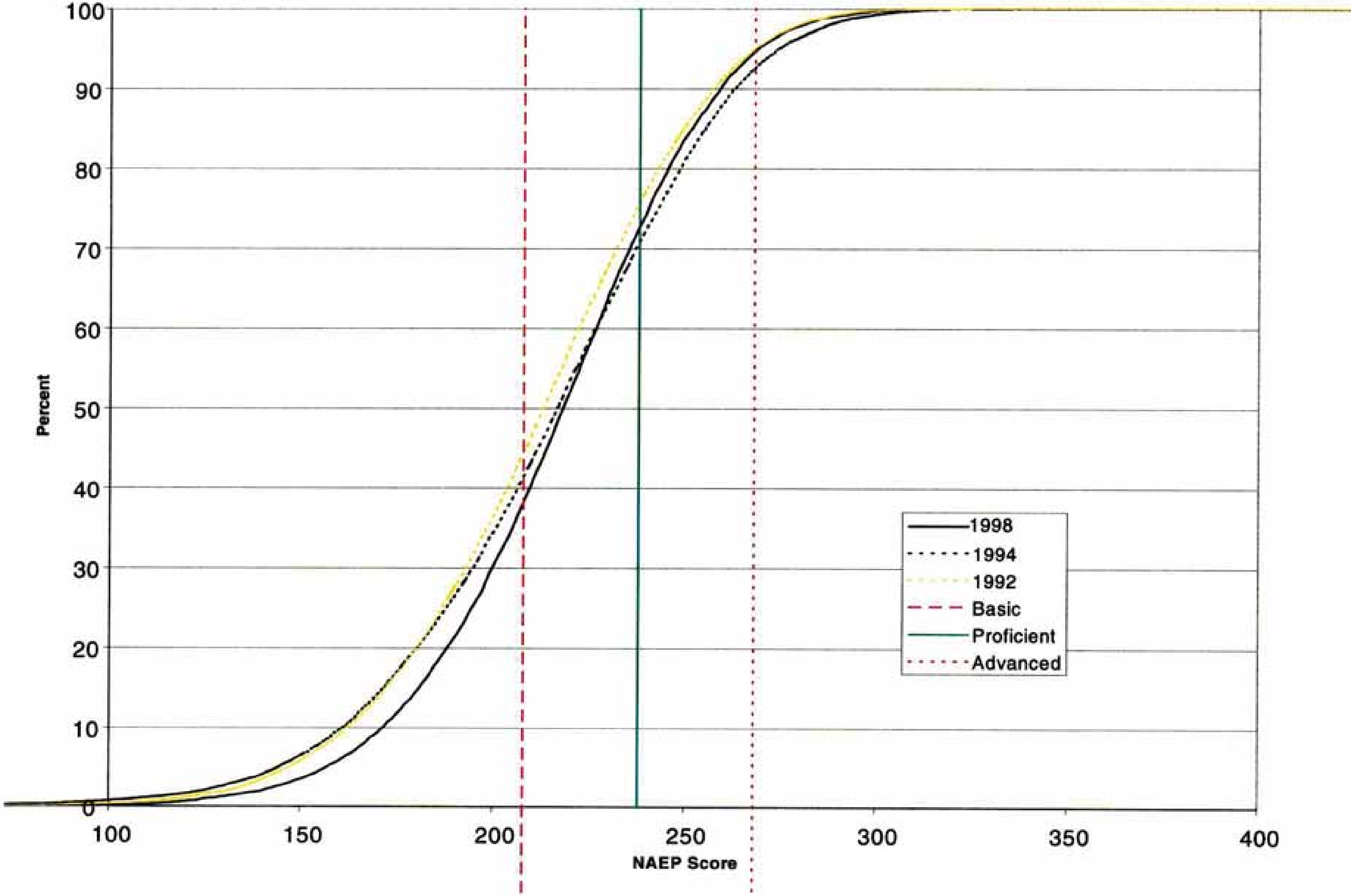
STATE
% White students at or above Grade Level⁺ on State test

NAEP
% White students at or above Basic on NAEP

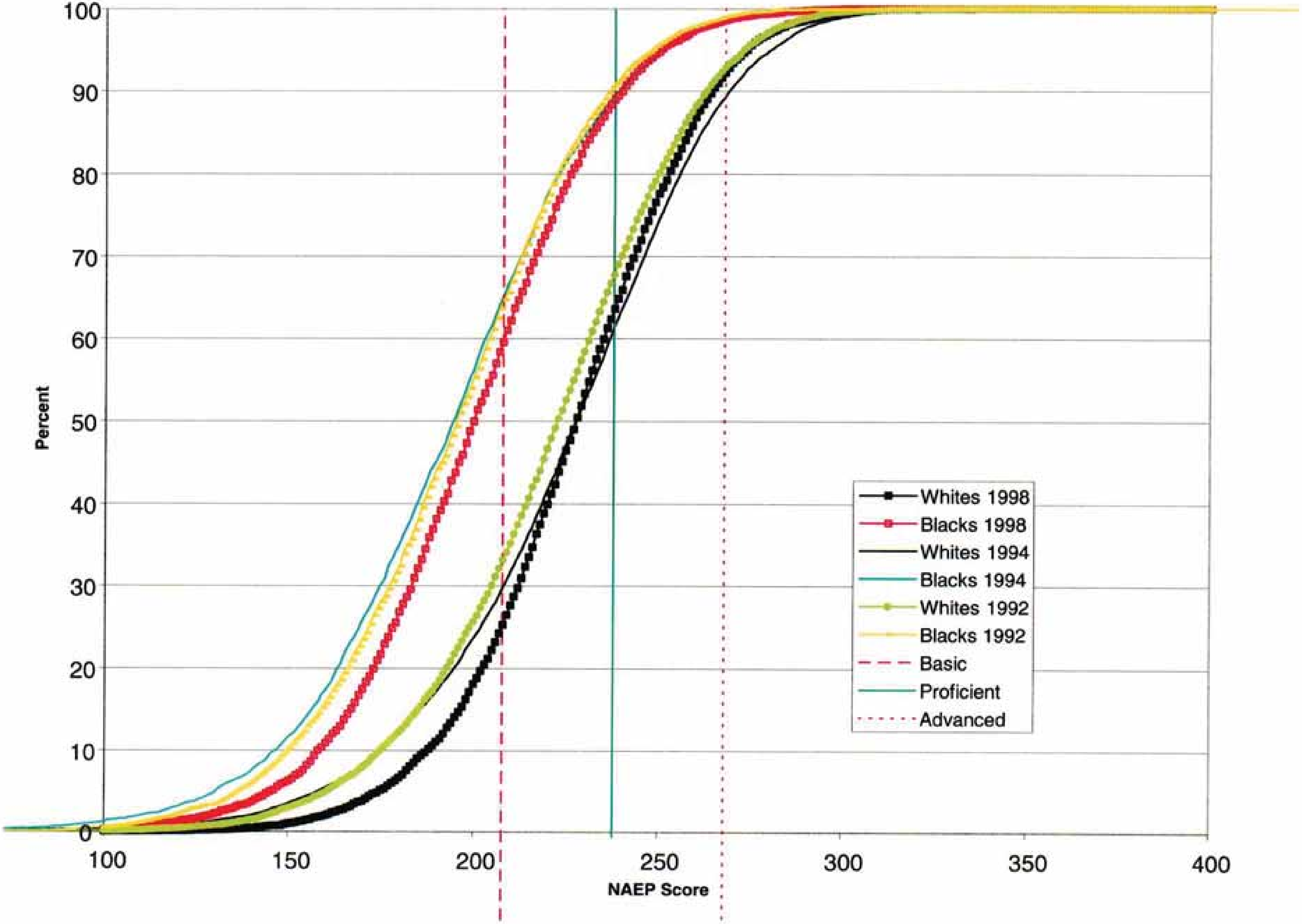
STATE
% Hispanic students at or above Grade Level⁺ on State test

NAEP
% Hispanic students at or above Basic on NAEP

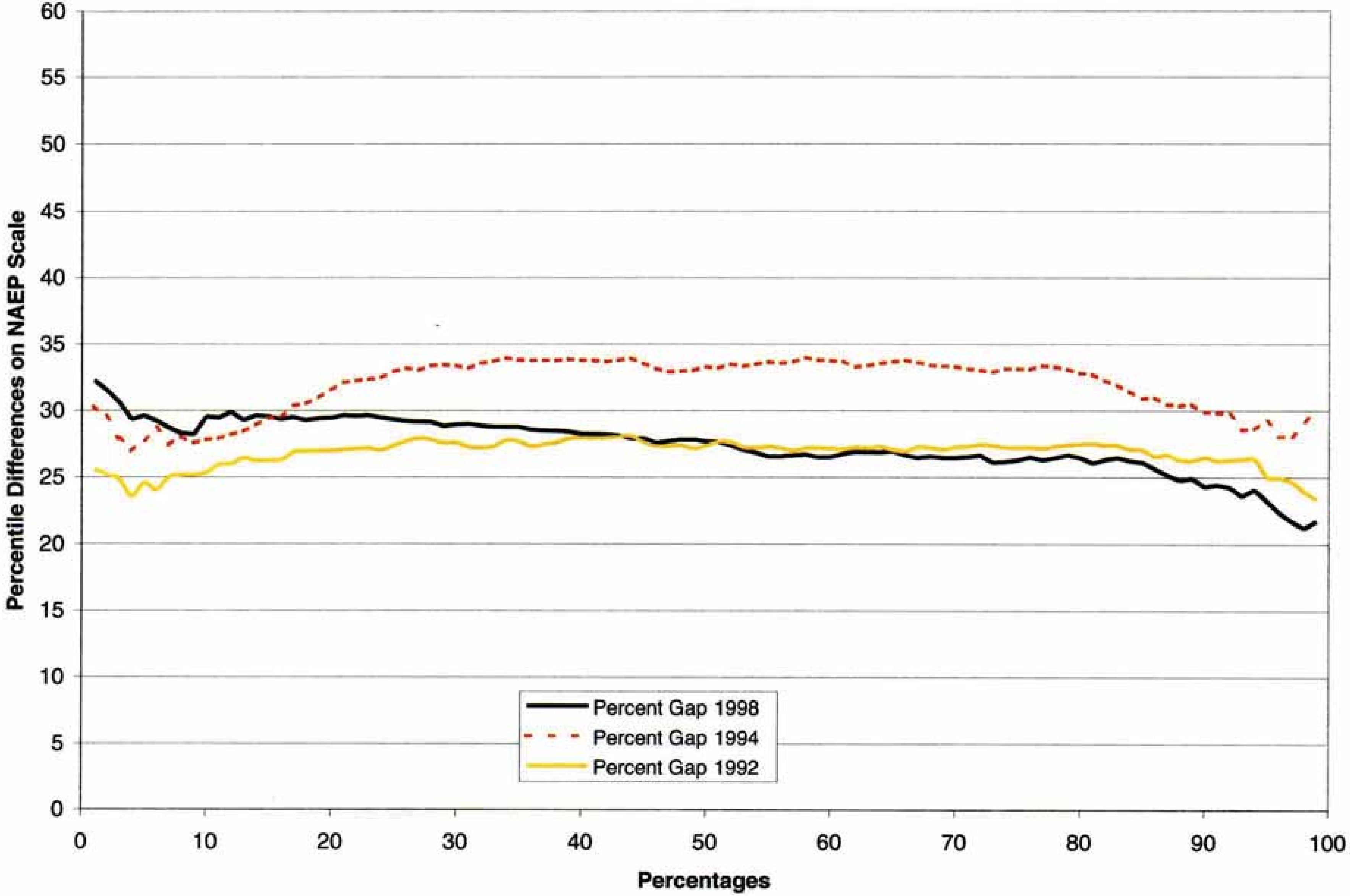
State A - Reading 1992, 1994 and 1998 Grade 4 Cumulative Distribution Function for Total Figure 9



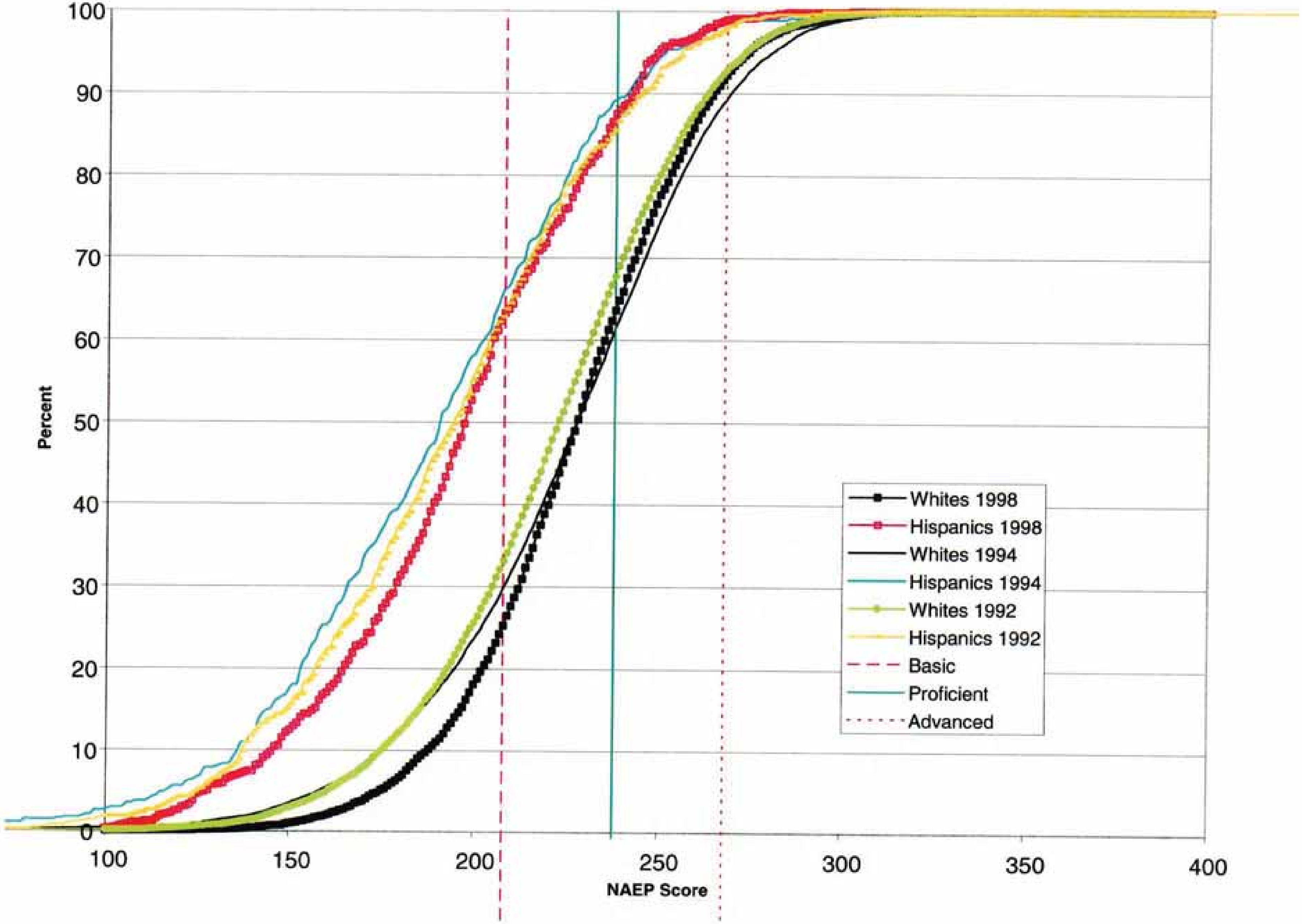
State A - Reading 1992, 1994 and 1998 Grade 4
Cumulative Distribution Function for Whites and Blacks
Figure 10



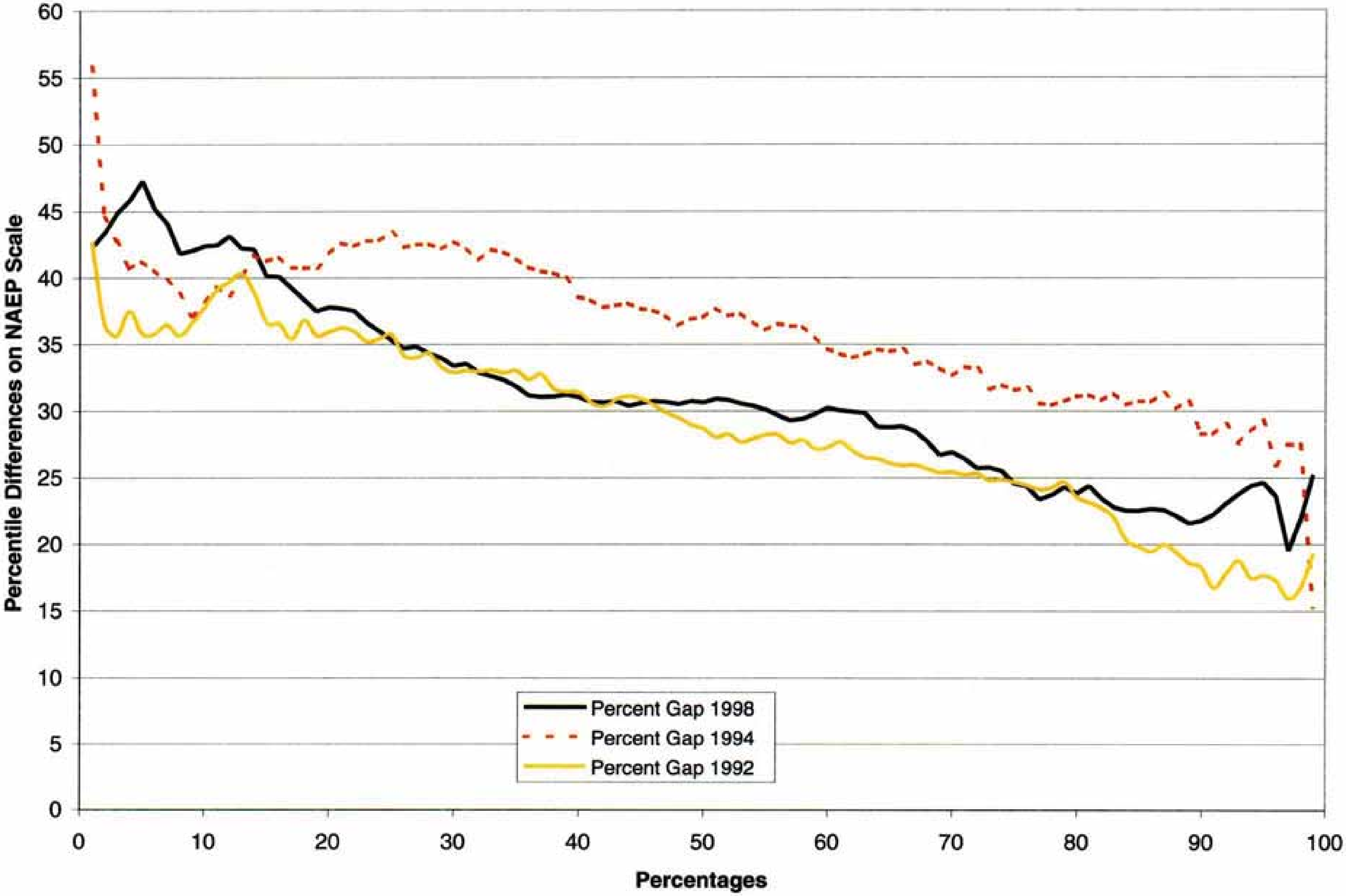
State A - Reading 1992, 1994 and 1998 Grade 4
Gap in Percents Between Whites and Blacks
Figure 11



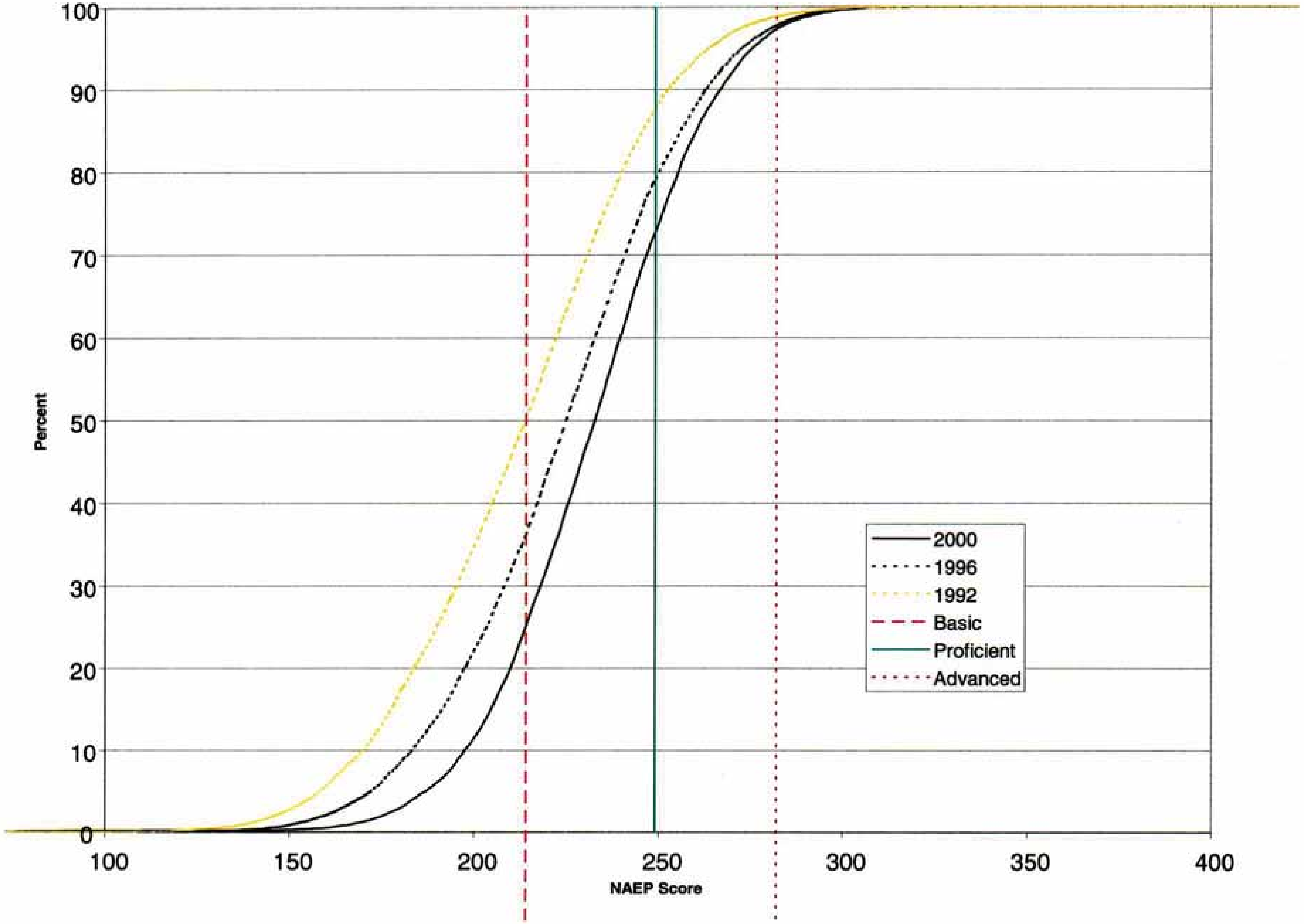
State A - Reading 1992, 1994 and 1998 Grade 4
Cumulative Distribution Function for Whites and Hispanics
Figure 12



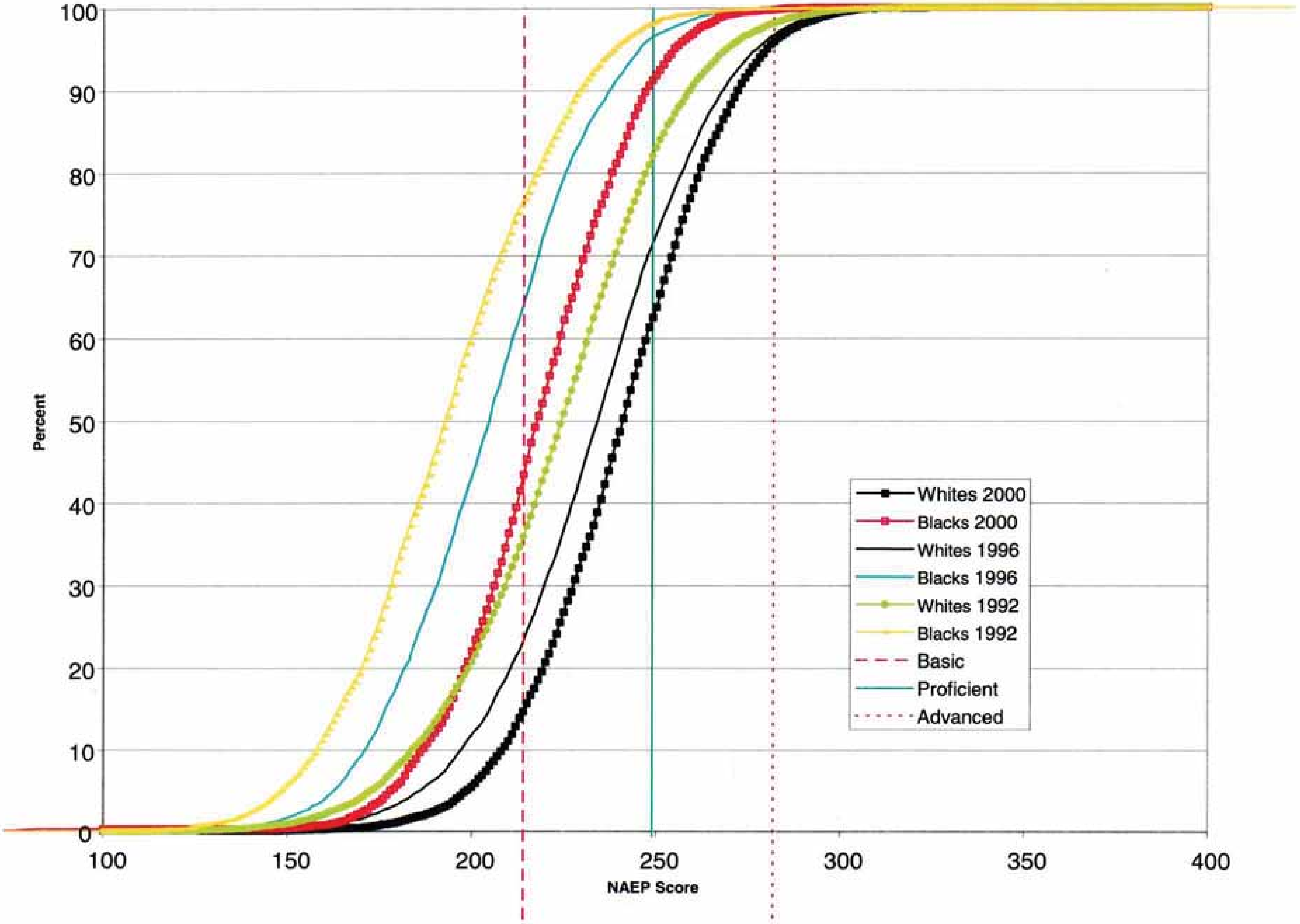
State A - Reading 1992, 1994 and 1998 Grade 4
Gap in Percents Between Whites and Hispanics
Figure 13



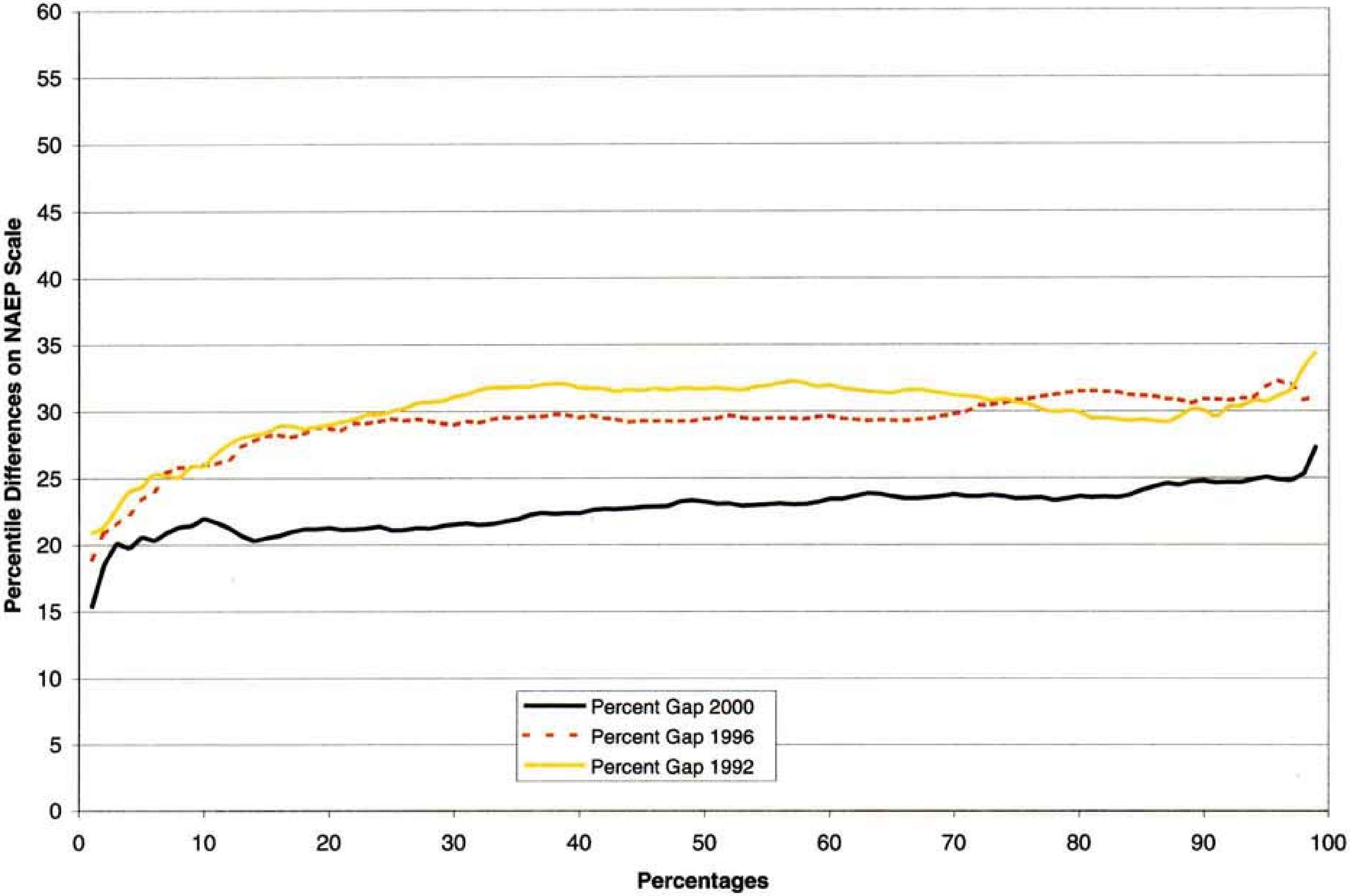
State A - Mathematics 1992, 1996 and 2000 Grade 4
Cumulative Distribution Function for Total
Figure 14



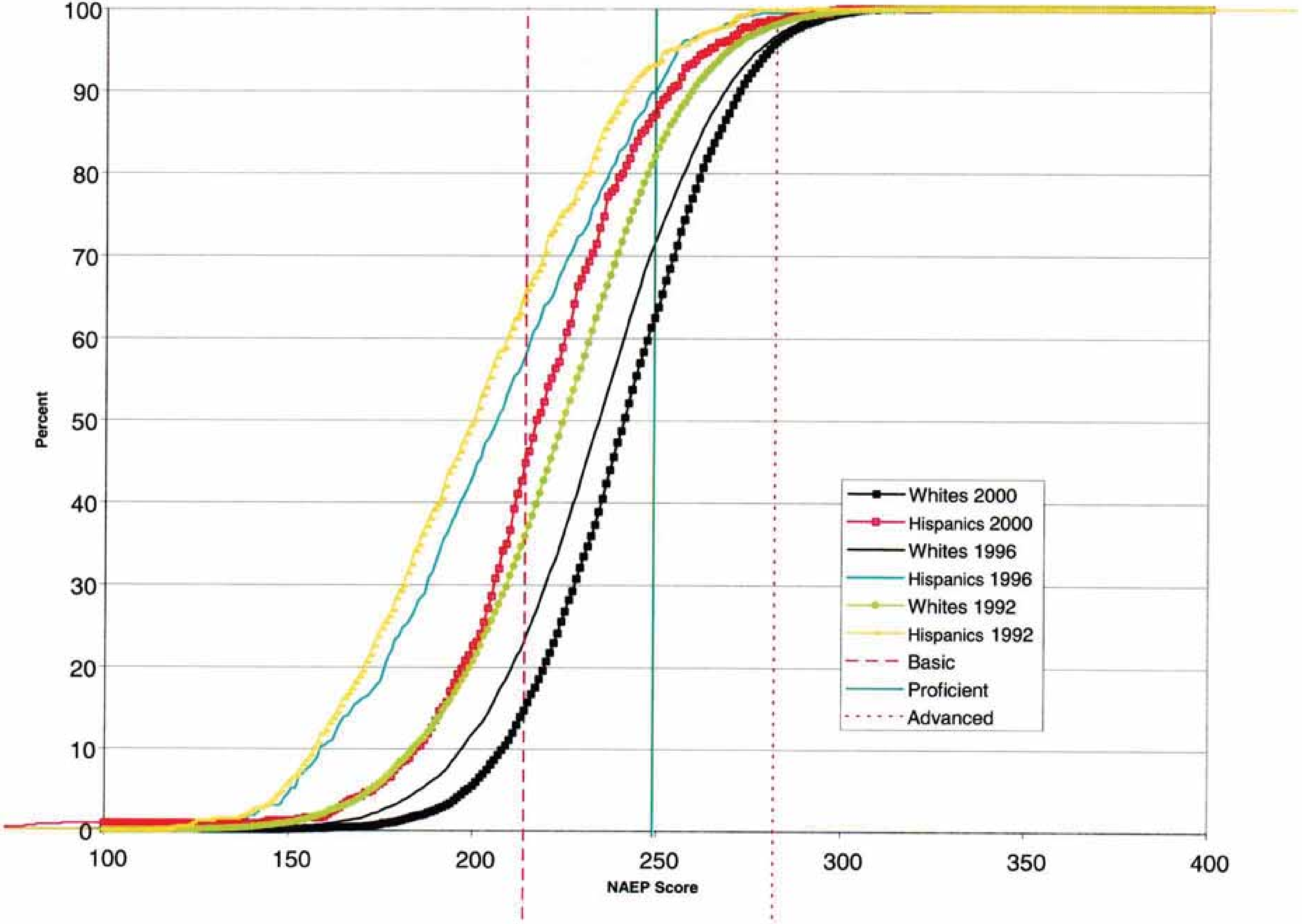
State A - Mathematics 1992, 1996 and 2000 Grade 4
Cumulative Distribution Function for Whites and Blacks
Figure 15



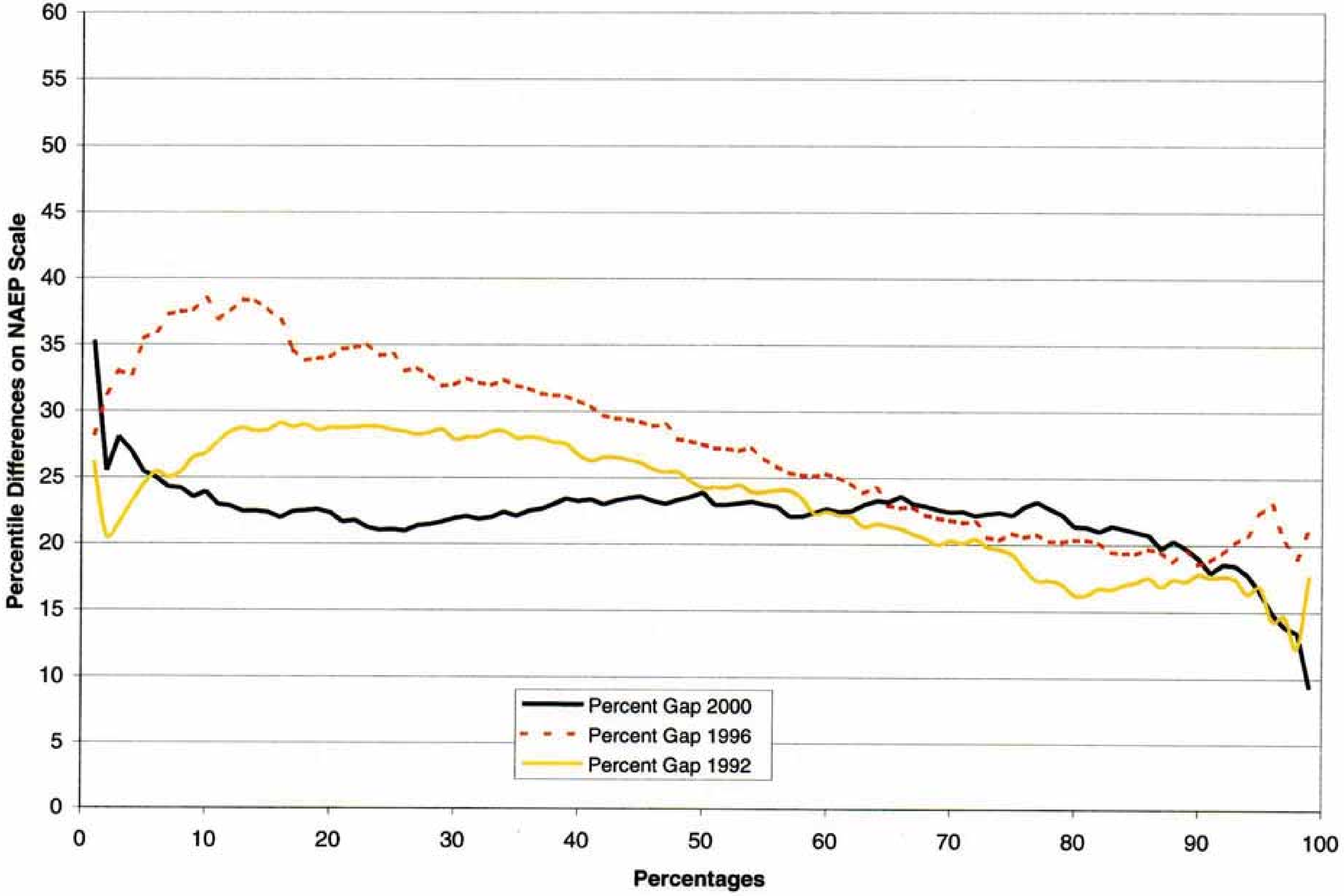
State A - Mathematics 1992, 1996 and 2000 Grade 4
Gap in Percents Between Whites and Blacks
Figure 16



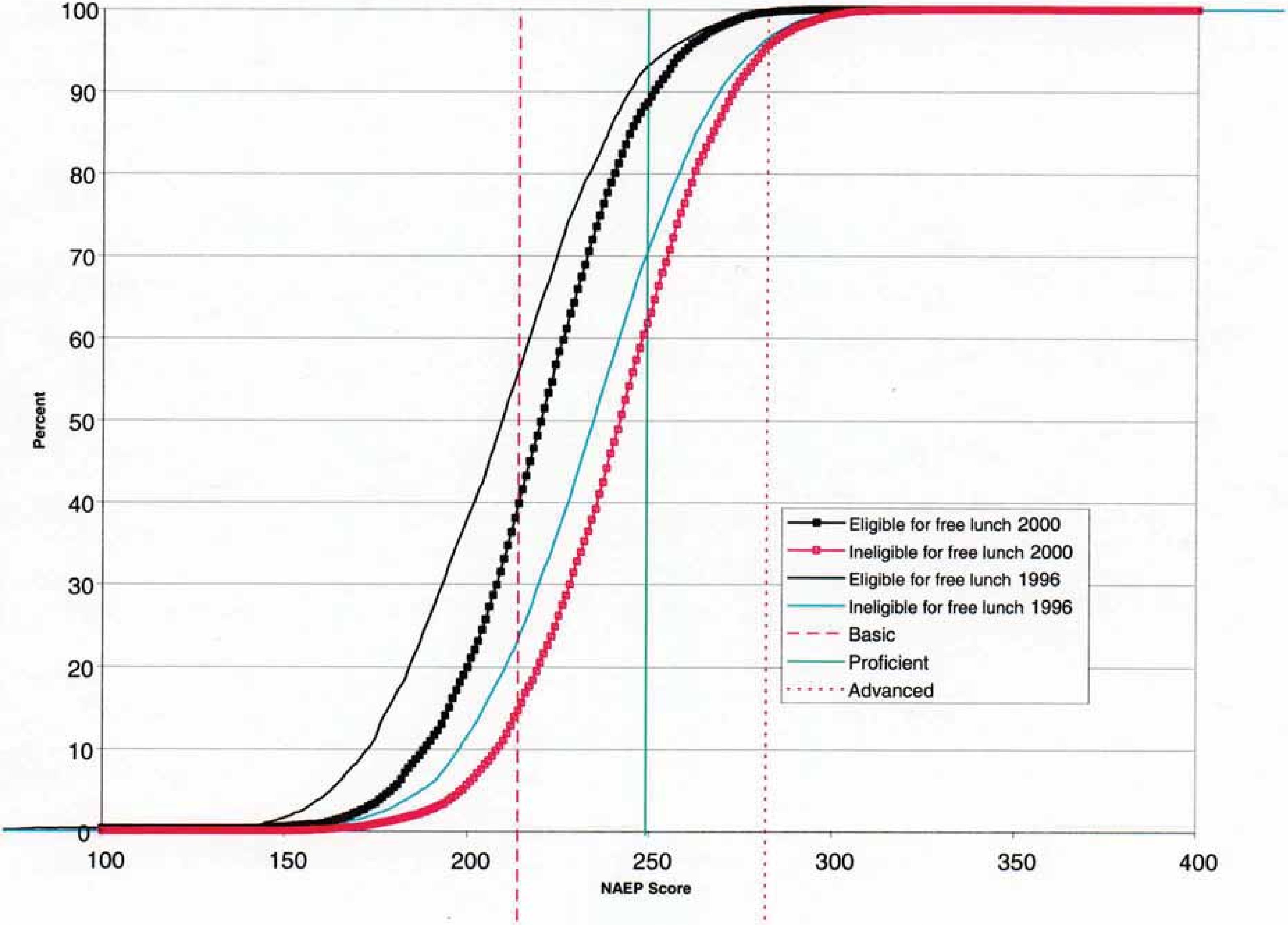
State A - Mathematics 1992, 1996 and 2000 Grade 4
Cumulative Distribution Function for Whites and Hispanics
Figure 17



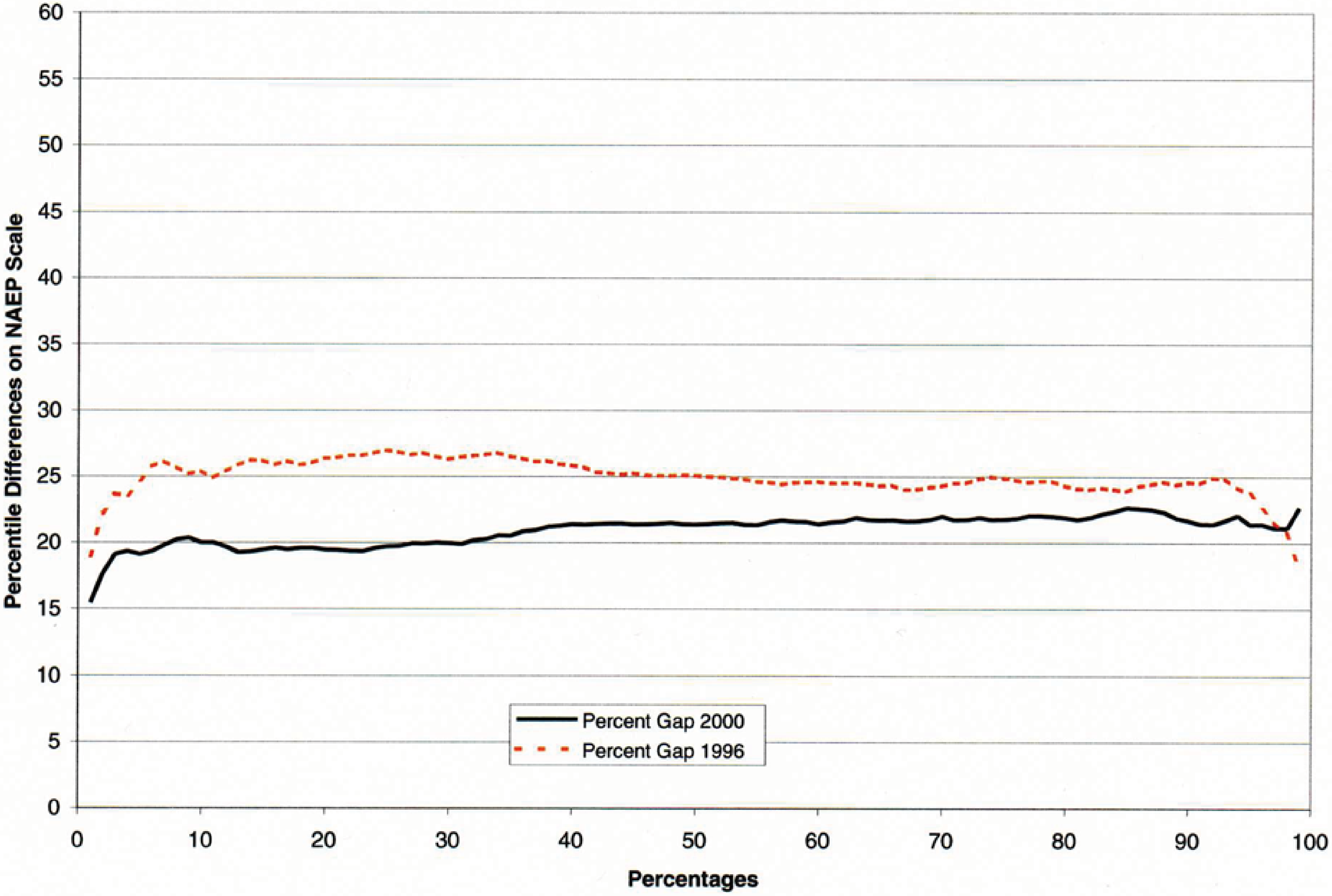
State A - Mathematics 1992, 1996 and 2000 Grade 4
Gap in Percents Between Whites and Hispanics
Figure 18



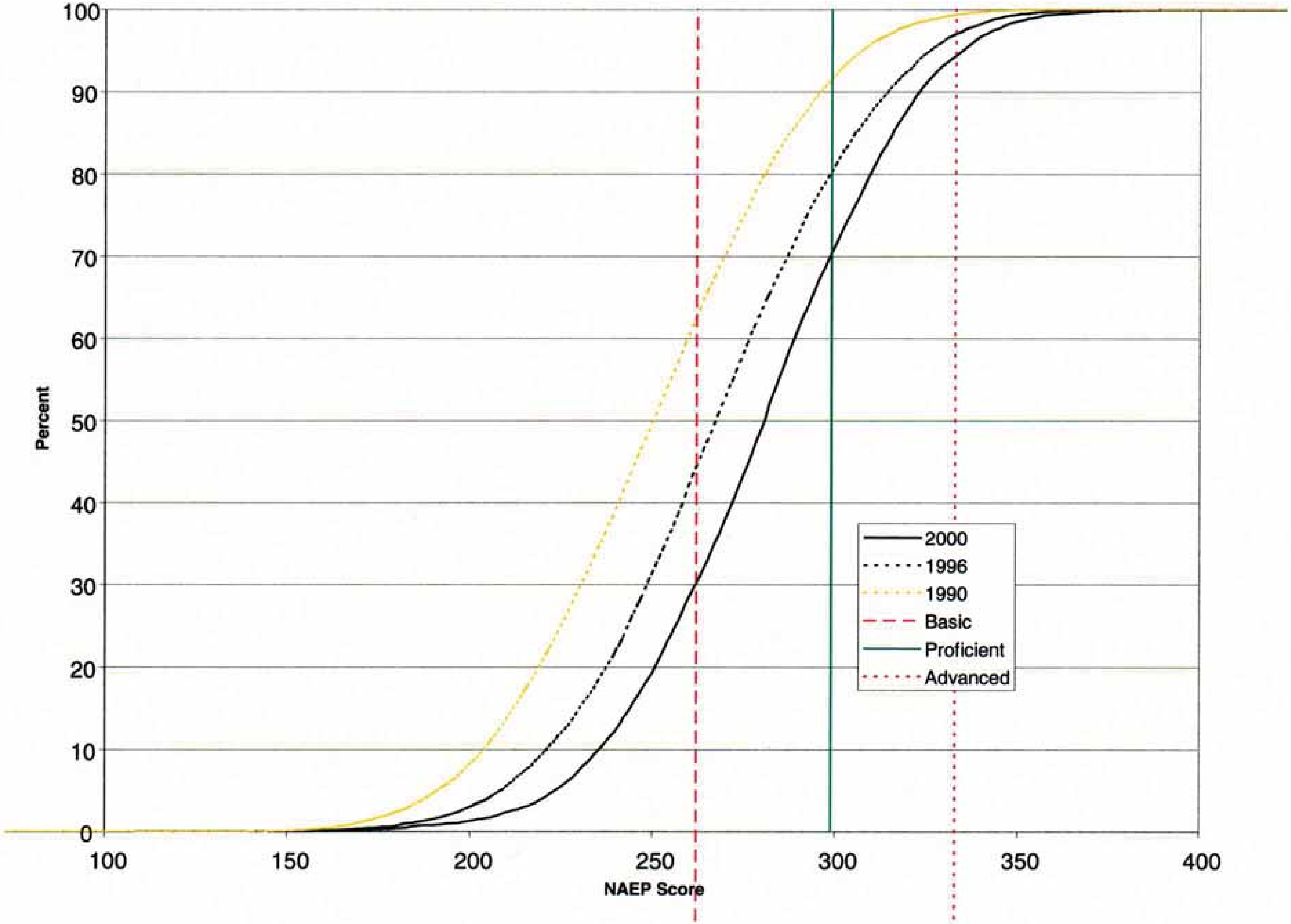
State A - Mathematics 1996 and 2000 Grade 4
Cumulative Distribution Function for Ineligible for free lunch and Eligible for free lunch
Figure 19



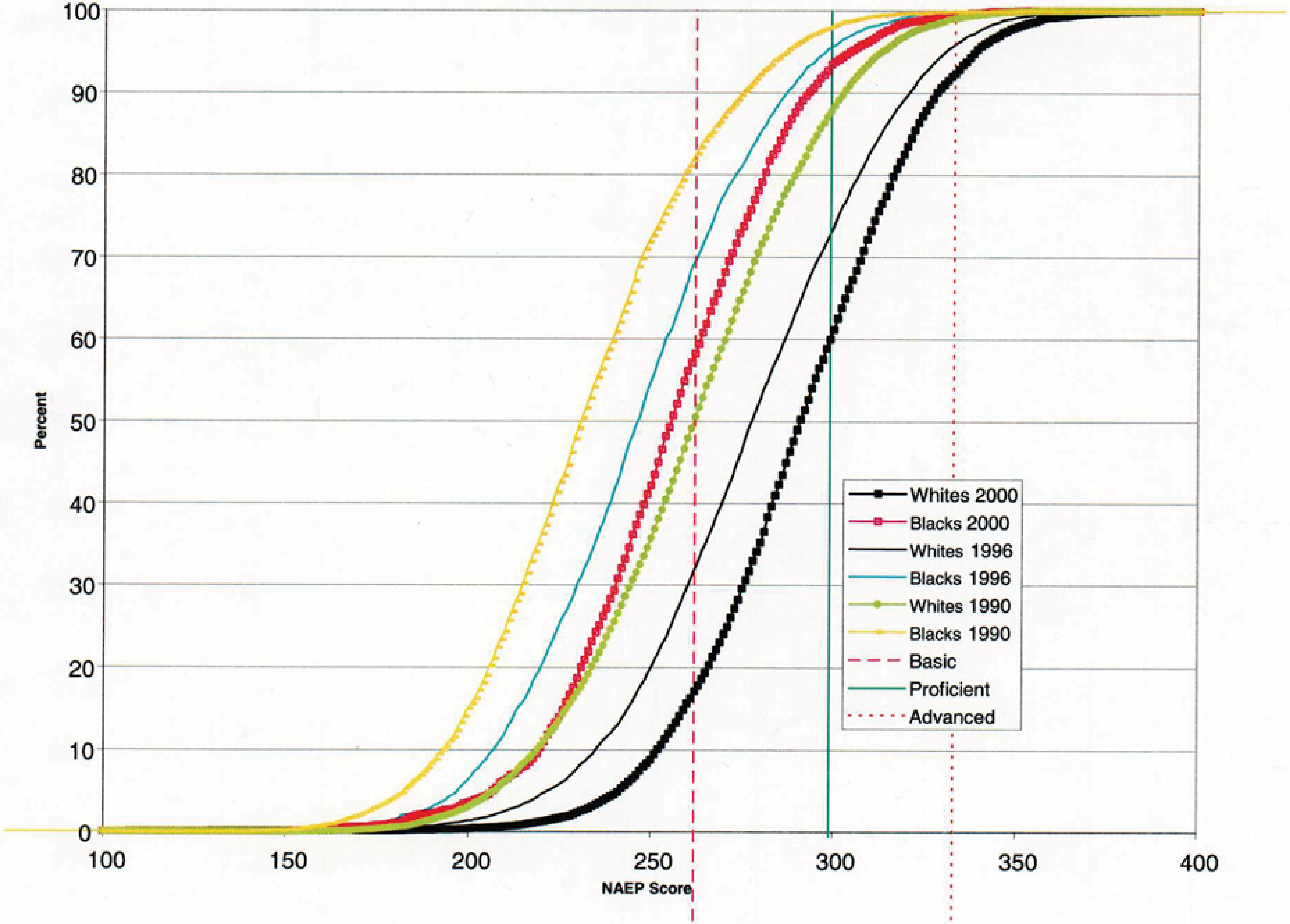
State A - Mathematics 1996 and 2000 Grade 4
Gap in Percents Between Ineligible for free lunch and Eligible for free lunch
Figure 20



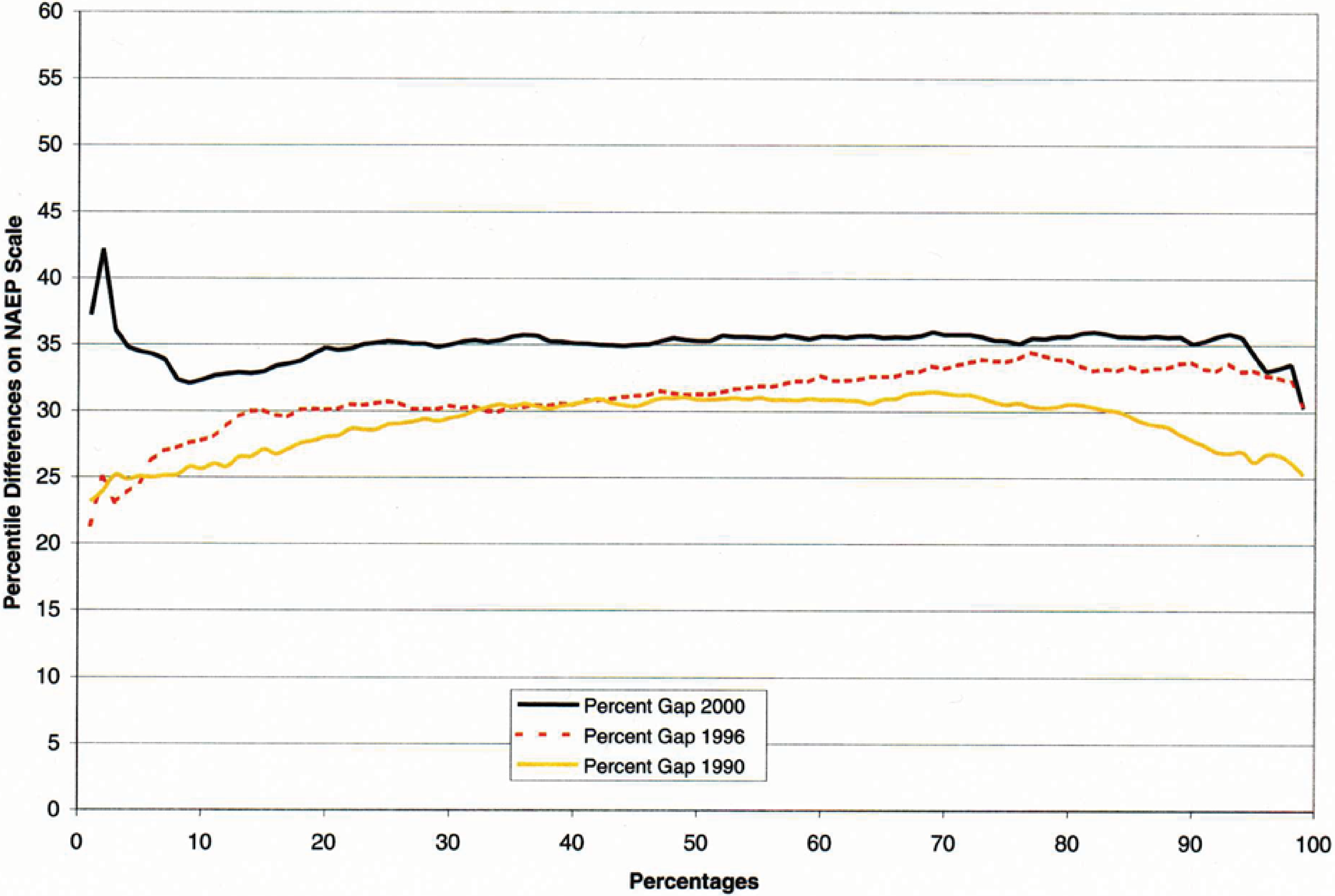
State A - Mathematics 1990, 1996 and 2000 Grade 8
Cumulative Distribution Function for Total
Figure 21



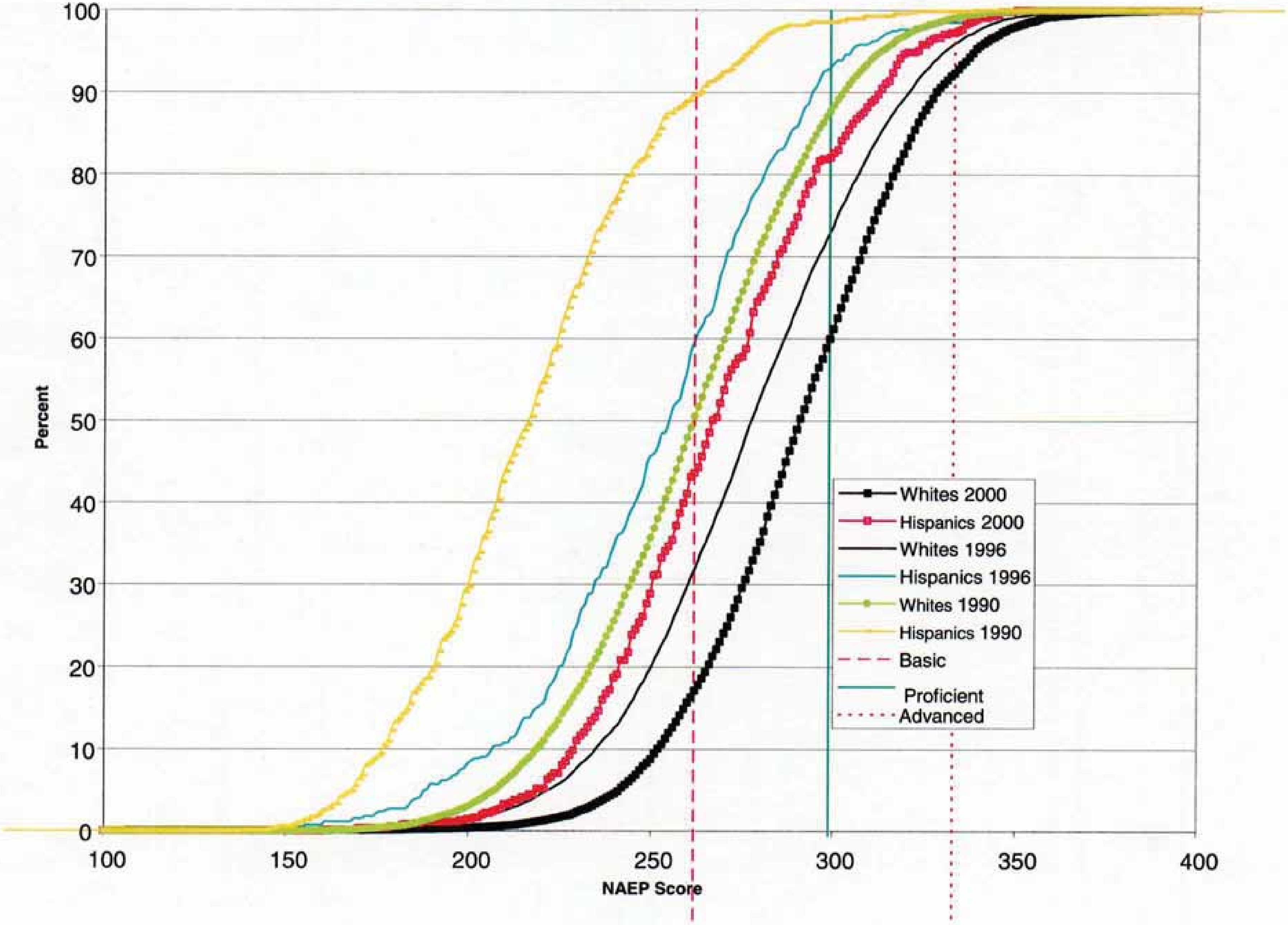
State A - Mathematics 1990, 1996 and 2000 Grade 8
Cumulative Distribution Function for Whites and Blacks
Figure 22



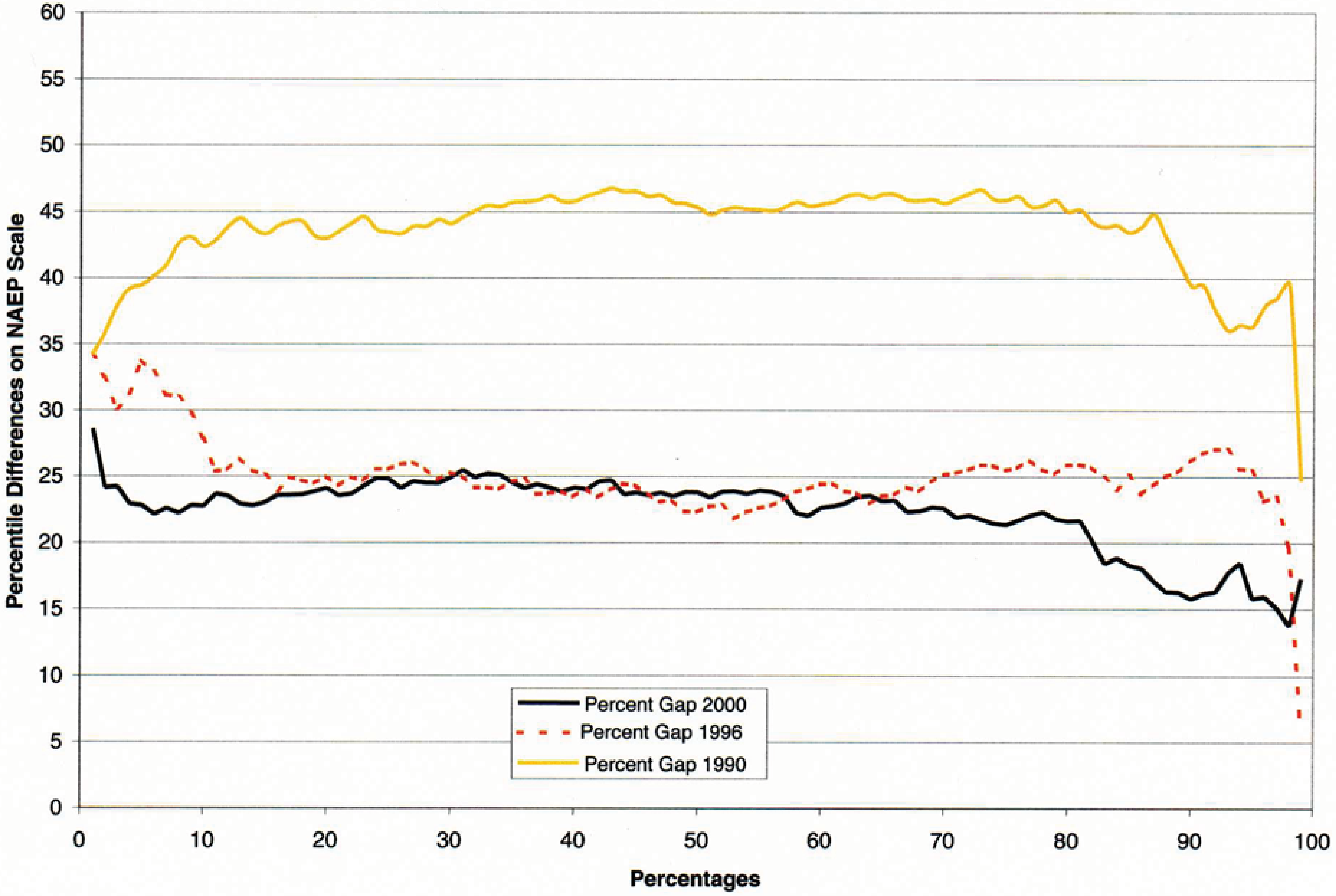
State A - Mathematics 1990, 1996 and 2000 Grade 8
Gap in Percents Between Whites and Blacks
Figure 23



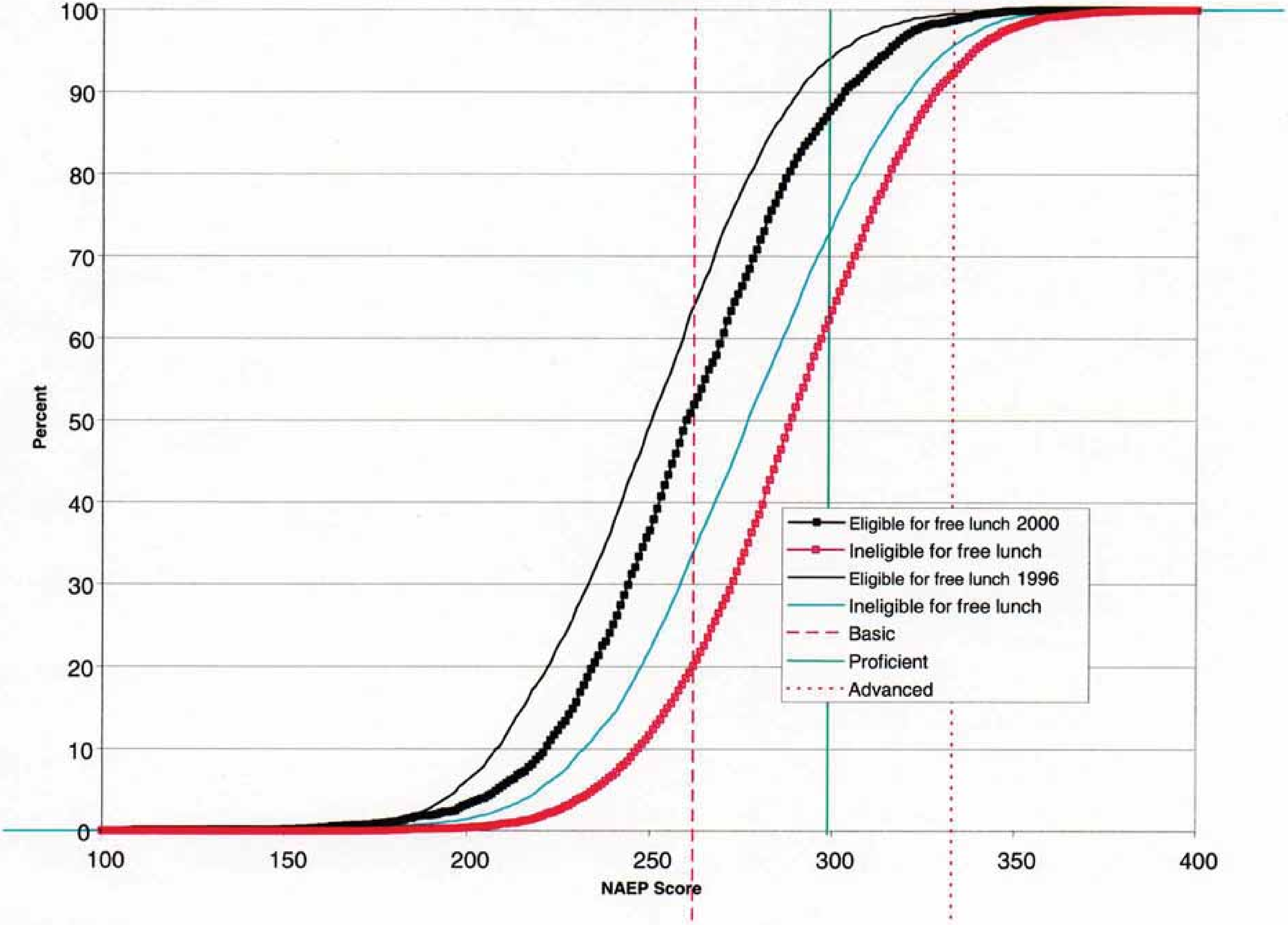
State A - Mathematics 1990, 1996 and 2000 Grade 8
Cumulative Distribution Function for Whites and Hispanics
Figure 24



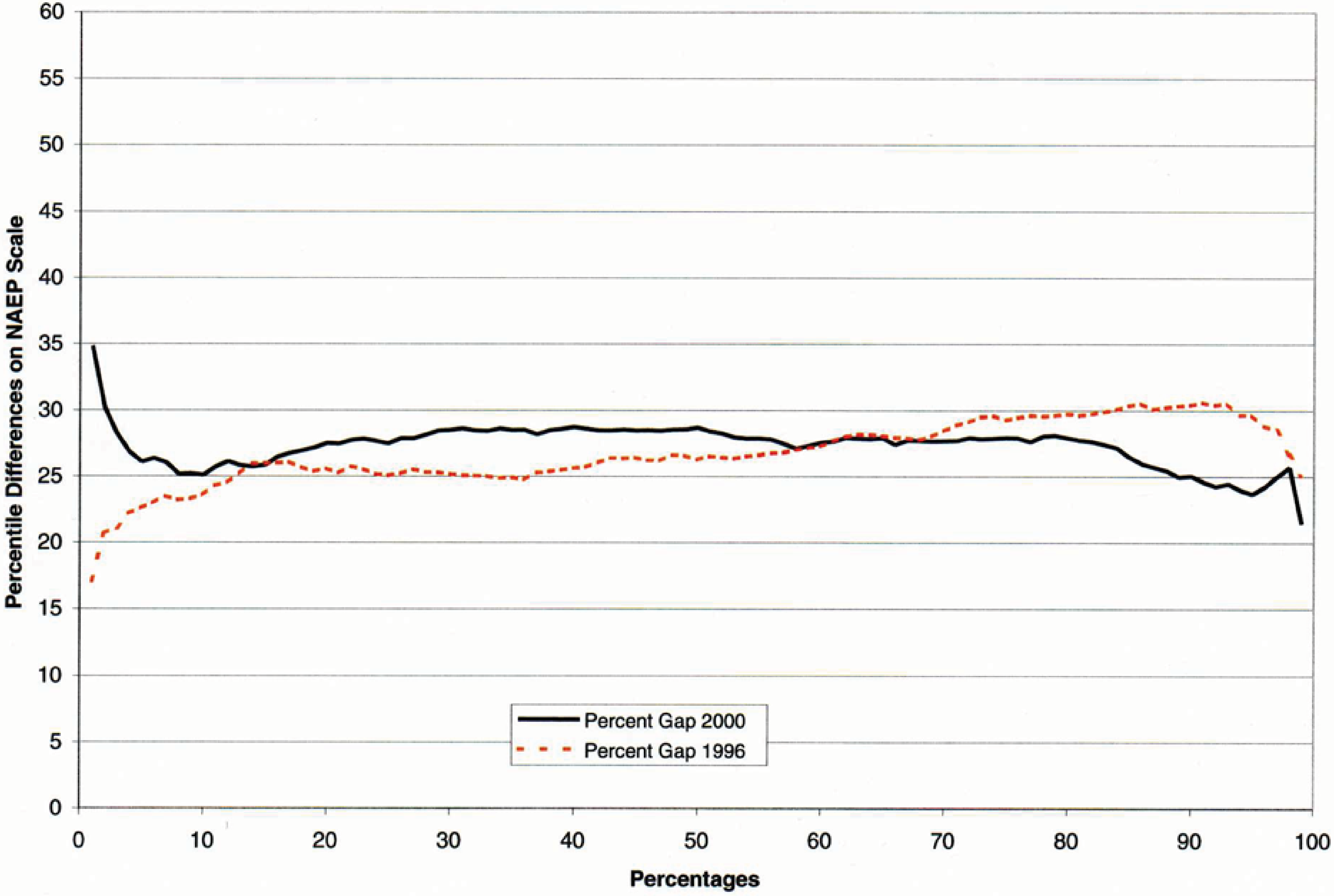
State A - Mathematics 1990, 1996 and 2000 Grade 8
Gap in Percents Between Whites and Hispanics
Figure 25



State A - Mathematics 1996 and 2000 Grade 8
Cumulative Distribution Function for Ineligible for free lunch and Eligible for free lunch
Figure 26



State A - Mathematics 1996 and 2000 Grade 8
Gap in Percents Between Ineligible for free lunch and Eligible for free lunch
Figure 27



Mark Reckase January 31, 2002

State B Argument for Meeting the Requirements of the ESEA Legislation

The December 12, 2001 draft of the ESEA legislation contains a number of requirements regarding performance standards and academic testing that states must meet if they are to obtain federal funding under the legislation. These requirements are used as a basis for demonstrating the kinds of evidence that can be provided as documentation that a state has meet the requirements of the law. A specific state was selected for the demonstration. That state is labeled State B in this document because the state has not had an opportunity to review this argument for accuracy. Therefore, the information provided in this document should only be considered as an example of what could be done to meet the requirements of the legislation rather than an accurate representation of the performance of a state. For State B, NAEP data is also considered to determine if State B results and the NAEP results are consistent with each other.

The requirements of the draft ESEA legislation are as follows:

1. The state must have challenging academic content standards.
2. The state must have challenging student academic achievement standards.
3. The standards apply to public elementary and secondary school children.
4. The standards exist for mathematics, reading or language arts, and science (beginning in 2005-2006).
5. The achievement standards should be aligned to the content standards.
6. The achievement standards should have two levels of high achievement, proficient and advanced. The proficient standard is the critical one that appears later in the legislation.
7. A basic level is included to help track the performance of lower performing students.
8. The state should have an accountability system that provides a definition for adequate yearly progress.
   a. The definition should be statistically valid and reliable.
   b. The definition should result in continuous and substantial academic improvements for all students.
9. States should work toward narrowing achievement gaps.
10. There should be separate measurable annual objectives for the following groups:
    a. Economically disadvantaged.
    b. Major racial and ethnic groups.
    c. Disabled.
    d. Limited English proficient.

11. The starting point for the process is the 2001-2002 school year. For that year, the base level is the percentage of students meeting or exceeding the state's proficient level for achievement. This level is defined as the higher of the following two options:
    a. The lowest achieving group,
    b. The school at the 20th percentile ranked by the percentage of students at the proficient level.
12. The state will have 12 years to bring all student performance above the proficient level. This seems to imply 12 years of education to reach the proficient 12th grade standard.
13. There should be annual measurable objectives for mathematics and reading.
14. Each target group should have its own minimum.
15. There should be intermediate goals that include equal increments over the 12 years.
16. There is a 10% cushion for meeting the standard. A state may miss the target by 10% in a given year.
17. At least 95% of each group must take the assessment.
18. The state can average up to three years of data when reporting results.
19. Before 2005, students must be assessed once in 3 – 5, 6 – 9, and 10 – 12.
20. Starting in 2005, students must be assessed at grades 3 through 8, inclusive.

State B was selected from a set of possible states for use in this demonstration because it already met many of the legislative requirements and because it had an approved assessment plan under previous ESEA legislation. State B also had some features that would help highlight challenges to state assessment/NAEP comparisons. State B makes heavy use of performance assessments in its testing program, reports results according to standards, uses a rolling average for reporting results, and uses a variety of subscores rather than a single total score when reporting student performance. Further, results are reported for a variety of subgroups. The State B assessment program and reporting system is summarized below.

1. State B has developed extensive academic content standards and student performance standards. While the State has produced documents that describe content standards including the State B State Frameworks for English language arts and mathematics, each school district selects or adapts from among a number of options. These include, the content standards provided by the National Council of Teachers of English, the National Council of Teachers of Mathematics, the New Standards for English and mathematics, and content standards from the State B Skills Commission. However, the assessments are aligned with The State B Standards as described in the state frameworks, and the standards provided by the X Standards Program, a national program that defined performance standards. The use of these many sets of national standards in the production of the state standards would allow the state to argue that their content standards meet the requirement to be challenging.

2. State B has adopted the performance standards from the X Standards Program for language arts and mathematics, and has developed their own standards for their writing assessment. For this report, only the reading and mathematics standards are discussed because only those are required when the legislation was adopted. The standards are set for grades 4, 8, and 10 as required by the legislation for the period of time before 2005.

   The labels for the performance standards from the X Standards Program are: achieved the standards with honors, achieved the standard, nearly achieved the standard, below the standard, and little evidence of achievement. The state set their performance standards at the levels suggested by the X Standards Program. That is, if students were either at the level of achieved the standards, or achieved the standards with honors, they were considered to have met the state performance standards. The "achieved the standard" level is equivalent to "proficient" and the "achieved the standard with honors" level is equivalent to "advanced". There are two levels below the proficient level to allow showing progress for students who have not met the state's performance standard.

   Performance standards were set for three content areas within mathematics: skills, concepts, and problem solving. Performance standards were set for two content areas within reading: basic understanding, and analysis and interpretation. Performance of subgroups of the population of students was compared using the average percentage exceeding the performance standards for each of the content area subscales on the assessment.

3. Each school district sets the level of adequate yearly progress for their own students following non-mandatory guidelines established by the Department of Education. These targets indicate the percentage of students who will move from below achieving the standard to achieving the standard or higher, and the percentage that will move out of the lowest level of achievement. The targets are to be achieved over a three year period. The state uses three year rolling averages to determine if the school has achieved its goals.

4. Achievement gaps are reported for the following subgroups: poverty/non-poverty; Asian-Pacific Islander, Black, Hispanic, Native American, White, Multi-racial; male/female; and special education, LEP, general education. The results are reported using a three year rolling average using total performance on the state assessment in each curriculum area.

The State B Results

The State B state assessment system uses the Criterion Referenced Examinations as a major component of its state assessment system. These examinations are augmented by custom tests to complete the coverage of the state curriculum frameworks. The components of the system that are relevant to the ESEA legislation are the Criterion Referenced Examinations in reading and mathematics. The standards set by the X Standards Program are used as the state standards on the examinations.

Reading

Results for the Criterion Referenced Examination in reading are reported for two different curriculum areas: Basic Understanding and Analysis and Interpretation. The results are reported according to performance categories rather than a numerical score scale. The performance categories are: Achieved the Standards with Honors, Achieved the Standard, Nearly Achieved the Standard, Below the Standard, and Little Evidence of Achievement. The state requires that students perform at either the Achieved the Standard or the Achieved the Standards with Honors category to be considered to have met the state standard.

The National Assessment of Educational Progress administered the Reading test in State B in the years 1992, 1994, and 1998. Only the 1998 year overlaps with the State B results reported here. However, to get some sense of trend, both the 1994 and 1998 results will be considered. NAEP results are reported in a number of different ways. The closest correspondence to the State B method of reporting is the percent at or above an achievement level. There is no exact correspondence between the NAEP achievement levels and the X Standards Program reporting categories so NAEP results are reported using both the Basic and Proficient categories. These seem to be the closest to the State B performance standard.

Reading Gains. The state assessment results are reported as a three-year rolling average percent above the standard. That is, the percent above the performance standard is computed for each of the past three years, and the results are averaged. When the next year's results are available, the oldest year's results are dropped and the average is recomputed on the newest three years results. This approach is used to smooth out year-to-year fluctuations in the student population. The basic results from the reports for 2000 and 2001 are summarized below for grades 4, 8, and 10.

Percent of Students Meeting or Exceeding Reading Performance Standards
for Years 1998 to 2000 and 1999 to 2001

	Grade 4		Grade 8		Grade 10	
Years	Basic Understanding	Analysis and Interpretation	Basic Understanding	Analysis and Interpretation	Basic Understanding	Analysis and Interpretation
98-00	74	58	49	26	NA	NA
99-01	76	62	49	23	34	27
Change	2	4	0	-3	NA (2)	NA (6)

NA indicates that data are not available for the three year rolling average. The numbers in parentheses are the growth implied by the results spanning years 1999 to 2001.

These results show that there is growth in Grade 4 and likely growth in Grade 10, but Grade 8 shows no change in Basic Understanding and a decline in Analysis and Interpretation. Because these percentages are essentially population statistics, that is, all eligible students are included, all differences are statistically meaningful.

The overall performance on NAEP Reading for a sample of students from State B is given in the following table. The results are available for both 1998 and 1994 for Grade 4, but only for 1998 for Grade 8.

Percent at or above NAGB Achievement Levels
for NAEP Reading by Year and Grade

Achievement Level	Year	Grade 4	Grade 8
Proficient	1998	32	30
	1994	32	
Basic	1998	65	74
	1994	65	

Note that there was no change in the Grade 4 performance over the 4 years between 1994 and 1998 for either the Proficient or Basic levels. This is somewhat inconsistent with the State B results that shows slight growth at Grade 4 after 1998, but since the years covered are not the same, it is certainly possible that changes to the State B program after 1998 could have brought about growth that is undetected by NAEP.

The standards used by State B and NAEP are not quite consistent, even on an equipercentile basis. The Basic level on NAEP Reading seems similar to the State B standard at Grade 4, but the Proficient level of NAEP seems closer the State B standard for Grade 8. These comparisons are difficult because of differences in the content, scoring and reporting for the two testing programs.

Reading Gaps. State B reports results for several subgroups of students. These results can be used to determine whether the gaps in performance between these subgroups are declining over time. The reporting subgroups are White, Asian/Pacific Islander (A/PI), Black (B), Hispanic (H), and American Indian (AI). The following table shows the difference in the percentage meeting or exceeding the state standard between the White population and each of the other groups by grade and year.

Gaps in Performance in Percent at or above Standard

	Grade 4				Grade 8				Grade 10			
Year	A/PI	B	H	AI	A/PI	B	H	AI	A/PI	B	H	AI
1998	18	31	35	30	8	22	26	29	NA	NA	NA	NA
1999	7	27	30	7	5	15	13	11	5	13	13	6
2001	14	27	32	22	14	25	27	26	15	14	26	24

NA indicates that the data are not available for Grade 10 in 1998.

These results are extremely variable. The difference between the 1999 year and the other years suggests that different definitions might have been used to define the subgroups in that year. Overall, the results for reducing gaps are inconsistent and it is difficult to determine if there is a trend over the three years for which data are available.

State B also reports results for students below the poverty level, for students in special education and for students with limited English proficiency. The difference between the percentage meeting or exceeding the performance standard for these groups and for the remainder of the student population are given below.

Gaps in Percentage above Standard
between Identified Group and Regular Student Population

	Grade 4			Grade 8			Grade 10		
Year	Below Poverty Level	Special Education	LEP	Below Poverty Level	Special Education	LEP	Below Poverty Level	Special Education	LEP
1998	NA	35	48	NA	36	39	NA	NA	NA
1999	25	35	35	NA	24	21	NA	19	20
2001	18	36	51	NA	33	37	NA	28	30

NA indicates that the data were not available for the subgroup grade combination.

The results across years are not consistent, so it is difficult to make inferences about the closing of gaps between the identified groups.

NAEP data are available for some of the subgroups reported by State B -- Black, Hispanic, and Asian American – although it is not clear that the methods used for identifying group membership are the same. The data for the differences in percentages for the groups from the White sample are given in the following table.

Differences in Percentage at or above Achievement Levels
from the White Sample for NAEP Reading

		Grade 4			Grade 8		
Achievement Level	Year	B	H	AA	B	H	AA
Proficient	1998	26	30	14	18	24	-2
	1994	26	25	20			
Basic	1998	33	44	23	35	34	-4
	1994	33	34	29			

The magnitudes of the gaps between groups are roughly comparable for Grade 4 and Grade 8 Blacks and Hispanics, but the pattern of results for Asian American students seems different than those for the group labeled Asian American/Pacific Islander in State B. The differences may be due to differences in the criteria for placing students into the groups for the two testing programs.

Mathematics

Results for the Criterion Referenced Examination in mathematics are reported for three different curriculum areas: Skills, Concepts, and Problem Solving. The results are reported according to performance categories rather than a numerical score scale. The performance categories are: Achieved the Standards with Honors, Achieved the Standard, Nearly Achieved the Standard, Below the Standard, and Little Evidence of Achievement. The state requires that students perform at either the Achieved the Standard or the Achieved the Standards with Honors category to be considered to have passed the state standard.

The NAEP Mathematics Assessment was administered in 1996 and 2000. These years do not quite correspond to the reporting years for State B, but the results should give a rough indication of the trends in performance from 1998 on.

Mathematics Gains. The state assessment results are reported as a three-year rolling average percent above the standard. This approach is used to smooth out year-to-year fluctuations in the student population. The basic results for the last two years are summarized below for grades 4, 8, and 10.

Percent of Students Meeting or Exceeding Mathematics Performance Standards

	Grade 4			Grade 8			Grade 10		
Year	Skills	Concepts	Problem Solving	Skills	Concepts	Problem Solving	Skills	Concepts	Problem Solving
98-00	56	24	18	52	20	24	43	19	15
99-01	58	29	21	50	18	25	37	21	16
Change	2	5	3	-2	-2	1	-6	2	1

These results show that there is growth in Grade 4 and mixed results for Grades 8 and 10. Because these percentages are essentially population statistics, all differences are statistically meaningful. The decline of 6 percentage points in skills at Grade 10, and the increase of 5 percentage points for concepts at Grade 4 are particularly notable especially because they are based on the three year rolling averages.

The percentages at or above the Basic and Proficient achievement levels for NAEP Mathematics are given in the following table.

Percentage at or above the Achievement Levels
by Grade Level and Year

Achievement Level	Year	Grade 4	Grade 8
Proficient	2000	23	24
	1996	17	20
Basic	2000	67	64
	1996	61	60

The pattern of increase is fairly clear from the table. For both grades, the percent at or above the achievement levels increased over the years covered by NAEP testing. The Proficient level seems comparable to the results for Concepts and Problem Solving on the Criterion Referenced Examinations and the Basic level seems more comparable to the Skills content. The trend in performance is similar for Grade 4, but not for Grade 8.

Mathematics Gaps. State B reports mathematics results for several subgroups. These results can be used to determine whether the gaps in performance between groups are declining over time. The reporting subgroups are White, Asian/Pacific Islander (A/PI), Black (B), Hispanic (H), and American Indian (AI). The following table shows the difference in the percentage meeting or exceeding the state standard between the White population and each of the other groups by grade and year. For the year 2001, the state averaged the results from the two content areas so there is only a single number in the table for each subgroup in 2001.

Gaps in Performance in Percent at or above Standard

	Grade 4								Grade 8								Grade 10							
Year	A/PI		B		H		AI		A/PI		B		H		AI		A/PI		B		H		AI	
	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS	S	PS
1998	10	5	40	15	42	16	35	17	12	2	32	21	32	22	27	21	6	6	29	17	30	17	31	16
1999	6	5	34	24	32	23	11	7	8	9	31	26	29	25	22	17	8	2	37	16	36	16	17	10
2001	10		28		29		20		8		27		28		24		6		26		26		24	

S indicates Skills and PS indicates Problem Solving.

These results are extremely variable. Overall, the results for reducing gaps are inconsistent and it is difficult to determine if there is a trend over the three years for which data are available.

State B also reports results for students below the poverty level, for students in special education and for students with limited English proficiency. The percentage point gaps for these groups with the remainder of the student population are given below. Again, the results for 2001 are an average gap rather than the gaps on the individual content areas.

Gaps in Percentage above Standard
between Identified Group and Regular Student Population

	Grade 4						Grade 8					Grade 10				
Year	Below Poverty Level		Special Education		LEP		Below Poverty Level	Special Education		LEP		Below Poverty Level	Special Education		LEP	
	S	PS	S	PS	S	PS		S	PS	S	PS		S	PS	S	PS
1998	NA	NA	26	11	52	16	NA	47	9	30	6	NA	36	16	47	16
1999	25	22	27	18	41	20	NA	43	27	53	31	NA	37	17	32	18
2001	19		21		32		NA	29		33		NA	25		24	

NA indicates that the data were not available for the subgroup grade combination. S indicates Skills and PS indicates Problem Solving.

Because of the inconsistent pattern across years, it is difficult to conclude that gaps are increasing or decreasing. It is possible that the results only reflect year to year changes in the student population.

The differences in NAEP Mathematics performance for subgroups are given in the following table. The differences in percentage at or above the NAGB achievement levels for Black, Hispanic, and Asian American students from the results for the White students are provided.

Difference in Percentage at or above NAGB Achievement Levels
on NAEP Mathematics for Identified Group and the White Sample by Grade

		Grade 4			Grade 8		
Achievement Level	Year	B	H	AA	B	H	AA
Proficient	2000	26	25	9	23	25	8
	1996	17	13	4	17	20	6
Basic	2000	42	46	24	41	42	11
	1996	43	33	20	36	40	11

The results from the NAEP data show that the gaps in performance are increasing from 1996 to 2000 in almost every case. The exception is for the Asian American students in Grade 8. This result is counter to that for the Criterion Referenced Examination where the gaps generally do not change or get smaller. These differences could be due to differences in test content, the location of the cutscores for the standards, or to different ranges in the coverage of years.

Adequate Yearly Progress

State B allows each school to set its own targets for adequate yearly progress within state guidelines. The state guidelines include the following conditions:

- Targets are based on reducing percentages of
    - students below standard
    - students in the lowest performance level

- The percentages are to be reduced between 9% and 15% over three years, starting in 2000
- The school will use the three-year rolling average to measure progress
- Each school will set targets for closing equity gaps.

The schools will not be held to the standards until 2003 using the 2000 rolling average as the base year.

Discussion

The interpretation of the results presented here is extremely challenging because of the differences in the tests used for assessing student performance, the way that the State B program is implemented, and differences in the time periods covered by the reporting of results. These issues will be discussed in more detail later in this report. First, the general results for State B will be summarized. Then the consistency of the NAEP results with Criterion Referenced Examination results will be discussed. Finally, the issues in making such comparisons will be listed with the goal of providing cautions about the over interpretation of observed differences in results.

Results According to State B Performance Goals

State B has specified goals for its schools to reduce the percent of students who are performing below standard by 3% to 5% each year. It is not clear how this school level goal applies for the interpretation of overall state results. Many Title I schools could meet the goal, but the state as a whole might not meet the goal, depending on the proportion of students in the state from schools that met the goal and the proportion of students who were below standard in the school. For the year 2000, 36% of the students were below standard in Basic Understanding in Reading. Using the 3% reduction requirement, .35 x .03 = .01, a 1% reduction in proportion below the standard is needed. The state reduced the percentage below the standard by 2%. It is difficult to determine how the rolling average procedure used by State B affects these results. It may make it more difficult to achieve the stated goals.

The state requires specific plans for reducing gaps when the gaps between groups are greater than 15%. For Reading, the gap between White performance and that the Black, Hispanic, and American Indian Grade 4 and 8 populations are greater than 15% for 2001. At Grade 10, the gap is also present for Hispanic and American Indian populations. Gaps are also present for Below Poverty Level, Special Education, and LEP populations at all grade levels. Mathematics results show a similar pattern. Also, gaps exist for the Black, Hispanic, American Indian, Poverty Level, Special Education, and LEP populations at all grade levels.

The NAEP results are inconsistent with the Criterion Referenced Exam results, but not in a regular pattern. This suggests that the NAEP assessments have different characteristics than the Criterion Referenced Exams in terms of content, level of standards, and technical features such as reliability and conditional standard error.

Issues

Comparing results for State B on the Criterion Referenced Examinations and NAEP is very difficult for a number of reasons. These are:

1. The assessments have different content specifications.
2. The assessments have different proportions of open-ended tasks. The Criterion Referenced Exams have greater proportions of performance tasks than NAEP does.
3. The scoring process for the open-ended items is different for the two assessments.
4. The reliability and conditional standard errors for the two assessments are likely very different.
5. Individual student scores are reported for the Criterion Referenced Examinations, but they are not for NAEP.
6. NAEP results are for a sample of students. The Criterion Referenced Exam is administered to all eligible students.
7. The Criterion Referenced Exam reports results on multiple curriculum areas using performance categories only. NAEP reports results on a single score scale using both numerical scores and achievement level categories. The results are reported at no level lower than the state.

Because of all of these issues, it is likely that NAEP and Criterion Referenced Exam results are only moderately correlated. Also, the standards of performance are likely quite different. Further, because the Criterion Referenced Examinations are tied to state policy, performance on the examinations are likely to be more sensitive to instruction than NAEP.

Because of the differences between NAEP and the State B assessment program, the results should be compared with great care. First, some understanding of the content coverage of the two programs is needed. Second, it should be understood that the relationship between instructional emphasis and content coverage can result in legitimate differences in results on the two testing programs. If State B instruction emphasizes content that is on the state assessment, but that is not on NAEP, the state assessment may accurately reflect performance gains when NAEP does not. Differences in stakes for the two testing programs will also influence the results.

The differences in testing programs suggest that NAEP can only be used to confirm state assessment results in a very general way. Trends in the same direction, or lack of directly contradictory information should be considered as confirmation.

Using NAEP data to confirm progress in "State C"

Paul W. Holland
Planning Work Group
1/7/02 (revised 1/28/02)

1. Introduction

This is a draft of a "partial argument" that might be put forward by a state to show that it is in compliance with the requirements of the "No Child Left Behind" (NCLB) legislation within the reauthorization of the ESEA in 2002. It is based on data from a real state, but this state is called "State C" here in order to focus on the argument rather than on the correctness of the details of the data from the state. Furthermore, it is limited by my current (as of the first date on this draft) understanding of the NCLB legislation, which I am only familiar with through second-hand reports. The purpose of this exercise is to see what issues arise when trying to apply the requirement of the NCLB legislation to a real case and thereby stimulate discussion within the NAGB Ad Hoc Committee on Confirming Test Results. It should not be construed in any way beyond this limited objective.

Furthermore, the NCLB requirements include many things that I will not say much about here. These include: how challenging the state's content standards and achievement standards are, the degree of alignment between the achievement and content standards (though this seems to be a serious effort in State C), and the alignment between achievement and content standards (though this seems to also be seriously considered in State C). After briefly reviewing the recent history of state assessment in State C, I will concentrate on progress as measured by test results in grades 4 and 8 in Reading and Mathematics.

2. Brief history of assessment in State C

While State C's assessment system was somewhat consistent over the 1990s, it was by no means a simple repeat each year of what had been done the year before. The grades that were tested each year sometimes changed and new tests were added, especially in the upper grades. By the end of the decade a new set of objectives and student expectations were formulated in English language arts, mathematics, science and social studies for most grades. This resulted in a new statewide curriculum and the state assessments were altered, aligned more closely with this curriculum and given a new name.

By 2005, Reading and Mathematics will be assessed in grades 3 to 9 (this fits with the NCLB requirements), and promotion requirements for grades 3, 5 and 8 will be tied to the state assessments for the cohort starting first grade in 2000-2001. Currently the state assessment is part of the high-school graduation requirement. (These pre-2005 testing points also fit with the requirements of the NCLB legislation).

My point in giving this brief history is to document the fact that state assessment systems can change quite a bit over a ten-year period. The changes in State C are very much driven by the accountability movement, and for this reason are organized towards a producing a stronger and stronger accountability system for the education of the students in State C using unified curricula and carefully monitored and well aligned state assessments as part of the system. This is a case of a state which has been committed to the accountability movement for some time and which has committed considerable resources to it. It provides us an example where many of the ideas of NCLB have been implemented for some time.

3. Overall Performance Trends

I decided to use the data from 1993-94 to 1999-2000 from State C because this was all available for most of the groups. I also wanted to avoid any startup problems with the state assessment system and to have a range of years where there was NAEP data in both subjects for comparison. Thus, I will attempt to use 1994 as the base year to emulate the NCLB use of 2001-2002 as a base year. Rather than go through various groups, I will look only at one, Black examinees, for reasons that will be mentioned below. What I do here could be done for each group of interest.

3.1 Adequate yearly progress: From the information that I had, I could not find a statement about what adequate yearly progress meant from the perspective of State C. For this reason, I followed the information I was given about the NCLB legislation and devised a system of my own. This is just an example and might be differently arranged in the real setting. Furthermore, this is not an attempt to evaluate State Cs progress, but merely to illustrate what it looks like, and to discuss some of the issues that arise in measuring it.

In the materials I had for State C, I found at least three different references to performance that might be considered by some as “proficient”. In order to proceed, I chose the least stringent one, which is considered a “passing” score for the grade by State C. The other standards ranged up to a level that was worthy of academic recognition and that seemed more like an “advanced” standard to me.

I had data for the following groups, White, Black, Hispanic, and "economically disadvantaged".

In 1994 the group with the minimum percent meeting the proficient standard that I used for State C was Black examinees with 56% in grade 4 Reading, 58% in grade 8 Reading, 36% in grade 4 Mathematics and 32% in grade 8 Mathematics. I decided to focus on Black examinees in terms of this analysis of overall progress because they had the farthest to go. Much the same story could be told for the other groups, but they were, in the base year, closer to meeting the standard that I chose to regard as the 12-year target for students in State C.

In order to meet the 12-year target of 100% of Black students meeting the proficient standard, it will take an average yearly increase in these percentages of from 3.5 to 5.7 percentage points per year, depending on the grade and subject. Thus, I give in Table 1, the target percentages for each year (and grade and subject) along with the actual percentages for Black examinees in State C achieving the proficient level over the 7-year period for which I have data. I did not try to put values in for the other, less stringent, yearly targets that are mentioned in the NCLB legislation. They would be lower than the ones in Table 1.

Table 1: Target and actual percentages meeting the proficient standard for Black examinees in State C. In each cell, the number *above* is the actual and the number *below* is the 12-year target.

Year	**'94**	**'95**	**'96**	**'97**	**'98**	**'99**	**'00**
Grade 4 Reading	56 56	61 60	60 63	66 67	77 71	79 75	82 78
Grade 8 Reading	58 58	57 62	60 65	70 69	71 72	81 76	83 79
Grade 4 Math	36 36	47 41	57 47	62 52	69 57	73 63	75 68
Grade 8 Math	32 32	30 38	44 43	55 49	66 55	74 61	81 66

I did not know what more to do with this, so I will leave with the observation, that according to its state assessment results, and using my made up standard and rules, State C is achieving *and exceeding* the target percentages of its lowest performing group on the way to 100% proficient in 12 years. They are doing this more or less in accord with the NCLB requirement of approximate equal increments per year, except that in the earlier years (95 and 96) there were some cases where the goals were not being met.

Presumably in the full case we would see similar tables for all of the groups of interest. The yearly targets would be different for different groups and the case I examined is the group that has to increase to most in the 12-year period.

Table 1 can be supplemented with more detailed graphs such as Figure 2, 6, 10 and 14 in the "Aspen book" for State C. (These four Figures are included here and labeled as they are in the Aspen book, with the State called State C.) These Figures show the CDFs for the entire distribution of scores for both Black and White examinees over the period 1994 to 2001 for each relevant grade and subject,. They include the cut-score for achieving the "proficient" standard I am using here. They show that the improvement is across the board at every score level.

(Figures 2, 6, 10, 14 go about here)

If I had set a higher standard for "proficient", then in the base year there would be fewer examinees achieving it so the annual rate of improvement would have to be higher than what I used. The effect of this may be seem by imagining the vertical dotted line in the Aspen Figures moving to the right. If this is done, then more of the students will be to the left of the dotted line indicating that fewer of them are meeting the standard. This may explain some of the discrepancies between the targets and the actuals in Table 1, (i.e., State C may be trying to meet a higher standard than the one I used).

3.2 Does NAEP support or confirm this improvement? None of the NAEP Achievement Levels corresponds closely (in terms of the percent of students in State C achieving them) to the standard for "proficient" that I used. For example, in 1994, the percent of all 4th grade students in State C achieving NAEP Basic or above in Reading was 58% while those achieving NAEP Proficient of above was 26%. On the other hand, in 1994, the percent of students achieving what I have called "proficient" in the above analysis on the State Assessment was much higher, at 73%. Similar discrepancies between the percents of students achieving the “proficient” level that I used and NAEP achievement level percents exist for the other grade and subject in the two sets of years where there is relevant State NAEP data—‘94 and ‘98 in Reading and ‘96 and 2000 for Mathematics. Table 2 shows the relevant NAEP data for Black examinees using the percent above the NAEP Basic achievement level as the comparison value.

Table 2: Percents above Basic for State C

Year	**‘94**	**‘96**	**‘98**	**‘00**
Grade 4 Reading	38		38	
Grade 8 Reading	NA		54	
Grade 4 Mathematics		47		60
Grade 8 Mathematics		31		40

Without any data in 1994 for Grade 8 Reading, from the data in Table 2 we can not say anything about Reading except that there was no change for Black students in Grade 4 regarding the percent above NAEP Basic over the two relevant assessment years. This is not the pattern that we see in Table 1, or in the Aspen Figures.

The results for Mathematics are different. There are large changes in the percents above the NAEP Basic achievement level for Black examinees in both grades. These changes are in the same direction that the State assessment data show. NAEP clearly supports the conclusion of improvement in the performance in Mathematics for the Black examinees in State C.

In Reading, the strongest statement I could make is that the NAEP results do not show a decrease, based on the data for the percent above Basic as the criterion. However, if we examine the full distribution of scores for Black examinees in the two years we see a somewhat different story. Table 2.5 shows that there was positive change over time for Black 4th graders in Reading, but that it is mostly at the lower score levels of NAEP. Table 2.5 gives several percentiles that range from low scores (10th percentile) to high scores, the (90th percentile).

Table 2.5 Six Spaced Percentiles and the Mean for Reading Score for Black 4th graders in State C on NAEP

Percentile	10th	25th	50th	75th	90th	Mean
1994	139	168	196	219	237	191
1998	154	175	198	220	240	197

Thus, in order to see where NAEP shows positive improvement over the two years it is necessary to look more closely at the data than just the percent achieving the NAEP Basic level.

4. Gaps between the performance of subgroups of students

I will limit my discussion of Gaps to the Black/White comparison.

4.2 The Black/White Gap: Table 3 shows the percents achieving the "proficient" level for White and Black examinees for '94, '97, and 2000 (just to limit the amount of data in the table) as well as the difference in these percents for these three years.

Table 3: Difference in percents achieving proficient on the State C assessment, White minus Black. In the cells the subtractions are shown.

Year	'94	'97	'00
Grade 4 Reading	83 – 56 = 27	86 – 66 = 20	95 – 82 = 13
Grade 8 Reading	86 – 58 = 28	89 – 70 = 19	95 – 83 = 12
Grade 4 Math	67 – 36 = 31	86 – 62 = 24	93 – 75 = 18
Grade 8 Math	70 – 32 = 38	83 – 55 = 28	95 – 81 = 14

Table 3 shows a clear decrease over the 7-year period in the gap between White and Black examinees on the State C Assessment.

I think this is a clear-cut example of "gap closing". The effects are huge and they are not localized at a few score levels but are across the board. If we examine the CDFs in Figures 2, 6, 10, 14, it is clear that the gaps are closing because the lower scoring group is progressing at a faster rate than the higher scoring group is.

4.2 Does NAEP support the gap closing? Table 4 combines features of Tables 2 and 3. It shows, where available, the difference in the percent above the NAEP Basic Achievement Level for White and Black examinees.

Table 4: Difference in percents above the NAEP Basic Achievement Level, White minus Black. In the cells the subtractions are shown.

Year	**‘94**	**‘96**	**‘98**	**‘00**
Grade 4 Reading	73 – 38 = 35		80 – 38 = 42	
Grade 8 Reading	NA		87 – 54 = 33	
Grade 4 Mathematics		85 – 47 = 38		89 – 60 = 29
Grade 8 Mathematics		78 – 31 = 47		83 – 40 = 43

In Grade 4 Reading, where we have enough data to make a comparison over time, the trend in Table 4 based on the percents achieving NAEP Basic is contrary to the good news of Table 3. In Mathematics, we do get trends that are in the same direction shown in the State C Assessment results.

I wondered if a more careful look using some of the “gap graphs” that I have suggested elsewhere would shed any light on the Grade 4 results. Figures C1 and C2 are the corresponding “difference in percents” and “difference in percentiles” plots. They cover three years of data and include NAEP Reading 1992 assessment. These graphs show that between 94 and 98 the gaps did get bigger both from the point of view of differences in percents, Figure C1 (where the black curve for 1998 is higher than the dotted curve for 1994, and therefore the gap is bigger) and from the point of view of differences in percentiles, Figure C2 (where the black curve for 1998 is higher over most of the range than the dotted curve for 1994).

(Figures C1-C3 go about here)

Figure C3 shows that these changes in gaps are due to an upwards shift in the distribution of scores for White examinees from ‘94 to ‘98 but an upwards shift only in the lower scores for Black examinees. This is also seen in Table 4 where the percent of White examinees above the NAEP Basic achievement level goes up over the two year, 73 to 80, but the corresponding percentages for Black examinees is unchanged at 38.

Thus, NAEP Reading shows an increase in scores for both Black and White examinees, but only for the lower scoring part of the distribution for Black students. This is a different type of picture than what is shown for NAEP Mathematics. These types of changes show how complex the issues can be for using NAEP results to compare to State assessment results.

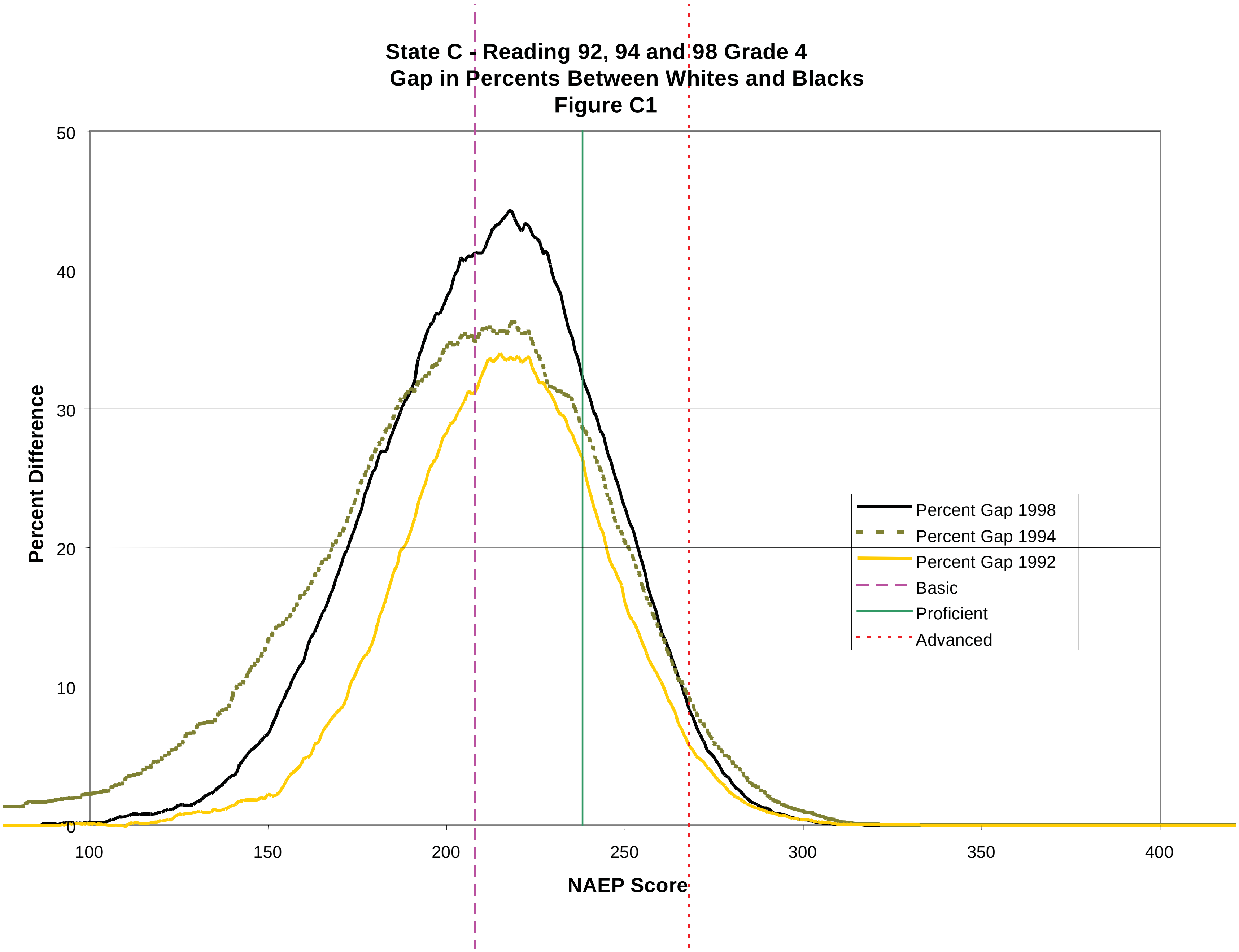
5. How many students are tested?

(I included this because I had some relevant data available, it is not central to the analysis.) According to the NCLB legislation, a state must assess at least 95% of the students in each group. I do not have data on this, but the following data are relevant. The state reports total K-12 enrollment and the number of students tested in a subset of seven of these grades (3-8, 10) by group—All, Black, Asian, Hispanic, Native American, White, Economically disadvantaged. I have the data for 97-98 to 99-00, 4 years. I took the total K-12 enrollment and multiplied it by 7/13 to get an estimate of the enrollment in the tested grades. Table 5 shows the number of tested students in each group as a percent of these estimates.

Table 5: Number of students tested each year as a percent of estimated enrollment using estimate described in the text.

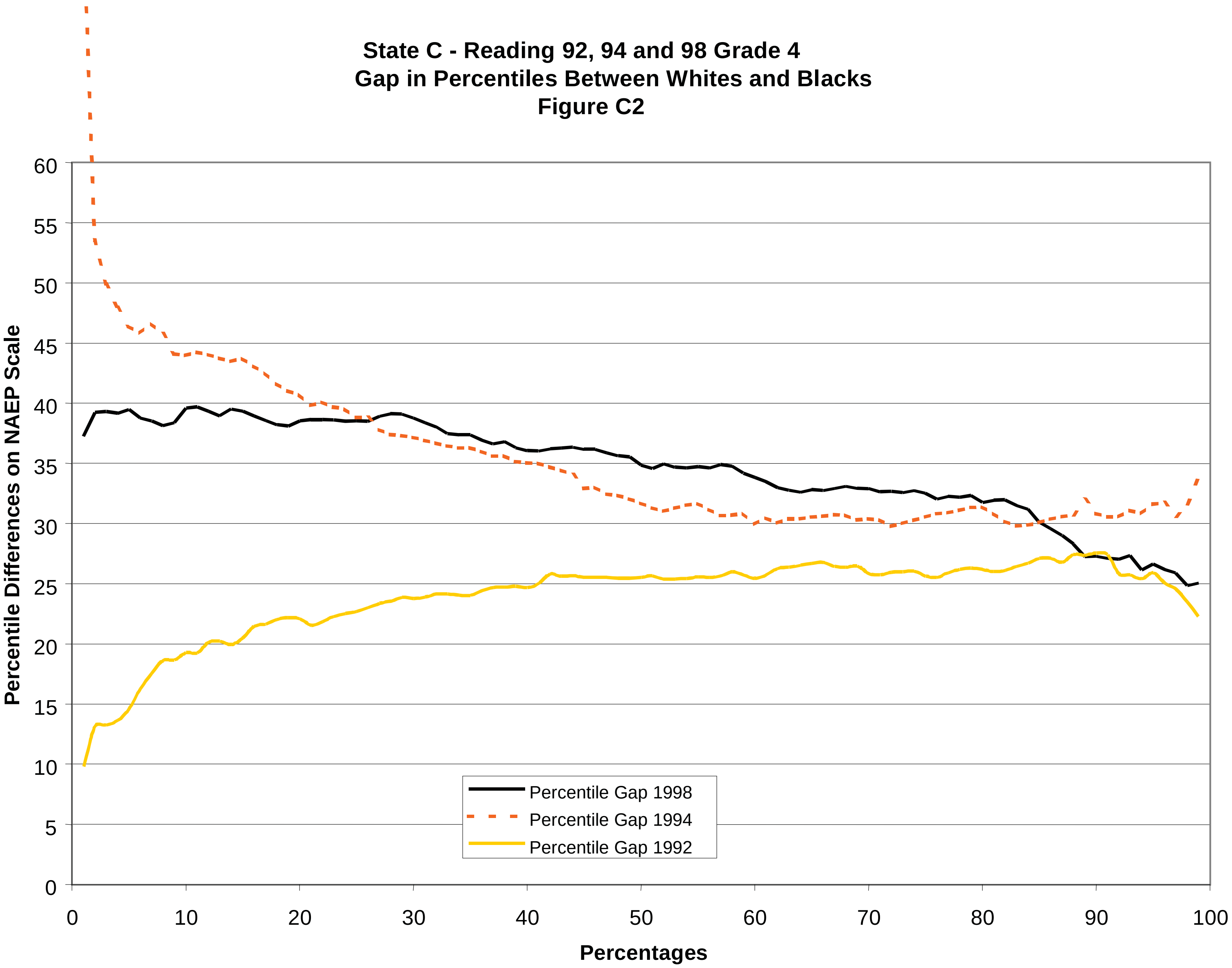
Year	97-98	98-99	99-00
All	98	98	98
Black	96	97	97
Asian	99	98	98
Hispanic	95-	95-	95
Native American	105	100-	104
White	100+	101	101
Econ. Disadvant.	95	95	96

Table 5 indicates that in State C a very large proportion of students in each of the groups is tested. The estimate gives unreasonably high values in certain cases but these results suggest that the difference in the rates of testing of the different groups are quite in line with the requirements of NCLB. Of course, in a real case we would like to see better numbers than my estimates that are based on assumptions that must be very wrong some of the time.



State C - Reading 92, 94 and 98 Grade 4
Gap in Percents Between Whites and Blacks
Figure C1

State C - Reading 92, 94 and 98 Grade 4
Gap in Percentiles Between Whites and Blacks
Figure C2



State C - Reading 92, 94 and 98 Grade 4
Cumulative Distribution Function for Whites and Blacks
Figure C3

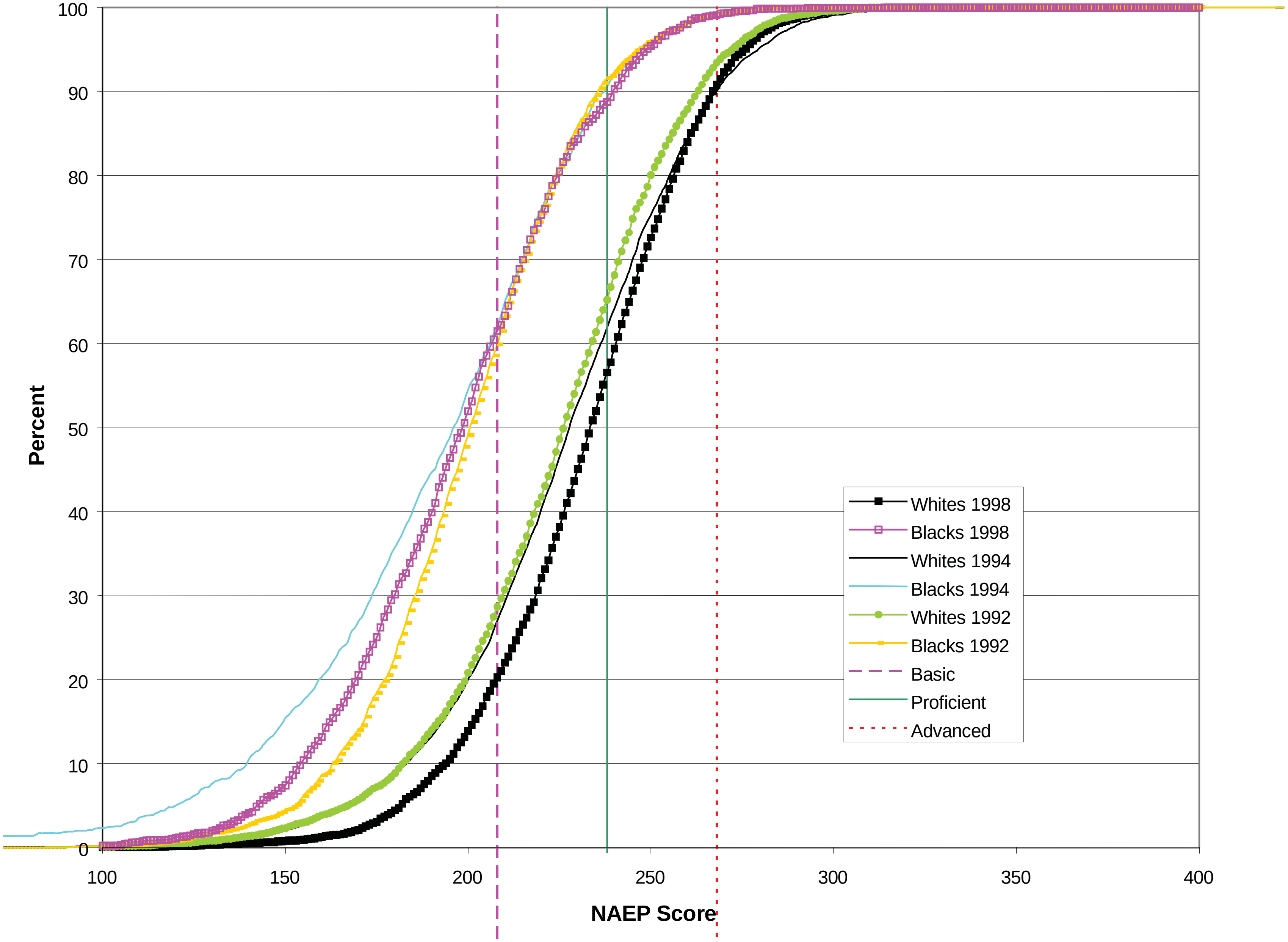


Figure 2
CDF: State C Learning Index 4th Grade Reading Scores by Race/Ethnicity (Black/White): 1994 and 2001

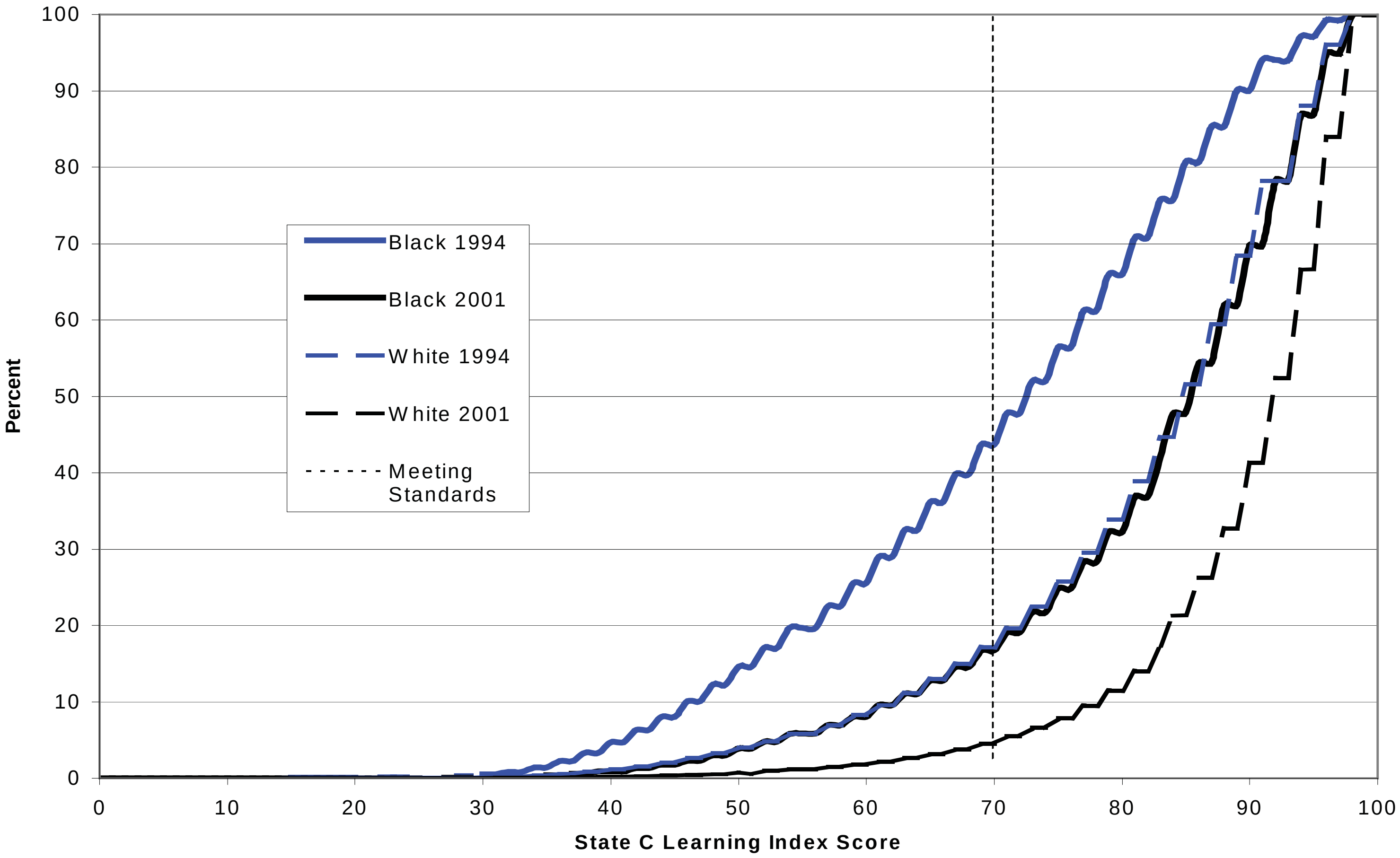


Figure 6
CDF: State C Learning Index 4th Grade Math Scores by Race/Ethnicity (Black/White): 1994 and 2001

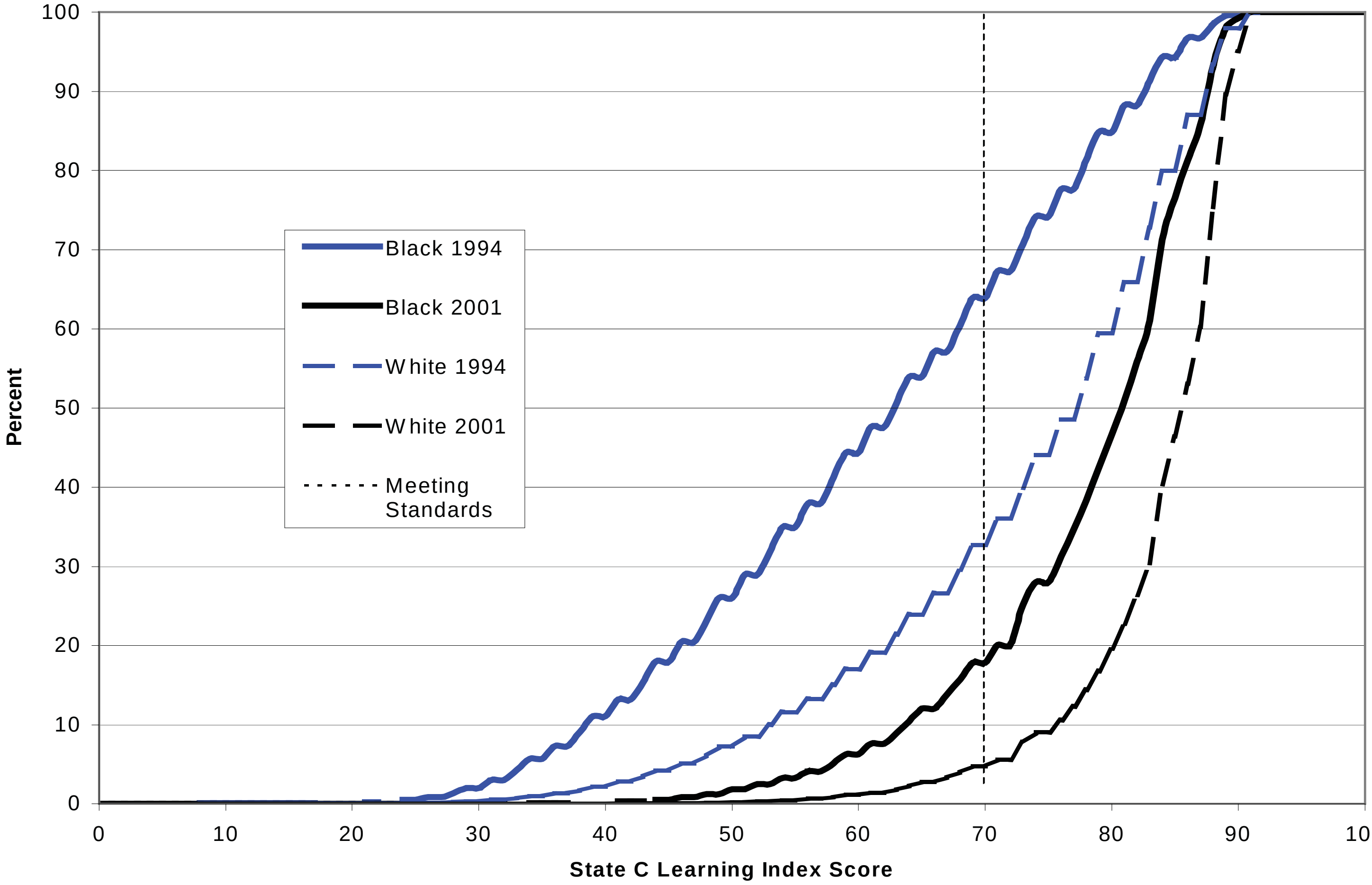


Figure 10
CDF: State C Learning Index 8th Grade Reading Scores by Race/Ethnicity (Black/White): 1994 and 2001

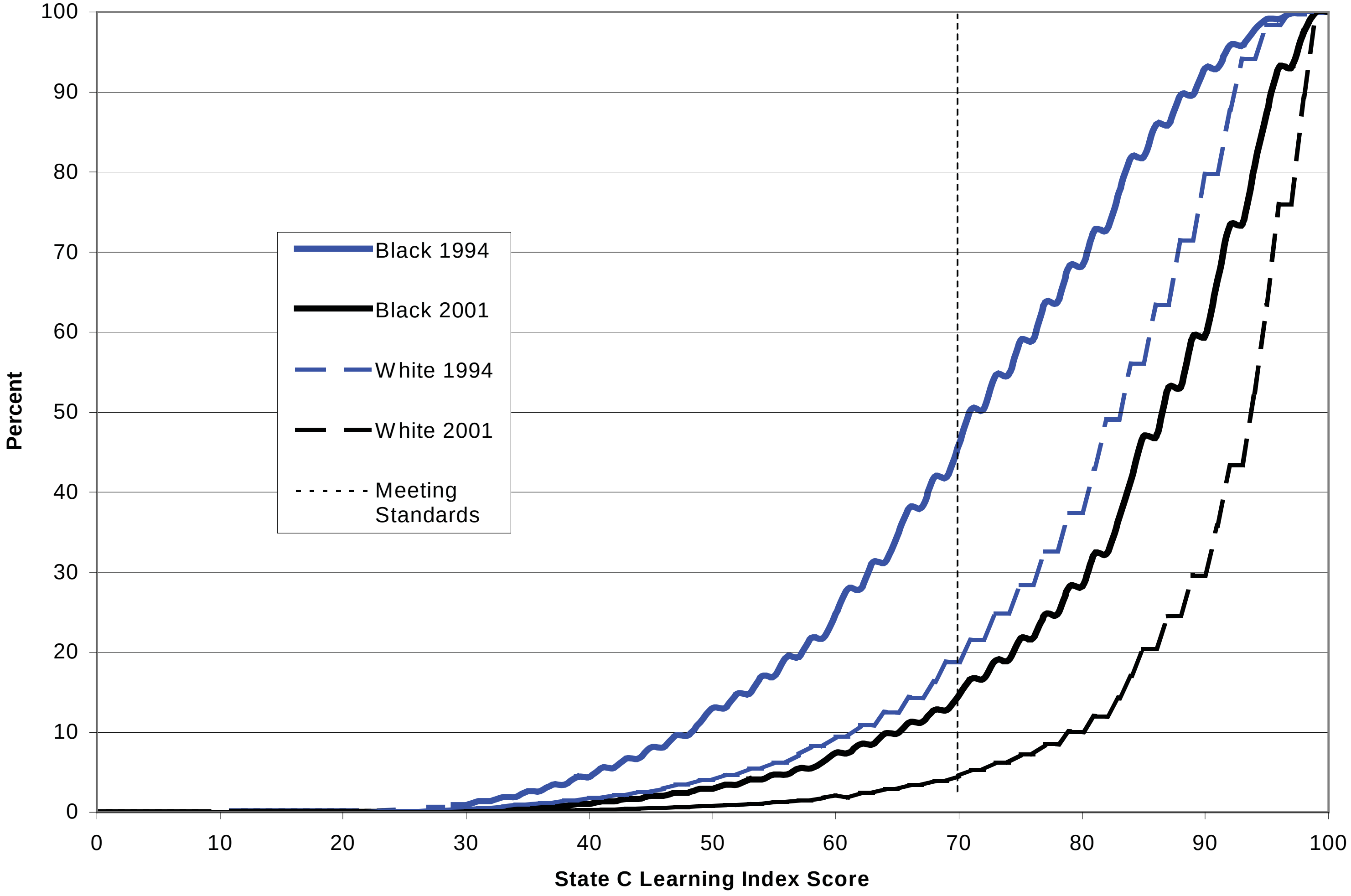
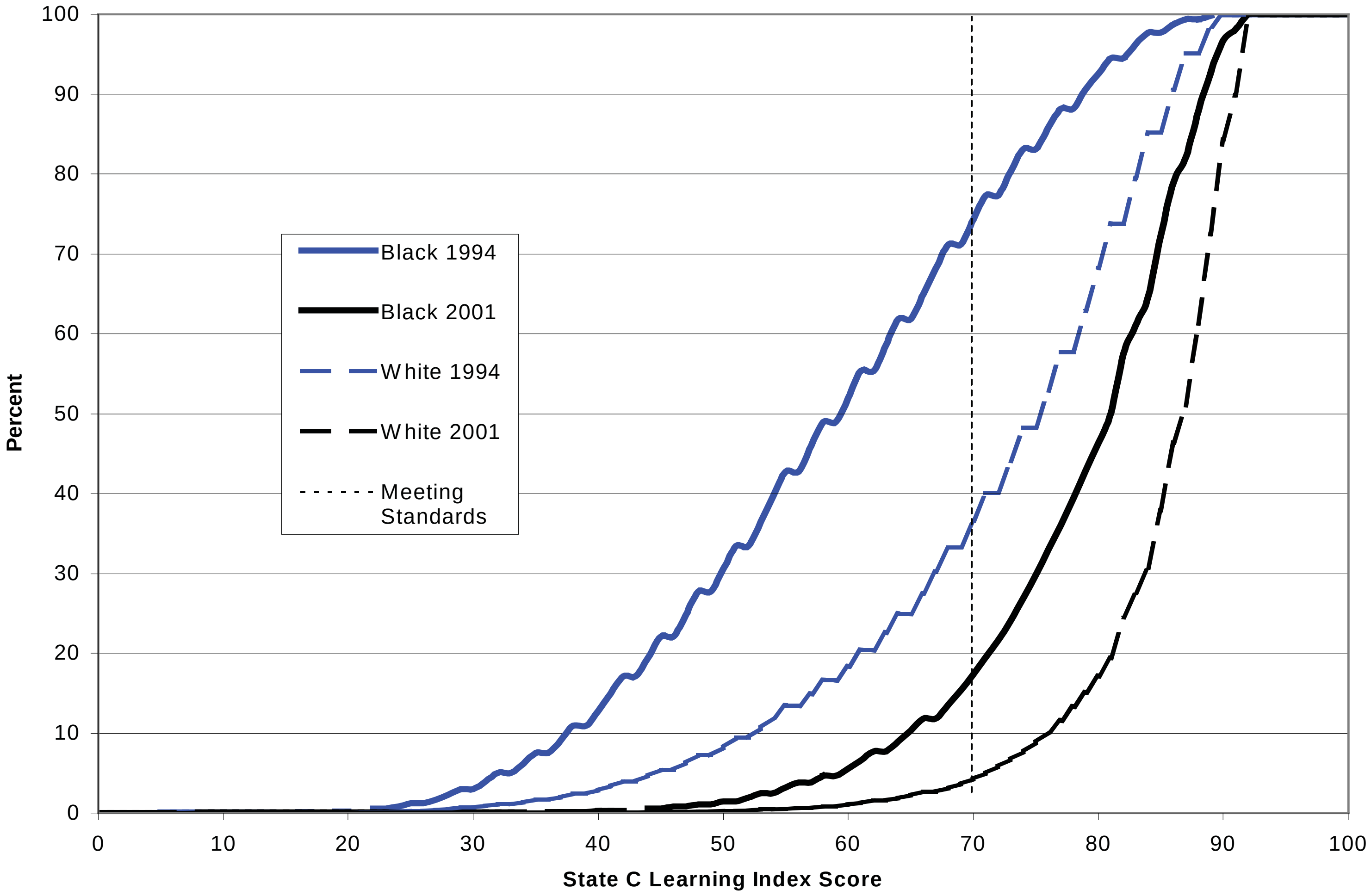


Figure 14
CDF: State C Learning Index 8th Grade Math Scores by Race/Ethnicity (Black/White): 1994 and 2001


Percent
State C Learning Index Score
Black 1994
Black 2001
White 1994
White 2001
Meeting Standards