

Analyzing, Disaggregating,
Reporting, and Interpreting
Students' Achievement Test Results:

A Guide to Practice for Title I and Beyond

Richard M. Jaeger
and
Charlene G. Tucker



PRE-PUBLICATION COPY



Analyzing, Disaggregating,
Reporting, and Interpreting
Students' Achievement Test Results:

A Guide to Practice for Title I and Beyond

Richard M. Jaeger*
Center for Educational Research and Evaluation
University of North Carolina at Greensboro

and

Charlene G. Tucker
Connecticut Department of Education

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

*This paper was completed while the first author was a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University. I am grateful for the partial support provided by the Spencer Foundation under Grant Number 199400132.



The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues.

Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center is a permanent, central part of the Council of Chief State School Officers. The Center was established through a resolution by the membership of CCSSO in 1984.

This report is part of a series sponsored by the State Education Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Comprehensive Assessment Systems for IASA Title I/Goals 2000. The series addresses issues related to the standards and assessment provisions of Title I.

The Title I requirements to disaggregate student assessment data raise a number of policy and technical issues that state and local education agencies must consider. This report addresses many of these issues, acknowledging that they are likely to be encountered in most large-scale assessment programs regardless of the specific focus of those programs. As assessment programs are designed or modified to include all children, and fewer children are excluded for reasons such as limited facility in the English language and various handicapping conditions, additional issues related to disaggregation will arise. One example is the biasing effect of differing student participation rates across schools and school districts. The SCASS project plans to address such questions in a series of tightly-focused reports on such topics as assessing English-language learners in ways that satisfy Title I requirements.

This report is also available on CCSSO's web site (<http://www.ccsso.org>).

Preparation of this report was supported in part by the U.S. Department of Education, Office of Elementary and Secondary Education. The views and opinions expressed in this report are not necessarily those of the Council of Chief State School Officers, the U.S. Department of Education, or the State Collaborative on Assessment and Student Standards.

Council of Chief State School Officers
One Massachusetts Ave., N.W., Suite 700
Washington, D.C. 20001-1431

Gordon M. Ambach, Executive Director

Edward R. Roeber, Director,
Student Assessment Programs

Wayne N. Martin, Director,
State Education Assessment Center

Phoebe C. Winter, Project Director,
State Collaborative on Assessment
and Student Standards

ISBN # 1-884037-45-3

Copyright © 1998 by the Council of Chief State School Officers, Washington, D.C.

Acknowledgments

It is with great pleasure that we recognize the assistance of a number of colleagues in the preparation of this report. First, and foremost, Dr. Phoebe Winter of the Council of Chief State School Officers provided invaluable help in several ways — from skillful management of the entire project, to obtaining essential data from several state departments of education and local education agencies, to a very thoughtful and thorough review of earlier drafts of the report. It was written with the support of the State Collaborative on Assessment and Student Standards, Comprehensive Assessment Systems for IASA Title I/Goals 2000, Study Group on Disaggregation of Assessment Data. All members helped in defining the purposes of the report, in guiding its development, in reviewing initial drafts, and in providing many helpful suggestions for improvement. Those members are

Dr. Alan Sheinker of the Wyoming Department of Education (Chair);
Dr. Dale Carlson of the California Department of Education;
Dr. Linda Hansche of Georgia State University; and
Dr. Grace Ross, Dr. Elois Scott, and Dr. Hugh Walkup of the United States Department of Education.

This report has been greatly enriched through the provision of student achievement data by several state departments of education and local education agencies. We are indebted to Dr. Douglas Rindone of the Connecticut Department of Education, Dr. Louis Fabrizio and Dr. Eleanor Sanford of the North Carolina Department of Public Instruction, and Dr. Mary Ellen Farmer of the Fairfax County (Virginia) Public Schools.

Finally, we are grateful to the Council of Chief State School Officers and the U. S. Department of Education for providing the support that enabled the preparation of this report. Although the report has benefited from the excellent advice of colleagues, we bear sole responsibility for any remaining errors it may contain.



Table of Contents

Background and Introduction	7
Disaggregation of Achievement Test Results	7
Who Should Read this Report.....	7
The Structure of the Report.....	8
How to Use this Report.....	10
A Word About Computing	12
Vignette 1: The Lincoln County School District	13
The Lincoln County School District.....	13
The State's Testing Program	13
The District Testing Committee	14
The Committee's Initial Deliberations and Actions.....	14
The Committee's Second Meeting.....	15
The Committee's Third Meeting	17
The Committee's Fourth Meeting.....	25
Vignette 2: The Harold Howe II High School.....	33
The Harold Howe II High School	33
The State's Testing Program	33
The School's Mathematics Curriculum Task Force	34
The Committee's First Meeting	35
The Committee's Second Meeting.....	36
The Committee's Third Meeting	38
The Committee's Fourth Meeting	43
Vignette 3: The State of Euphoria	49
The State of Euphoria	49
Euphoria's Statewide Testing Program	49
Euphoria's State Testing Office and Its Charge	51
Conclusions	63
Appendix A: Computing a Confidence Interval Around a Sample Percentage	64
Appendix B: Using Confidence Intervals to Compare Two Sample Percentages and Conducting a Hypothesis Test on the Difference Between Two Percentages....	65
Appendix C: Constructing a Confidence Interval Around a Sample Average	67
Appendix D: Testing the Null Hypothesis that Two Population Averages are Equal	68
Appendix E: Testing Simultaneous Null Hypotheses on Differences Between Pairs of Population Means.....	70
Appendix F: Computing the Effect Size of the Difference Between Two Sample Averages	71



Background and Introduction

When it passed the Improving America's Schools Act of 1994, the United States Congress created an obligation and an opportunity for states, local educational agencies, and schools to disaggregate and analyze data resulting from assessments of their students' achievement in ways that would be useful and illuminating. In particular, in Part A, Subpart 1 of Title I of the Act, the Congress required that (Section 1111 (b) (1) (A)):

Each state plan shall demonstrate that the State has developed or adopted a set of high-quality, yearly student assessments in at least mathematics and reading or language arts, that will be used as the primary means of determining the yearly performance of each local educational agency and school served under this part . . . Such assessments shall . . . enable results to be disaggregated within each State, local educational agency, and school by gender, by each major racial and ethnic group, by English proficiency status, by migrant status, by students with disabilities as compared to nondisabled students, and by economically disadvantaged students as compared to students who are not economically disadvantaged.

This portion of the Title I legislation raises a series of interesting analytic and policy issues that involve, among others, specification of appropriate definitions for categories of students, protection of students' individual privacy rights, appropriate generalization of test results beyond the group of students tested, valid interpretations of test results within and among categories of students, and portrayal of test results in ways that are interpretable and useful to a variety of audiences. The purpose of this report is to address a number of these issues as they arise when achievement test data are disaggregated and reported — hopefully in ways that will be useful to personnel in schools, local educational agencies, and states who bear responsibility for determining the role of school testing programs, the reporting of achievement test results, and the interpretation and use of such information.

Disaggregation of Achievement Test Results

Schools and school districts could view the Title I requirement to disaggregate achievement test data merely as an obligation — dutifully reporting the test performances of girls and boys; students with limited English proficiency and students who are fully proficient in English; students who identify themselves as African American, Asian or Pacific Islander, Hispanic, Native American or Alaskan Native, white non-Hispanic and so on — without further considering their findings. Or, they could capitalize on the opportunity created by the Title I requirements by analyzing their students' test performances in new and creative ways for a host of purposes, including analyses of current school policies, evaluation of the effectiveness of current school programs, planning the development of innovative instructional interventions, and reporting to parents and citizens on students' collective achievement status and progress.

This report contains illustrations of analyses and summaries of achievement test results by schools and school districts that have chosen the latter course, a route that offers many potential benefits but requires diligence and appropriate caution. If the objectives of this report are realized, it will inspire and stimulate novel ways of analyzing, presenting, and using students' achievement test results. It also will clarify some ways achievement test data can be appropriately and inappropriately used and interpreted.

Who Should Read this Report

It might seem presumptuous of any author to include a section with the heading of this one. However, it is often helpful to know whom the authors had in mind when they were writing.

They could capitalize on the opportunity created by the Title I requirements by analyzing their students' test performances in new and creative ways.

This report is principally intended for professional personnel in schools, school districts, and state departments of education.

This report is principally intended for professional personnel in schools, school districts, and state departments of education who share responsibility for analyzing and reporting on students' collective achievement test results and for using such results. We include teachers, school principals, school testing coordinators at the district level, other district administrators, and state education personnel in testing, evaluation, or administrative roles. Another audience that should benefit from considering the substantive and interpretive issues raised in this report (although they might be somewhat less interested in some of the technical concerns raised) are school board members and other education policymakers at the school district and state levels. With these audiences in mind, we have attempted to avoid esoteric statistical jargon and to present important technical issues in settings and terms that lend themselves to practical application.

The Structure of the Report

The body of this report is composed of a series of vignettes that illustrate the analysis, reporting, and interpretation of disaggregated achievement test results. Each vignette is a small case study that describes a context in which achievement testing took place, the nature of the testing program used, the purposes that guided those who disaggregated students' collective test results, the critical choices made in disaggregating test results, and the ways in which resulting test data were displayed, reported, and interpreted. In each vignette, the case study portrayed is used as a vehicle for raising and discussing important technical and interpretive issues that materially affect the usefulness and validity of reports on students' collective school achievement. Table 1 provides an overview of each vignette in terms of its context and the major questions it addresses.

Table 1. Context and major questions addressed in each of the vignettes

Vignette #1, Page 9	
Context:	School District: Lincoln County Subjects Tested: Reading, Writing, Mathematics Grades Tested: 4, 6, 8 Perspective: District Testing Committee
Questions:	<ul style="list-style-type: none">• How can test results be disaggregated to better understand the factors associated with students' performances?• How can students' test results be displayed to convey interesting features of their performances?• How large must a group of students be to support reporting their collective test performances?• How trustworthy are differences among the average test performances of students in various groups?• What conclusions can one reach concerning the causes of students' test results?

Vignette #2, Page 40

Context:

School: Harold Howe II High School
Subjects Tested: Algebra I, Biology, English, U.S. History, Social Studies
Grades Tested: At end of course, regardless of high school grade
Perspective: Mathematics Curriculum Task Force

Questions:

- How can test results be disaggregated to better understand the factors associated with students' collective progress across two school years?
- How can students' test results be compared across two school years, to illuminate a school's progress in achieving statewide achievement goals?
- How large must a group of students be to support reporting their collective test performances?
- How trustworthy are differences among the average test performances of students in successive school years?

Vignette #3, Page 62

Context:

State: Euphoria
Subjects Tested: Reading, Mathematics, Writing
Grade Tested: 6
Perspective: Office of Testing and Assessment, State Department of Education

Questions:

- How can students' test results be analyzed across four years, so as to illuminate medium-term trends?
- How can test results be disaggregated to better understand the factors associated with students' collective progress across four school years?
- When are differences in average achievement across years statistically reliable and educationally meaningful?
- What factors can contribute to cross-year trends in student achievement?
- How can trends in students' collective test results be modeled?

In a report of finite length it is impossible to include illustrations that encompass every conceivable situation and circumstance. You will probably find that the cases described in this report include some circumstances that are quite similar if not identical to yours, but others that differ materially. For example, the school district described in a vignette might be larger or smaller than yours, might differ from yours in terms of its racial and ethnic composition, or might include or exclude some groups of students who are either missing from or prevalent in your school population. Never mind. It is the principles and strategies conveyed through the vignettes that are the “real beef” of this report, and they will apply regardless of the degree to which your school or school district matches those used for illustration.

It is the principles and strategies conveyed through the vignettes that are the “real beef” of this report.

The principles discussed and illustrated in this report should be applicable to the full range of disaggregations called for in the Title I legislation.

We have used real student achievement test data, graciously provided by several states and school districts, in all of the vignettes in this report. However, our vignettes are works of fiction. Although they make use of real achievement test data, the results portrayed in the vignettes do not represent those found in any real school, school district, or state. Since states and school districts have not yet begun collecting the information needed to address all Title I requirements — they might not test in all required grade spans or subjects or collect information needed to support all required disaggregations of achievement data — our examples are illustrative rather than exhaustive. We also have focused principally on issues related to data analysis, reporting, and interpretation, often ignoring such issues as requirements or desiderata for inclusion of important representatives or groups in decisions pertinent to meeting Title I requirements. However, the principles discussed and illustrated in this report should be applicable to the full range of disaggregations called for in the Title I legislation. Although the vignettes illustrate types of analysis that can be used to satisfy those requirements, they also illustrate investigations and interpretations that go beyond those requirements. As we noted earlier, we regard the Title I requirements as an important opportunity for educators in schools, school districts and state departments of education. They provide a framework for important and interesting analyses of students' achievement test data that can be used for evaluation, for planning, and for informing the public about the accomplishments and needs of the public schools. The vignettes reflect this perspective.

As an aid to initial reading and later reference, each vignette is structured in the same way. First, the school, school district, or state used for illustration is described in terms of its size and the composition of its student body. Second, the testing program that produced students' achievement data is described in terms of grade level and subject-matter focus, and in terms of score reporting and duration. Third, the objectives and purposes of professional personnel in analyzing and disaggregating students' achievement test results are discussed. Fourth, strategies for disaggregating, analyzing, reporting, and interpreting students' achievement test results are illustrated, compared, and critiqued. In the course of this discussion, a number of technical and interpretive issues are raised and, often, advice is provided concerning essential cautions and the most effective and prudent courses of action.

How to Use this Report

We hope this report will be used in several ways. First, we hope that members of our intended audiences will at least skim the report in its entirety to stimulate their thinking about the myriad ways students' achievement test results can be analyzed, portrayed, and interpreted. Second, we hope that the report will serve as a focus of discussion among education personnel with responsibility for analyzing, reporting, and interpreting students' collective test results, and by education personnel with policy-making responsibility in such venues as committee meetings, in-service workshops, and planning sessions. Finally, we hope that the report will provide a useful reference for education personnel who want to review particular technical or interpretive issues associated with their analysis, reporting, and use of students' achievement test results. The table below is intended to aid those readers who wish to locate or relocate information related to particular topics or issues.

at least skim the report in its entirety.

Table 2. A guide to topics of interest

Topic of Interest	A Bit of Detail	Pages
Identifying Questions	from teachers and principals	15, 16
	brainstormed by math committee	35
	raised by analysis	36, 37
	posed by legislature	51
Finding Information	ERIC	14, 17
	Regional Comprehensive Assistance Centers	17
	Title I	15
	State Department of Education	47
Definitions of Categories	importance of clear definitions and problems with inconsistency	53-54
Group Size	protection of privacy	28, 37
Displaying Data	box-and-whisker charts	17-18, 29, 31
	bar graphs	19-21, 26, 38-39, 43-46
	confidence intervals	26, 28
	tables	36, 37, 38-40
	advantages of graphs and tables	38, 46
	line graphs	52-53, 54-55, 56-59
Error	sampling error	22-23, 40
	measurement error	26-27
Comparing Data	percentages at performance levels	18-25, 43-47
	average scores for groups	25-29, 38-43
	group score distributions	29-31
	trends over time	55-59
Confidence Intervals	percentages at performance level	22-23, 55-56
	to compare percentages for groups	23, 24-25
	group averages	26-28
	averages over time	52-53
Statistical Analysis	test the difference between percentages for groups	23-24
	correlation vs. causation	24, 30
	frequency distribution	36-37
	t-test to test difference in average scores between groups	40-41
	relationship between sample size and statistical significance	41-42, 60
	effect size	42, 43
	linear regression analysis	59-62

A Word About Computing

The vignettes in this report make use of three computer programs:

- (1) Microsoft Excel is a spreadsheet program with many built-in statistical procedures that is part of the Microsoft Office Suite. Excel also produces graphs that are of publication quality that can be embedded in a word processing file. The program is available for Windows and Macintosh microcomputers.
- (2) SPSS is an extensive and sophisticated statistical analysis program that will do almost any statistical analysis that educators and educational researchers could desire. This program also is available for Windows and Macintosh microcomputers. It is menu driven, reasonably intuitive in its commands, and produces graphs that are of publication quality. SPSS graphs can easily be copied and pasted into word processing files. SPSS is available from SPSS, Inc., 444 N. Michigan Avenue, Chicago, IL 60611; Phone: (800) 543-2185; FAX: (800) 841-0064; or from the SPSS web site: <http://www.spss.com>. The educator's price for the SPSS base module (which is sufficient to conduct any of the analyses illustrated in this report) is currently \$495, but this price may increase. More sophisticated analyses can be conducted with additional modules that are separately priced.
- (3) DataDesk is another extensive statistical analysis program that is principally designed to facilitate graphical exploration of data. DataDesk is very easy to learn and use; it comes with a set of manuals that include a basic statistical analysis text. A demonstration version of DataDesk for Windows or Macintosh computers is available for downloading at no cost at the Data Description Inc. web site: <http://www.lightlink.com/datadesk>. The Data Description phone number is (800) 573-5121, and orders may be faxed to (607) 257-4146. The educator's price of DataDesk currently is \$390; the commercial price currently is \$650. Personnel in school districts and state departments of education will qualify for the educator's price. Please verify with Data Description all current prices before ordering, as they are subject to change.

There are a number of statistical computer program alternatives that are less expensive than either SPSS or DataDesk. One is the VisTa program, provided free of charge by Professor Forrest Young at the University of North Carolina at Chapel Hill. The program and its documentation can be downloaded from Professor Young's web site at <http://forrest.psych.unc.edu/research/vista.html>. The program runs on Windows and Macintosh microcomputers, has an intuitive interface and comes with excellent documentation. It will do all of the statistical analyses described in this report.

For a wonderful resource on statistical analysis computer programs, on-line statistical analysis courses, and links to a wide variety of free and commercially-available computer programs for statistical analysis, go to the web site: <http://www.math.yorku.ca/SCS/StatResource.html>

A school district may be tempted to hire a computer programmer to create a customized program for disaggregation of Title I achievement test data. We strongly recommend that this alternative be avoided since it would be costly, the program likely would require frequent maintenance, and small changes in Title I requirements could render the program obsolete.

school district may be tempted to hire a computer programmer to create a customized program for disaggregation of Title I achievement test data. We strongly recommend that this alternative be avoided.

Vignette 1: The Lincoln County School District

Important Questions Addressed in this Vignette:

- How can test results be disaggregated to better understand the factors associated with students' performances?
- How can students' test results be displayed to convey interesting features of their performances?
- How large must a group of students be to support reporting their collective test performances?
- How trustworthy are differences among the average test performances of students in various groups?
- What conclusions can one reach concerning the causes of students' test results?

The Lincoln County School District

Lincoln County is in a southeastern state that operates a statewide assessment program in reading, mathematics, and writing for students in Grades 4, 6, and 8. The Lincoln County schools enroll just under 6,700 students in 12 elementary schools and four middle schools. The elementary schools enroll 679 fourth-graders. The middle schools have sixth-grade and eighth-grade enrollments of 733 and 714, respectively.

The county is predominantly rural but has two large towns that serve as local centers of commerce, entertainment and services for a population that, for the most part, supports itself through agricultural production and sales. Truck farming, tobacco farming and pork production and processing predominate.

Lincoln County is socio-economically and racially diverse. During the 1995-96 school year, almost one-fourth of the students received free or reduced-price lunch. About one-fifth of the students identified themselves as African American, 14 percent identified themselves as Hispanic, and 45 percent identified themselves as white, non-Hispanic. Ten percent of the students responded "Other" to a question on racial and ethnic background that included all of the categories typically used by the U. S. Office for Civil Rights. Less than two percent identified themselves as Asian American and less than one percent classified themselves as Native American¹.

Only two percent of Lincoln County students were classified as limited-English proficient, so almost all students in the district were regarded as fluent in the English language. Lincoln County offers special education programs that enrolled nine percent of its students and a program for gifted students that enrolled seven percent of the total student body during the 1995-96 school year.

The State's Testing Program

As noted earlier, Lincoln County students in Grades 4, 6, and 8 are tested in reading, mathematics, and writing through an assessment program operated by the state. The state also is developing assessments in these three areas for students in Grade 10. Multiple-choice items predominate in the state's mathematics test, but to answer a few items, students must complete an answer grid or provide brief written solutions to problems presented in an open-ended format. The state's reading test also contains a mixture of multiple-choice and short-answer, open-ended items, with more than 90 percent of the items in multiple-choice format. To complete the state's writing test, students must respond to two essay questions, each of which is scored holistically on a six-point

¹About eight percent of Lincoln County students did not identify themselves by racial or ethnic group nor did they choose the "Other" category that was made available in a question concerning racial or ethnic-group membership. In subsequent analyses involving racial and ethnic-group membership, achievement test data will be used only for students who defined themselves as members of one of the categories listed here, including the "Other" category.

Lincoln County is socio-economically and racially diverse.

As noted earlier, Lincoln County students in Grades 4, 6, and 8 are tested in reading, mathematics, and writing.

Dr. Crawford asked the Committee to determine how assessment data for Lincoln County's fourth, sixth, and eighth-graders could be analyzed and reported to produce the most useful information.

scale. Students' scores on the two essays are summed to produce an overall writing test score. Since the state's assessment program was completely redesigned in 1994, the current program has been in operation only since the 1995-96 school year. Therefore, only two years of test data are available for students in grades 4, 6, and 8.

The District Testing Committee

When the state's assessment program was redesigned, Dr. Molly Crawford, the Lincoln County Superintendent of Schools, appointed a District Testing Committee composed of the principals of two of the district's elementary schools (Mr. James Bluford and Dr. Janet Clinton), one middle-school principal (Ms. Ann Salisbury), one elementary-school teacher (Ms. Aleshya Jackson), one middle-school teacher (Mr. Bobby Pinnix), and a knowledgeable and interested parent who agreed to serve as a representative of the school system's Parent-Teacher Association (PTA), Ms. Joyce Mineka.

Dr. Crawford asked the Committee to determine how assessment data for Lincoln County's fourth, sixth, and eighth-graders could be analyzed and reported to produce the most useful information for the following:

- 1) instructional planning by principals;
- 2) instructional planning by teachers;
- 3) school board policymaking;
- 4) dissemination of information to the county's parents and taxpayers; and
- 5) analysis and reporting in accordance with the requirements of Title I of the Improving America's Schools Act.

Dr. Crawford agreed to meet with the Committee at their request, on a consultative basis, and agreed to seek input from the U. S. Department of Education's Regional Comprehensive Assistance Center that served the state. Finally, Dr. Crawford suggested that the Committee ask the district's Technology Specialist for access to the Educational Resources Information Center's (ERIC) Assessment and Evaluation web site, where members could search for research articles on analyzing and reporting test results.

The Committee's Initial Deliberations and Actions

When the District Testing Committee first met, it took a bit of time to determine how to begin its work. Since part of Dr. Crawford's charge was to propose analysis and reporting of assessment data to satisfy the requirements of Title I, Ms. Jackson suggested that the Committee obtain a copy of the Title I requirements and use them as a point of departure. With the Committee's agreement, Ms. Jackson offered to obtain this from Dr. Crawford and to present the requirements to the Committee at its next meeting.

Mr. Bluford suggested that the three principals on the Committee might form one working group and the two teachers on the Committee might form another, to list the kinds of test results that they would find most useful for instructional planning. That idea was endorsed, and the two groups agreed to meet at least once during the following two weeks, prior to the next meeting of the full Committee.

Mr. Pinnix suggested that it would be helpful if each Committee member talked about their familiarity with school testing and with the analysis of test data. He said that he had completed a course in testing and measurement during his Master of Education program, and he thought he still had his textbook. Ms. Jackson said she had learned a bit about testing students in a course on educational psychology, and she was pretty familiar with the norm-referenced tests the district had used prior to the introduction of its new testing program. She felt comfortable telling parents about their children's test results. Dr. Clinton had finished her Doctor of Education program at the state university the previous year. She had taken a course in educational measurement and two courses in applied statistics as a part of that program, and she had used students' test results

as part of her doctoral dissertation. Although Ms. Salisbury and Mr. Bluford hadn't had any courses in educational measurement, each was familiar with the norm-referenced test results produced under the district's old testing program. Ms. Mineka stated that she hadn't any formal background in educational measurement and preferred to serve as a conduit between the Committee and Lincoln County's PTA.

Based on Dr. Clinton's recent use of test data in her dissertation research and her background in statistics, the other Committee members thought it would be good if Dr. Clinton served as Chair of the Committee. Her duties would include convening meetings, maintaining a "library" of reference materials and reporting to Dr. Crawford on the Committee's progress. Dr. Clinton agreed, with the proviso that she not also be asked to serve as Secretary of the Committee. Mr. Bluford volunteered to assume that role. Having defined its initial work plan, the Committee agreed to meet again in two weeks.

The Committee's Second Meeting

As the first order of business at the Committee's second meeting, Ms. Jackson reported on what she had learned about the Title I requirements for reporting student assessment data. She told the Committee that she had obtained from Dr. Crawford a copy of the applicable Federal Rules and Regulations that had been published in the *Federal Register*, Volume 60, Number 127, on July 3, 1995. Under Paragraph 200.4, the rules said that each state must assess students annually in at least mathematics and reading/language arts, and that test results must be analyzed separately for each school district and each school in the state. The Rules and Regulations further said that students must be assessed at least once during Grades 3 through 5, during Grades 6 through 9, and during Grades 10 through 12. Most important, the Rules and Regulations said that assessment results must be analyzed and reported separately for each school district and school for students categorized by gender, each major racial and ethnic group, English proficiency status, migrant status, students with disabilities compared to students without disabilities, and for economically disadvantaged students compared with those who are not economically disadvantaged.

The Committee members found the Rules and Regulations to be very helpful in defining groups of students for whom assessment results must be reported and compared.

Next, Ms. Jackson and Mr. Pinnix reported on their progress in defining assessment results that would be useful to teachers for use in instructional planning. They not only had met together, but each had met with other teachers in their respective schools. Through their discussion and consultation with colleagues, they developed the following list of suggestions:

- (1) Assessment results should be summarized for the students in each teacher's class.
- (2) One part of the summary should tell teachers how well their class was doing overall, in the three subjects tested.
- (3) Another part of the summary should let teachers know whether all students were performing at about the same level on the test, or whether some students were doing very well while others were doing poorly.
- (4) Just reporting a single overall reading score, a single overall mathematics score and a single overall writing score for a class wasn't as useful for instructional planning as reporting more-detailed scores. For example, in reading, it would be better to know how students in the fourth grade were doing with items that required them to identify the main idea of a reading passage versus items that required them to apply or extend the concepts discussed in a passage.
- (5) Teachers needed a frame of reference for interpreting their students' assessment results. For example, they needed to know what score on each test should be considered "good" performance and what should be considered "excellent" performance. They also wanted to know how well students in a given grade were doing this year, compared to how they did last year on the tests.

Assessment results must be analyzed and reported separately for each school district and school for students categorized by gender, each major racial and ethnic group, English proficiency status, migrant status, students with disabilities compared to students without disabilities, and for economically disadvantaged students compared with those who are not economically disadvantaged.

the state classified each student's test performance as not yet proficient, almost proficient, proficient, or advanced."

The Committee agreed that this was a very useful list of suggestions. Dr. Clinton, in particular, said that each of the teachers' suggestions had implications for how the assessment data should be analyzed and reported, as did the list of Title I Rules and Regulations that Ms. Jackson had described.

As Committee Chair, Dr. Clinton had called the State Department of Education to obtain information on how test results were made available to school districts. She was told that her school district was sent a computer file containing each student's name, grade, school name, the student's classification in terms of gender, racial or ethnic group, English language proficiency status, participation status in a subsidized school lunch program, and participation status in special education programs or programs for gifted students. Each student's record also listed the student's total score on each of the three achievement tests and the student's classification in terms of statewide performance standards on the three tests. The state classified each student's test performance as "not yet proficient," "almost proficient," "proficient," or "advanced," consistent with the Title I requirements for reporting the status of the state's progress in enabling children to achieve its performance standards. Although she agreed with the teachers' fourth suggestion (that more-detailed test results be reported for each subject), Dr. Clinton thought it probably wouldn't be possible to obtain that information, based on what she had been told about test information the state reported to each school district.

Mr. Bluford and Dr. Clinton had met to discuss the kinds of assessment results that principals would find useful for instructional planning. It is not surprising that their list was quite similar to the one compiled by the teachers' working group. Their suggestions were as follows:

- (1) Summarize test data for students in the fourth, sixth, or eighth grade for the entire school for each subject tested, so that each principal could learn how students in each school were doing, compared to the district as a whole, and compared to students in the same grade in other schools.
- (2) Summarize the test data so that principals could compare their students' test performances this year with how they did last year — providing some indication of yearly progress.
- (3) Summarize the data so that it would be possible to see how much students' test performances differed. That is, it would be helpful to know whether all students scored about the same or, if not, the range of scores earned by the weakest and the strongest students.
- (4) Report for their school — and for the district as a whole — the percentage of students whose test performances would be considered "poor," and corresponding percentages for those whose performances would be considered "mediocre," "good," and "excellent."
- (5) Summarize the test data so that principals would know how various groups of students in their schools were doing on the assessment. For example, report separately for girls and boys, for students in programs for the gifted, and for students receiving free or reduced-price lunch.

The Committee noted the similarity of the desires expressed by the teachers and the principals. They felt they were off to a good start, but Ms. Salisbury suggested that they review their charge from Dr. Crawford to see what else they should be doing. When that was done, it was apparent that the Committee so far had focused on three parts of their charge, but had not yet grappled with the other two: determining how assessment results should be reported to aid school board policymaking, and disseminating information to the county's parents and taxpayers.

Dr. Clinton suggested that the Committee proceed one step at a time. She thought that school board members and parents might find it difficult to suggest how students' assessment data should be reported for their use unless they had some examples of possibilities. With some examples, members of both groups could be asked how useful they found the reports and what else they would like to have. Dr. Clinton volunteered to develop some examples since she had recently used test results in her dissertation and felt she could call upon her professors at the uni-

versity for help if she needed them. She decided to talk to Dr. Crawford about working with the district's Computer and Information Services Coordinator, Jim Kelly, to produce some sample assessment reports, based on the Title I requirements, the teachers' ideas and the principals' ideas. In the meantime, she asked that other Committee members work with the District's Technology Coordinator in searching the ERIC system to find examples of assessment reports produced by other school districts and to contact the Regional Comprehensive Assistance Center that served their state. The Committee agreed to meet again two weeks later.

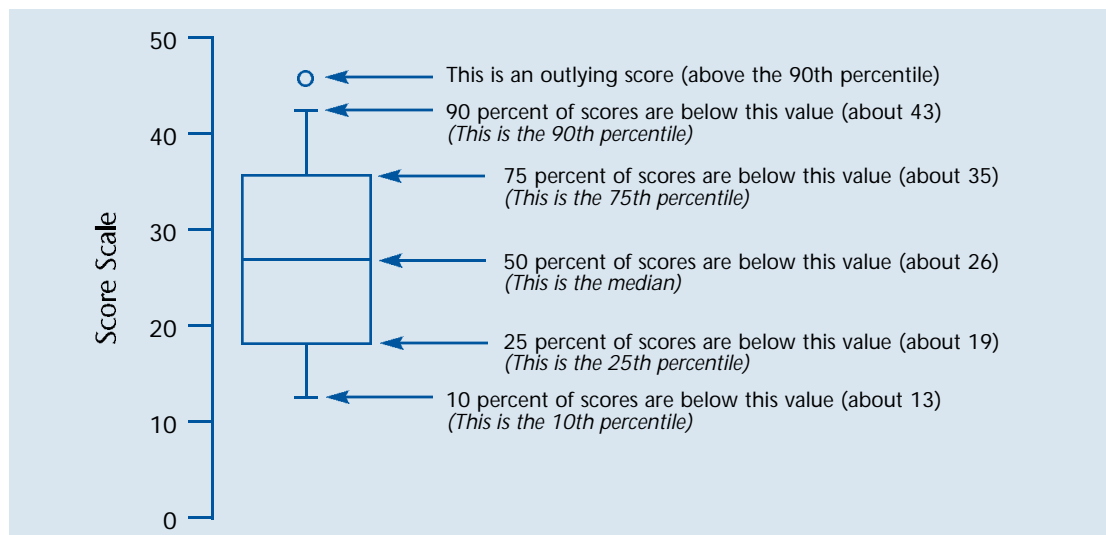
The Committee's Third Meeting

Most of the Committee's third meeting was devoted to review and discussion of examples of achievement data analyses and displays that Dr. Clinton and Mr. Kelly had prepared. It was a lively meeting that Committee members found to be instructive and stimulating.

When they reviewed the achievement test data available for Lincoln County's fourth, sixth, and eighth-graders, Dr. Clinton and Mr. Kelly realized that it could be disaggregated in many different ways. For the purpose of generating examples, they decided to produce results only for specified groups of students who were tested throughout the district. They reasoned that similar analyses could be conducted for students in individual schools if Committee members felt that a particular analysis was useful and informative.

The data analyses produced by Dr. Clinton and Mr. Kelly were of four basic types. First, they displayed the percentages of students in various groups — for example, boys and girls — who earned test scores that categorized them at various score levels, such as “not yet proficient,” “almost proficient,” “proficient,” and “advanced.” Second, they computed and displayed the average test scores earned by students in various groups. Third, they constructed graphs — known as “box-and-whisker charts” — that facilitated comparisons of essential features of score distributions for various groups of students. Finally, they computed statistics that allowed readers to see how much the average test score of a group would be expected to vary just due to random sampling fluctuation, and because the tests used to measure student achievement were not perfectly reliable. These statistics were essential in deciding whether differences between the average performances of groups of students were statistically trustworthy, or likely to reflect nothing more than the kinds of random fluctuations one would expect when students are sampled from a larger population of potential examinees. Again, more will be said about this issue when these statistics are discussed later.

A “box-and-whisker chart” is the wonderful invention of a man named John Tukey, the father of Exploratory Data Analysis. It is a graph that summarizes the most important features of a distribution of scores in a way that is very easy to read and understand. Furthermore, it greatly facilitates comparison of the score distributions of two or more groups, since box-and-whisker charts for several groups can be drawn side by side. A box-and-whisker chart looks like the drawing that follows. Note that essential features of the box-and-whisker chart have been highlighted:



These statistics were essential in deciding whether differences between the average performances of groups of students were statistically trustworthy.

In a box-and-whisker chart, the width of the box is not important, but the height of the box conveys useful information.

- The middle half of any distribution of scores falls between the 25th percentile and the 75th percentile. These are the values that, respectively, define the lower fourth of a distribution and the lower three-fourths of a distribution. In the box-and-whisker chart illustrated above, the bottom of the box (the 25th percentile) falls at about 19 on the score scale, and the top of the box (the 75th percentile) falls at about 35 on the score scale. If the chart illustrated a distribution of test scores, we could say that a fourth of the students earned scores at or below 19, that three-fourths earned scores at or below 35, and half the students scored between 19 and 35.
- The horizontal line inside the box falls at about 26 on the score scale. This horizontal line defines the 50th percentile of the score distribution (also called the median). This tells us that half the students earned test scores at or below 26 — and, therefore, half earned scores above 26. So 26 is one answer to the question, “What is the most typical score for this group of students?”

The vertical lines that extend above and below the box are called the whiskers.

- The bottom whisker begins at the 25th percentile and ends at the 10th percentile. In this example, the 10th percentile is a score of about 13, indicating that 10 percent of the students in this group earned test scores at or below 13.
- The top whisker begins at the 75th percentile and ends at the 90th percentile. The 90th percentile is at a score of about 43 in this example, so we know that 90 percent of the students earned test scores at or below 43 and only 10 percent earned scores above 43. One student earned a score of 46, indicated by the circle that appears above the upper whisker. In a box-and-whisker chart it is customary to show extreme scores (those beyond the 10th and 90th percentiles) with individual circles.

The Committee reviewed and discussed each type of data display, often asking questions and commenting on the usefulness of the display for various purposes. A summary of their conversation — and the data displays considered by the Committee — follows.

Dr. Clinton and Mr. Kelly explained that the first six data displays were in the form of bar charts. Labeled Figure 1 through Figure 6, these graphs show the percentage of Lincoln County fourth-graders, overall, and by group, whose Mathematics test scores placed them in the various achievement categories designated by Lincoln County’s state. These figures permit Lincoln County’s educators to determine how well their fourth-graders performed on the mathematics test overall and how the distributions of performances of students in different groups compared. This kind of disaggregation of achievement test data is useful in determining whether students in any group are doing materially better — or worse — on the state’s achievement tests than members of other groups. Figures 1 through 6 were as follows:

These figures permit Lincoln County’s educators to determine how well their fourth-graders performed on the mathematics test overall and how the distributions of performances of students in different groups compared.

Figure 1. Percent of Lincoln County Schools' fourth-graders, by mathematics proficiency category (1995-96)

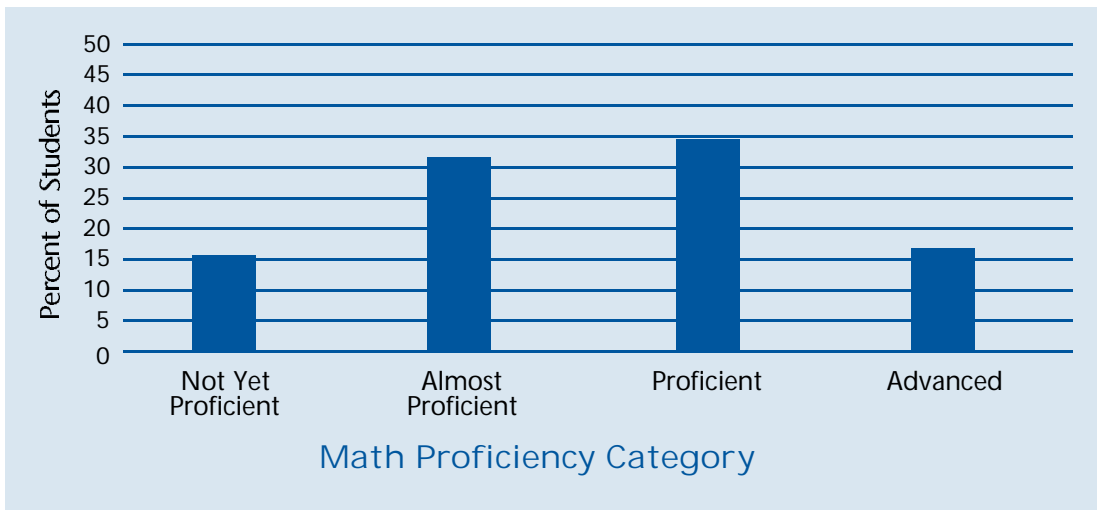


Figure 2. Percent of Lincoln County Schools' fourth-graders, by mathematics proficiency category within gender (1995-96)

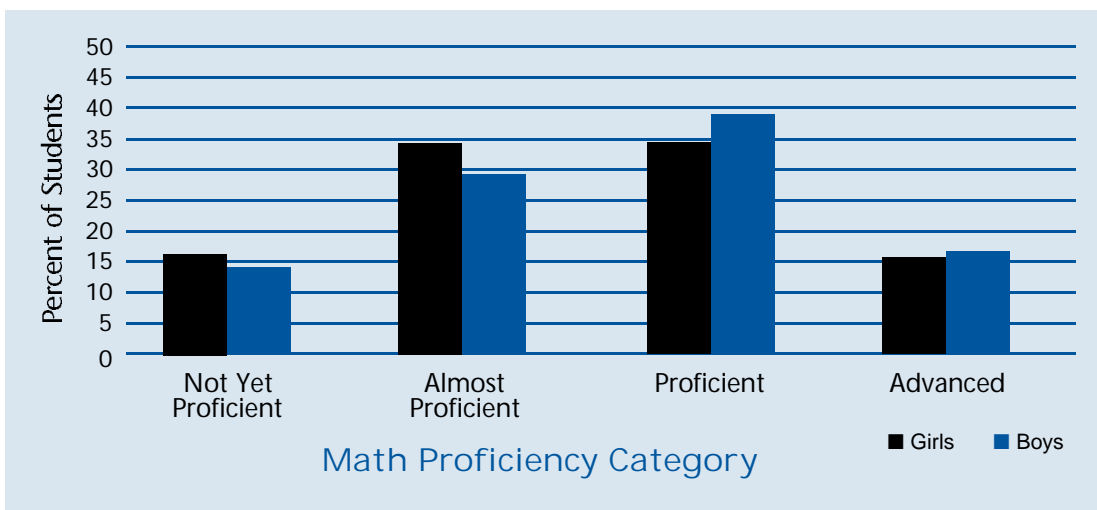


Figure 3. Percent of Lincoln County Schools' fourth-graders, by mathematics proficiency category within racial/ethnic group (1995-96)

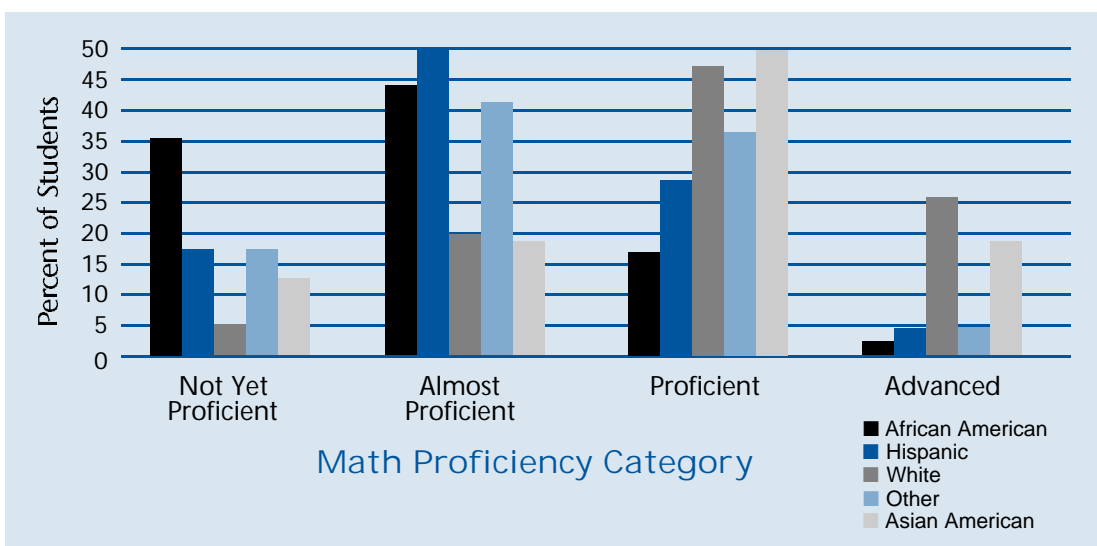


Figure 4. Percent of Lincoln County Schools' fourth-graders, by mathematics proficiency category within gifted and regular programs (1995-96)

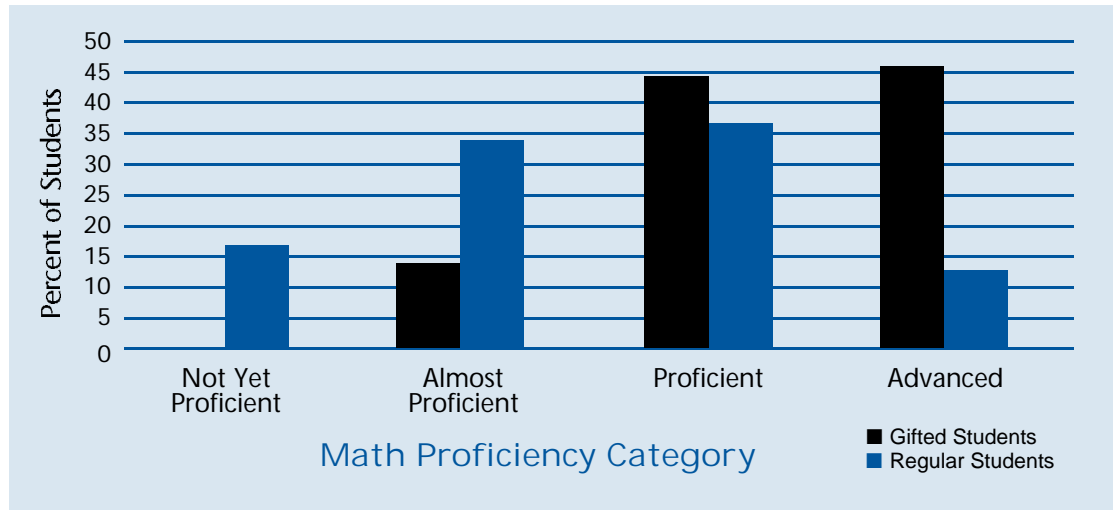


Figure 5. Percent of Lincoln County Schools' fourth-graders, by mathematics proficiency category within school-lunch program (1995-96)

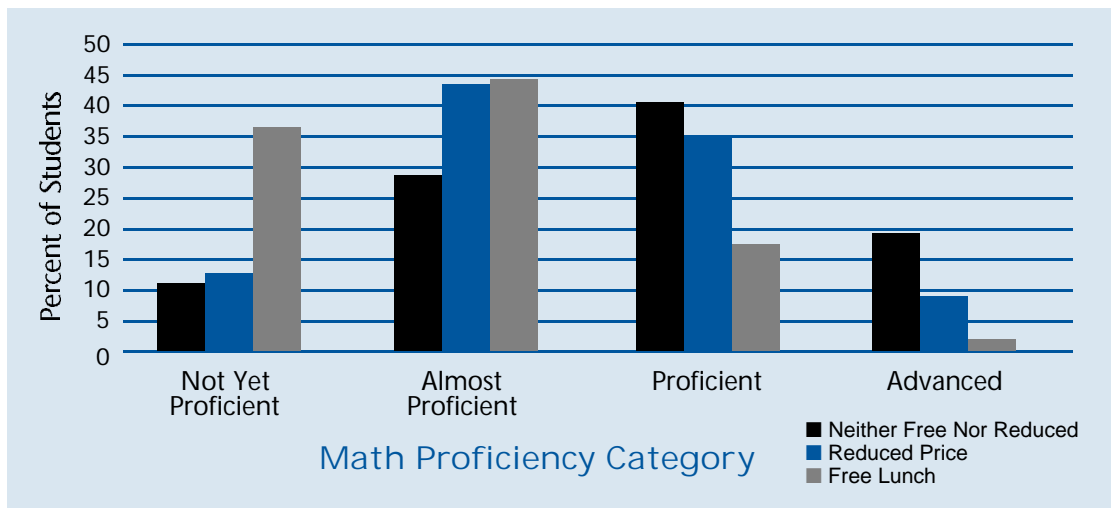
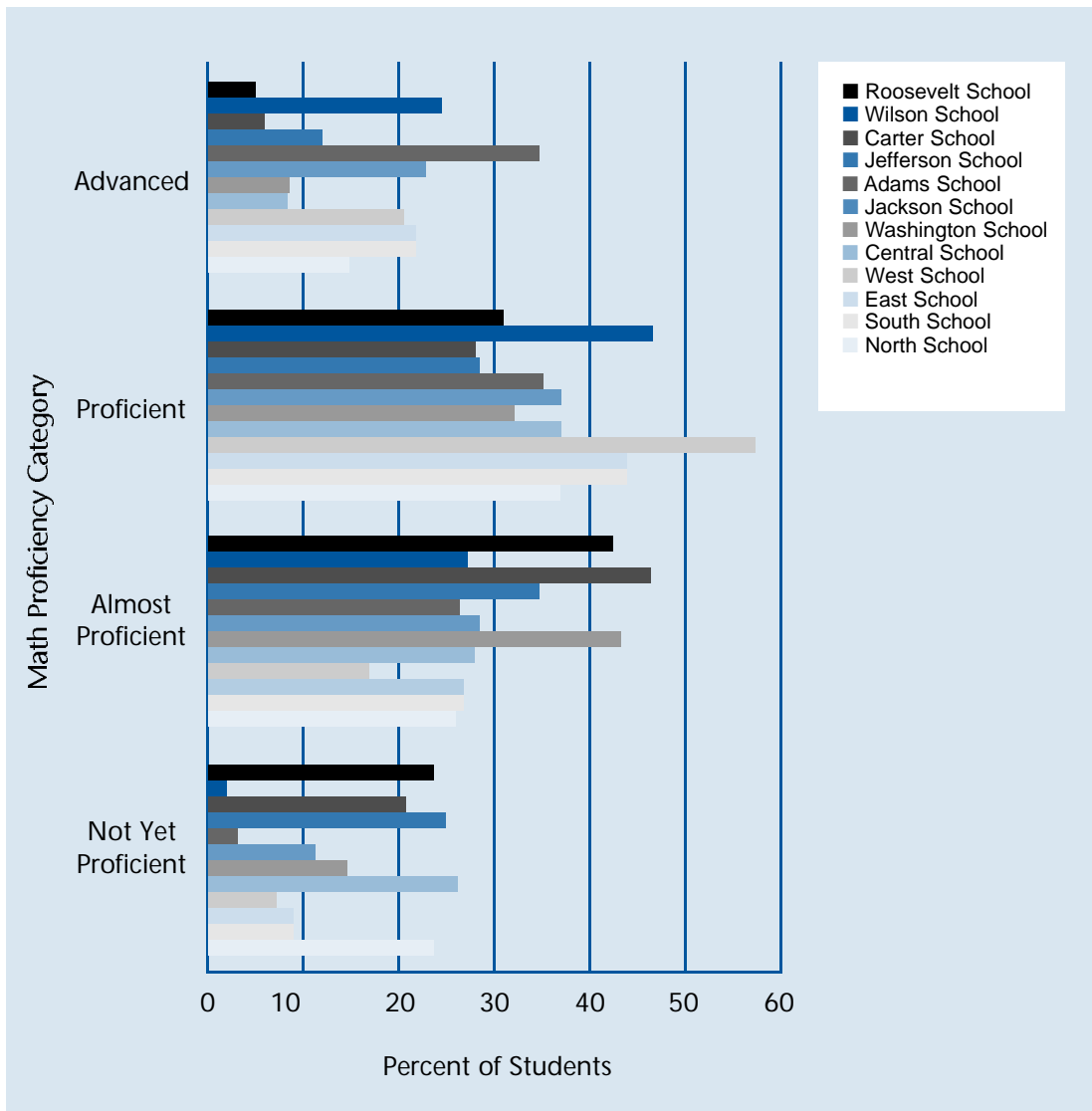


Figure 6. Percent of Lincoln County Schools fourth-graders by mathematics proficiency category within school (1995-96)



Although Figures 1 through 6 did not provide all of the answers to the questions posed by the teachers and the principals on the Committee, they were of substantial interest to Committee members. Figure 1 illustrates the distribution of mathematics test performances for all fourth-graders who were tested in Lincoln County. Upon reviewing this distribution, Mr. Bluford remarked, “There’s good news and bad news.” He pointed out that more than half the district’s fourth-graders had mathematics scores in the Proficient and Advanced categories, but almost as many earned scores that placed them in the Not Yet Proficient or Almost Proficient categories. Mr. Bluford concluded that the Lincoln County Schools still had a lot of work to do, especially with the 15.5 percent of fourth-graders who were classified as Not Yet Proficient.

Dr. Clinton agreed that having more than 15 percent of the 1995-96 fourth-graders in the Not Yet Proficient category was troubling, but she encouraged the Committee to consider several related issues. First, she suggested that the Committee pay close attention to the results shown in the other figures that she and Mr. Kelly had prepared. These figures show the achievement-category classifications of fourth-graders in various groups. By reviewing them, the Committee might learn whether some groups of students were more likely than others to be classified as Not Yet Proficient. Second, she raised a statistical question that hadn’t occurred to any of the other Committee members. That question had to do with generalizing the result that Mr. Bluford had identified as troubling.

Mr. Bluford remarked, “There’s good news and bad news.”

is important to ask whether these results could be considered typical for the Lincoln County Schools, or whether one would expect the results for another school year to be very different.

One should ask how much we would expect the percent of fourth-graders who are classified as Not Yet Proficient to vary from year to year, just due to random sampling fluctuation.

Although there was no question that 105 of the 679 Lincoln County fourth-graders who had completed the state's mathematics test during the 1995-96 school year had been classified as Not Yet Proficient, the results for this 15.5 percent of the tested fourth-graders represented just one year in the life of the Lincoln County Schools. It is important to ask whether these results could be considered typical for the Lincoln County Schools, or whether one would expect the results for another school year to be very different. Dr. Clinton was asking both an interpretive question and a statistical question.

Interpretively, one could regard the 1995-96 data as definitive for that school year. That is, since all eligible fourth-graders had been tested, there can be no question that the results observed tell the story of what happened to the Lincoln County Schools' fourth-graders when they encountered the state's mathematics test in 1995-96. A statistician would say that the 15.5 percent finding, in answer to the question, "What percent of Lincoln County fourth-graders earned mathematics scores in 1995-96 that placed them in the Not Yet Proficient category?" is a *population parameter*². Figures that define results for an entire population of individuals are regarded by statisticians as immutable. Since all eligible students were tested, rather than a sample of students, this population value will not fluctuate statistically. Its status as a population value earns it the title "population parameter."

But from another perspective, one could argue convincingly that the 15.5 percent result was not only a consequence of the quality of education provided to fourth-graders in Lincoln County, but occurred in part because of the particular students who happened to be fourth-graders in the school district during the 1995-96 school year. From this perspective, the 679 fourth-graders who were enrolled during the 1995-96 school year are regarded as a sample from the population of fourth-graders who enroll in the Lincoln County Schools across the years. When next year's test results are compared to this year's, any differences will be due not only to changes in the quality of education offered in Lincoln County, but also in part to differences between the backgrounds of students who happened to be enrolled in the fourth grade during the two school years. If the 1995-96 fourth-graders are considered to be a sample drawn from a larger population of fourth-graders who might be enrolled across the years, the 15.5 percent who were classified as Not Yet Proficient in 1995-96 would be regarded as a *sample statistic* rather than a population parameter. Sample statistics are not immutable. They vary across selected samples.

If the 679 fourth-graders who were enrolled during the 1995-96 school year are considered to be a sample from the population of Lincoln County fourth-graders who enroll across the years, one should ask how much we would expect the percent of fourth-graders who are classified as Not Yet Proficient to vary from year to year, just due to random sampling fluctuation. Or, as statisticians might put it, how much will the 15.5 percent vary, due to chance? Fortunately, well-known statistical theory provides an answer to this question. It is possible to compute what is called the *standard error* of the percent of fourth-graders who would be classified as Not Yet Proficient and, from that standard error, to construct what is called a *confidence interval* around the percentage. A standard error of a sample statistic indicates how much that statistic varies across samples, and a confidence interval around a sample statistic tells us limits within which we would expect to find the population parameter that a sample statistic represents. A more complete discussion of these issues can be found in a very readable book by Jaeger (1990)³ that contains no equations, and in any number of more technical statistics books, such as Glass and Hopkins (1997)⁴. Appendix A⁵ provides steps for computing the standard error of the 1995-96 percentage and the confidence interval around that percentage.

Dr. Clinton pointed out, from the computations in Appendix A, that just due to random sampling fluctuation from year to year, the Committee could expect the percent of fourth-graders

²In the balance of this report, statistical terms, when first introduced, will be printed in italics. They will either be defined in the text or in a footnote.

³Jaeger, R. M. (1990). *Statistics: A Spectator Sport*. Newbury Park, CA: Sage Publications.

⁴Glass, G.V. and Hopkins, K. (1997). *Statistical Methods for Psychology and Education*. New York: Prentice-Hall.

⁵In the balance of this report, any computations will be presented in appendices. The material in an appendix can be ignored by those who are not interested in computational details. Others might find it useful to learn how various statistical results are computed. In either case, it is the underlying conceptual issues discussed in this report that are of greatest importance.

with mathematics achievement in the Not Yet Proficient category to range from just under 13 percent to just over 18 percent. She pointed out that this result would be particularly useful when the Committee compared the fourth-graders' mathematics results for 1996-97 to those for 1995-96.

Ms. Jackson was particularly interested in the results shown in Figure 2. She commented that she had always heard that boys did better in mathematics than girls, but was still a bit surprised to see the differences between the percent of girls and boys who were classified as Proficient or Advanced. The percentages in the Proficient category were 34.2 for girls and 39.2 for boys, a difference of five percent. The percentages in the Advanced category were closer, but still favored boys: 16.9 percent for boys versus 15.2 percent for girls.

Dr. Clinton suggested that it might be worth seeing whether the difference in percentages of boys and girls in the Proficient or Advanced categories was statistically reliable. That is, whether the difference could be attributable to random sampling fluctuation or would likely be found for populations of boys and girls across school years. Here again, an important conceptual question must be addressed, along with the associated statistical question. It is indisputable that, for the 1995-96 school year, a higher percentage of fourth-grade boys than girls scored in the Proficient or Advanced categories. If we ask whether the observed difference is statistically reliable, we are really asking whether we would generally expect to see differences favoring boys across school years or, alternatively, whether the observed result for the 1995-96 school year could merely be a consequence of the kind of fluctuation in percentages one would expect for samples of the sizes tested.

Statistically, there are several ways of addressing this question. One approach would be to compute confidence intervals around each sample percentage, much as was done in the example just completed. Another approach would be to test the statistical significance of the difference between the percentages of fourth-grade boys and girls whose mathematics scores placed them in the Proficient or Advanced categories. We'll illustrate both procedures in Appendix B. However, we find the confidence interval approach to be more informative, since it provides information on the range of sampling variability associated with each sample percentage.

Using the results in Appendix B, Dr. Clinton pointed out that the apparent advantage in mathematics performance enjoyed by fourth-grade boys, at least in terms of the percent who were classified as Proficient or Advanced, merely could be due to random sampling fluctuation rather than any real mathematics advantage enjoyed by boys and was not statistically reliable. In other words, one might well expect a difference favoring girls in some other school year even though the 1995-96 Lincoln County mathematics results for fourth-graders definitely favored boys.

Upon reviewing Figure 3, Committee members concluded that fourth-graders in some racial and ethnic groups, particularly white, non-Hispanics and Asian Americans, did materially better on the state's mathematics test than did students in other racial and ethnic groups, particularly African American and Hispanic fourth-graders. Eight in ten African American fourth-graders were classified as Not Yet Proficient or Almost Proficient on the state's mathematics test in 1995-96, and the same was true of more than two-thirds of Hispanic fourth-graders. Mr. Pinnix raised the issue of statistical reliability; he was especially concerned about overinterpreting the results for Asian American students because he thought the size of this group must be quite small.

When the data were checked, Dr. Clinton agreed with Mr. Pinnix's concern. She noted that only 16 fourth-graders had identified themselves as Asian American. Therefore, the 50 percent of those students who were classified as Proficient only constituted eight students and the 19 percent classified as Advanced only constituted three students. With numbers this small, the percentages could not be regarded as reliable. That is, although the observed percentages for the 1995-96 school year were certainly correct, very different percentages might be found for Asian American students in another school year. For example, a 95 percent confidence interval around the 50 percent of Asian American students who were classified as Proficient would range from a lower limit of 26 percent to an upper limit of 75 percent. From the data at hand, we can be 95 percent confident that the percent of an underlying population of Asian American fourth-graders whose mathematics scores placed them in the Proficient category would range from about a fourth to about three-fourths. This is, indeed, a very wide interval.

The apparent advantage in mathematics performance enjoyed by fourth-grade boys, at least in terms of the percent who were classified as Proficient or Advanced, merely could be due to random sampling fluctuation.

will be important for the Lincoln County schools to follow up on this finding.

Correlation does not necessarily indicate causation.”

Nonetheless, the data in Figure 3 make clear that the 1995-96 mathematics performance of African American and Hispanic fourth-graders is materially lower than that of white, non-Hispanic fourth-graders. It will be important for the Lincoln County schools to follow up on this finding to see if it can be isolated to a particular set of schools, to see whether it is associated with a policy that fosters ability grouping, or to see whether some other school district policy or condition might be altered to improve the mathematics performance of fourth-graders with these racial and ethnic backgrounds.

The advantage of disaggregating achievement test scores is apparent in the results shown in Figures 1 and 3. Although a school district might not be content knowing that almost half its fourth-grade students scored in the Not Yet Proficient or Almost Proficient categories, it should be even less content knowing that overwhelming majorities of its African American and Hispanic fourth-graders did so.

The Committee regarded the results shown in Figure 4 as expected, and as a confirmation of the school district's identification of fourth-graders for its gifted education program. Members of the Committee were relieved to find that nine out of ten fourth-graders classified as gifted had mathematics scores that placed them in the Proficient or Advanced categories. A contrary finding would have called into question either the validity of the state's mathematics test or the process by which students were identified for the school district's gifted education program.

Ms. Jackson found the results shown in Figure 5 to be quite interesting. Her interpretation of this disaggregation of the school district's mathematics test results for fourth-graders linked students' socio-economic background and their test performances. She correctly pointed out that race, ethnicity, and economic status were strongly associated among students enrolled in the Lincoln County schools. She therefore questioned whether the association shown in Figure 3, between race, ethnicity, and mathematics test performance, really represented an association between students' economic status and their test performance, as shown in Figure 5.

Ms. Jackson's observation led to a spirited and illuminating discussion. Dr. Clinton repeated a rule that her statistics professor at the State University had "drilled into her head." Her graduate class was asked to repeat the following mantra: "Correlation does not necessarily indicate causation." The professor translated this to mean that just because two variables are associated with each other, one cannot conclude that one of the variables caused the other. In the case at hand, just because African American and Hispanic students more frequently scored in the Not Yet Proficient and Almost Proficient categories than did white, non-Hispanic and Asian American students, it does not mean that students' race or ethnicity is a cause of their mathematics test performance. Many of the things we identify and measure in educational research and policy analysis are surrogates for other variables. If students of one racial or ethnic group tend more frequently to be economically disadvantaged than do students of other racial or ethnic backgrounds, the test performance differences found to be associated with race and ethnicity could really be a function of relative economic disadvantage. To examine this possibility, it would be necessary to further disaggregate the test data by computing the distributions of test scores for students of differing economic status, within each racial or ethnic group. Alternatively, one could compute the distributions of test scores for students of different racial or ethnic background within each socio-economic-status category. Unfortunately, since only 11 fourth-graders in Lincoln County were eligible for free lunch and only 23 were eligible for reduced-price lunch, test results disaggregated by race, ethnicity, and lunch-program status would not be statistically reliable.

Recall that the first thing the Committee's principals wanted to know was how students in their school scored on the state tests, compared to students in other schools. The results shown in Figure 6, with fourth-graders' mathematics performances disaggregated by school, address this issue. Dr. Clinton asked Committee members what the graph in Figure 6 indicated. Mr. Bluford observed that, although the distributions of mathematics test performance for fourth-graders were different for a few schools, he was struck by the overall similarities more than the differences. Ms. Salisbury agreed but noted that fourth-graders in Wilson, Adams and West Schools seemed to be doing better, overall, than fourth-graders in most of the other schools. She based her comments on the percentages of fourth-graders in these schools who were classified as Advanced or Proficient.

Recalling the discussions the Committee had regarding statistical reliability and the issue of causality, Mr. Pinnix cautioned his colleagues about any conclusion that these three schools were more effective than others in their mathematics teaching for fourth-graders. He noted first that the numbers of fourth-graders in any one school ranged from a low of only 30 to a high of 78. Dr. Clinton agreed, noting that a 95 percent confidence interval around a finding that half a school's fourth-graders were Proficient would range from 36 percent to 64 percent for a school that enrolled 50 fourth-graders, and from 39 percent to 61 percent for a school that enrolled 75 fourth-graders. So the differences in mathematics performance the schools exhibited in 1995-96 might not hold in other school years, just due to the particular sample of students enrolled in a particular school in a given year. Mr. Pinnix pointed out that differences in students' mathematics performances across schools could be due to a host of things that had nothing to do with the instructional effectiveness of any school. For example, some schools might enroll more economically disadvantaged students than other schools, and some schools might enroll more gifted students than other schools. Earlier analyses showed that these characteristics of fourth-graders were strongly associated with their mathematics test performances.

Although she agreed with Mr. Pinnix's comments, Ms. Jackson cautioned the Committee about taking a "that's the way things are; there's nothing we can do about it" position. She encouraged the Committee to advocate use of the disaggregated test results as a basis for further investigation by Superintendent Crawford and others. If fourth-graders in some schools had better performances on the state's mathematics test than did fourth-graders in other schools, Ms. Jackson said it was essential that the Lincoln County Schools follow up with analyses and inquiry to find out why. If more effective instruction was the reason, she felt it essential that this be disseminated to all schools and teachers in the district. Noting that African American and Hispanic fourth-graders in the Lincoln County Schools didn't do as well as other students on the state's mathematics test, Ms. Jackson also encouraged analysis of the test data by race and ethnicity within each of the district's schools. She acknowledged that sample sizes would be very small, and that the results wouldn't have a lot of statistical reliability. Nonetheless, Ms. Jackson felt it was important to search for settings and conditions in which students who were members of these racial and ethnic groups did better. Others on the Committee agreed with Ms. Jackson's position.

The Committee decided to meet again to discuss the rest of the data displays Dr. Clinton and Mr. Kelly had produced. Members were encouraged by the results and looked forward to the next meeting in two weeks.

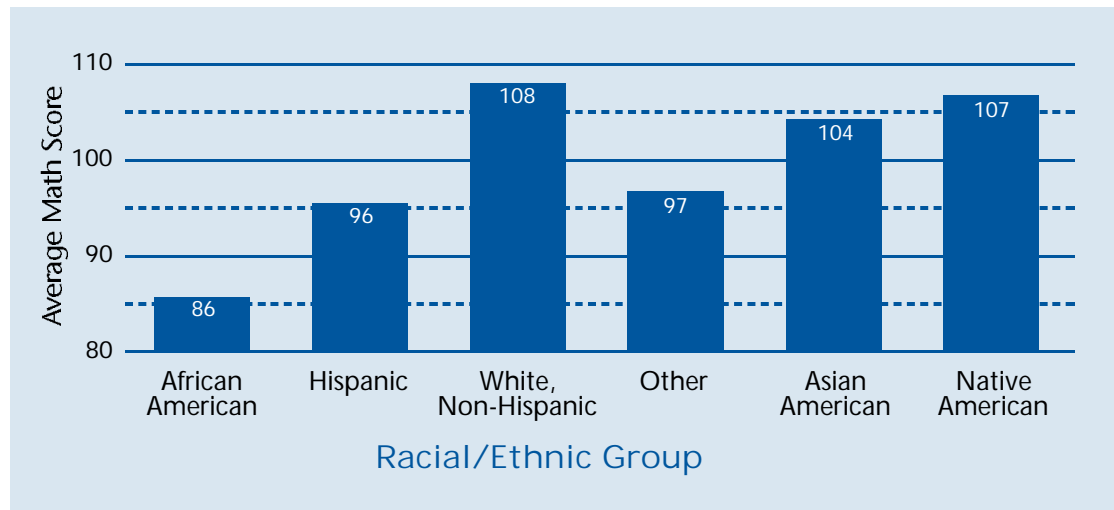
The Committee's Fourth Meeting

At the Committee's next meeting, Dr. Clinton and Mr. Kelly produced two more types of graphs. The first type displayed the average test scores earned by students in different ethnic groups. The second showed the box-and-whisker charts described earlier, side by side, for students who were members of different ethnic groups. Dr. Clinton suggested that Committee members look at one graph of the first type, to get an idea of the kind of information it conveyed, but spend more time reviewing the box-and-whisker charts because they conveyed more information than did plots of averages. Ms. Salisbury agreed, noting that a plot of averages might mask as much information as it revealed, since two groups could have very different averages even though their distributions of test scores overlapped quite a bit.

Figure 7 shows the average scores on the state's mathematics test earned by Lincoln County Schools' fourth-graders, by racial and ethnic group, during the 1995-96 school year. Dr. Clinton pointed out that this figure conveyed some of the information shown in Figure 3, but concentrated on the average earned by each group instead of the percent of each group that scored in each of the state's achievement categories.

Ms. Jackson cautioned the Committee about taking a "that's the way things are; there's nothing we can do about it" position.

Figure 7. Average Math Score for Lincoln County Schools' Fourth-Graders, by Racial/Ethnic Group



NOTE: Averages shown in this figure have been rounded to the nearest whole number.

The results shown in this figure suggest that white, non-Hispanic fourth-graders, on average, scored considerably higher than African American or Hispanic fourth-graders, that fourth-graders who classified themselves as “Other” scored about as well, on average, as Hispanic fourth-graders, and that Asian American and Native American fourth-graders scored about as well as white, non-Hispanic fourth-graders.

Ms. Salisbury was interested in the similarity of the average mathematics test performances of white, non-Hispanic, Asian American, and Native American fourth-graders. She wondered whether these results would likely generalize to other school years. To address this issue, Dr. Clinton looked up the sample size for each group, and then constructed confidence intervals around each of the averages. She said that averages with overlapping confidence intervals could not be regarded as reliably different from each other. In other words, if two averages had confidence intervals that overlapped, the group with the higher average might well have a lower average in a different school year, just due to sampling fluctuation. The details on how to construct a confidence interval around a sample average are given in Appendix C. The results of Dr. Clinton’s computations are shown in Table 3.

Table 3. Ninety-five percent confidence intervals around the average 1995-96 mathematics test scores of fourth-graders in Lincoln County Schools, by racial/ethnic group

Group	Sample Size	Upper 95 Percent Confidence Limit	Lower 95 Percent Confidence Limit
African American	146	89.6	82.6
Hispanic	88	99.0	92.4
White, non-Hispanic	335	109.0	106.2
Other	41	102.1	91.3
Asian American	16	111.8	96.5
Native American	2	122.7	91.3

Earlier it was noted that population parameters are considered “immutable,” in the sense that they do not vary across samples, as do sample statistics. It is now time to back away from that position ever so slightly. at least in terms of their definition of a population. To a statistician, an average score on an achievement test, computed for, say, every student in a particular

Although statisticians would hold that population parameters are immutable, specialists in the field of educational measurement would take somewhat different view.

racial or ethnic group in a school district, would be considered a population parameter. But to a measurement specialist, such an average would be a statistic that estimated what the students' true average score would be, were it possible to administer an infinite number of different forms of the achievement test to the population of students on an infinite number of occasions, provided the students' true achievement did not vary across test forms or occasions.

The difference in perspectives between the statistician and the measurement specialist is that the statistician only considers sampling fluctuation across samples of students to be a source of error in trying to estimate a population parameter. The measurement specialist also considers measurement error across test forms and testing occasions, regarding a single administration of one form of a test to be a sample of students' performance across all possible forms and occasions that leave the students' true performances intact.

In the examples we have considered thus far, and in those we will consider beyond this discussion, we will ignore measurement error as a source of variability in measuring the achievement of groups of students even though we would give it close attention were we testing individual students. The basis for this decision is the relative magnitude of sampling error and measurement error in realistic examples of school, school district, or statewide assessments. For samples of students that are of reasonable size — that is, samples that support separate reporting of results — sampling error will be far greater than measurement error, typically to the degree that ignoring measurement error will make little difference in the final interpretation of results.

Consider the following example in the context of the current discussion of the average Grade 4 mathematics achievement of students who are members of different racial or ethnic groups in the Lincoln County Schools. Lincoln County uses a mathematics test that has a *standard error of measurement* of four points. The standard error of measurement is like a standard deviation. It is a measure of how much an individual student's mathematics performance would fluctuate, were that student tested repeatedly on separate occasions, using different forms of the mathematics test, under conditions in which the student's true mathematics performance did not change. Although these conditions typically cannot be realized in practice, the standard error of measurement is a very useful statistic that can be estimated realistically in any of several ways. Some ways involve testing a group of students only once and others involve testing a group of students on two occasions, perhaps using two different forms of a test. Either approach is feasible in practice.

Now, standard errors of measurement behave like standard deviations — they are smaller for groups of students than they are for individuals. In fact, the standard error of measurement of a group average is merely the standard error of measurement of a test score for an individual student divided by the square root of the size of the group. So, for example, with a standard error of measurement of four points for an individual, the standard error of measurement of the average score of African American fourth-graders in Lincoln County (of which there are 146 according to Table 1), would be $4/\sqrt{146} = 4/12.083 = 0.331$. Since, as shown in Appendix C, the standard error of the average Mathematics achievement of African American fourth-graders in Lincoln County is 1.783, the standard error of measurement of the average is comparatively small (it is only 18 percent as large), and can safely be ignored.

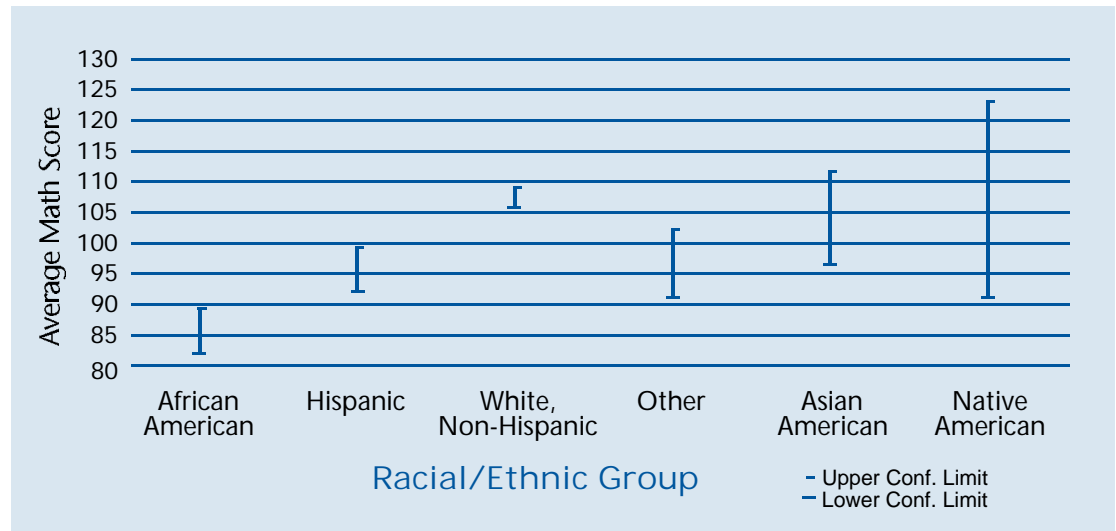
To be absolutely correct in computing a confidence interval around the average score of African American fourth-graders, one would square the standard error of the average for the 146 students, square the standard error of measurement of the average, add the two together, and take the square root of the sum. The result in this case would be $[(1.783)^2 + (0.331)^2] = (3.180 + 0.110) = 3.290 = 1.814$. This standard error is only trivially larger than the standard error of the average, 1.783, that considers only sampling error. In fact it is only 1.7 percent larger. Thus ignoring measurement error in this case, as in most realistic cases, makes little difference⁶.

Dr. Clinton decided that it would be easier to compare the confidence intervals in Table 1 if she graphed them. The resulting graph (Figure 8) is shown below. Dr. Clinton pointed out several

⁶Direct assessments of students' performances, such as assessments of students' productive writing skills, might be an exception to this general statement. In many such assessments, students respond to a single essay prompt and their responses are scored on a graded scale involving four to six points. If different essay prompts are used in successive school years, their difficulties might fluctuate enough to result in a relatively large standard error of measurement. In this case, the standard error of measurement for a sample average might be large enough that it should not be ignored when computing a confidence interval.

interesting features of the graph. First, the 95 percent confidence intervals around the averages for white, non-Hispanic students, Asian American students and Native American students overlap completely. Therefore, one cannot say that the average mathematics scores for these three groups differ reliably. Second, the 95 percent confidence intervals around the averages for Hispanic students and for students who classified themselves as “Other” overlap, so their averages are not reliably different. Third, the confidence interval around the average for the Native American students is so wide that it overlaps the confidence interval of every group except that of African American students.

Figure 8. Confidence intervals around average mathematics scores of Lincoln County Schools’ fourth-graders, by racial/ethnic group (1995-96 school year)



Ms. Jackson noticed that the average for Native American students was based on test results for just two students, and she was concerned about producing a report that contained their average. She felt that the test scores of individual students were the business of the student, the student’s parents, and the student’s teacher, and should not be reported in any way that identified the student. Although the average was based on the scores of two students rather than just one, Ms. Jackson felt that everyone in their school would know who the students were (the two Native American students happened to attend the school where Ms. Jackson taught), and that a report containing their average score would violate the two students’ rights to privacy. Other Committee members agreed with Ms. Jackson.

Mr. Bluford then raised the question of whether Committee members would feel the same way about an average based on the test scores of three students. What about four students? What about five? At what point were enough students in the sample that releasing their average test score would not violate individual privacy rights?

Dr. Clinton said that there is no statistical answer to this question, and she didn’t know of any generally-accepted rule. It was a policy decision. Following an extended discussion, the Committee decided to recommend to Superintendent Crawford that a group’s test performance not be reported publicly unless there were ten or more students in the group. This would protect students’ privacy rights. The Committee reached the same decision that the Educational Testing Service uses when it reports median Graduate Record Examination scores for applicants to a given major at a single college or university. If there are fewer than ten students in the group, no statistics are reported for that group.

Another factor that mitigates against reporting average scores for very small groups of students is the wide confidence interval around the average; typically, small sample sizes result in wide confidence intervals. Notice that the average mathematics score for Native American fourth-graders was 107, but the 95 percent confidence interval ranged from a lower limit of 91 to an upper limit of 123 (rounded). This interval is very wide, indicating that we have little idea of where

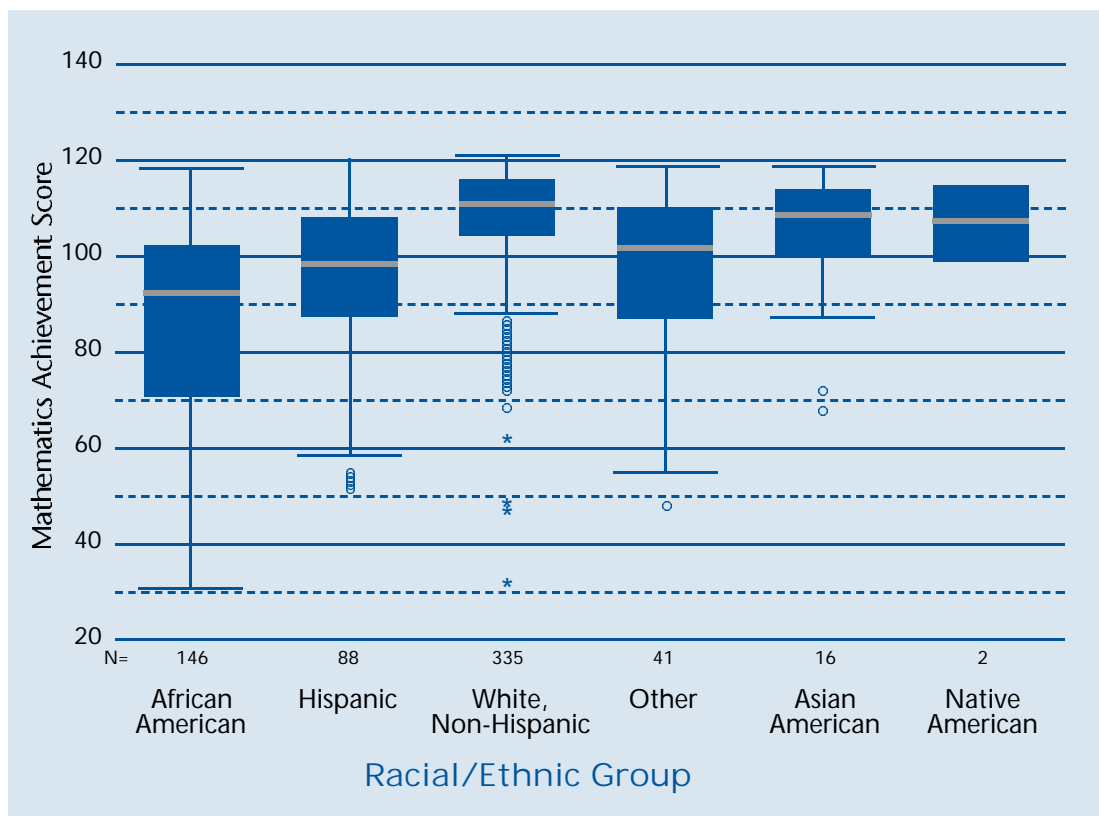
the test scores of individual students were the business of the student, the student’s parents, and the student’s teacher, and should not be reported in any way that identified the student.

the average mathematics score of a population of Native American fourth-grade students would fall. With this much uncertainty about where the average would be in any given school year, many would consider reporting an average for a given school year to be potentially misleading.

The wide confidence interval around the average mathematics score for Native American students led Ms. Salisbury to take another perspective on the issue of how large a group must be before its test results are included in a public report. Ms. Salisbury reasoned that the two Native American students' privacy would not have been harmed had the school district merely reported the confidence limits around their average mathematics score, rather than the average itself. She suggested an alternative rule for reporting test results for a group: Report the average test score for groups of ten or more, but the confidence limits around the average (without the average itself) for groups of five to ten students. In this way, she suggested, students' privacy would be protected and the most unreliable averages (those based on very small samples of students) would be avoided. The other Committee members thought Ms. Salisbury's suggestion had merit and were willing to modify their recommendation to Superintendent Crawford to include it.

Dr. Clinton suggested that the Committee look at two of the box-and-whisker charts (explained on pages 17-18) she had produced with Mr. Kelly. Figure 9 provides information that, in part, overlaps with that shown in Figures 3 and 7. It again shows the mathematics test performances of fourth-graders in the Lincoln County Schools by racial and ethnic group. But in this figure, more than the group's average scores are shown, so the graph provides more information about the groups than did Figure 7 and different information than did Figure 3.

Figure 9. Mathematics test score distributions for Lincoln County Schools' fourth-graders, by racial/ethnic group (1995-96 school year).



Looking at Figure 9, Committee members immediately saw that every group's distribution of scores overlapped with that of every other group to some degree. Figure 7 did not reveal this; it just showed the average scores for each group, and the overlapping wasn't highlighted by the results shown in Figure 3. However, the results shown in Figure 9 also indicated some important differences among the groups' test performances. First, the score distributions of the Hispanic fourth-graders and the fourth-graders who identified themselves as "Other" were virtually indis-

For each of these distributions of scores then, the box-and-whisker charts show that the upper 50 percent of the students scored pretty well, but many students in the lower halves of these distributions had a good bit of trouble with the mathematics test.

tinguishable from each other, but were somewhat lower than the distributions for white, non-Hispanic students and Asian American students. Even though the “distribution” for Native American students was shown on Figure 9, the Committee decided to ignore it because it was based on just two students.

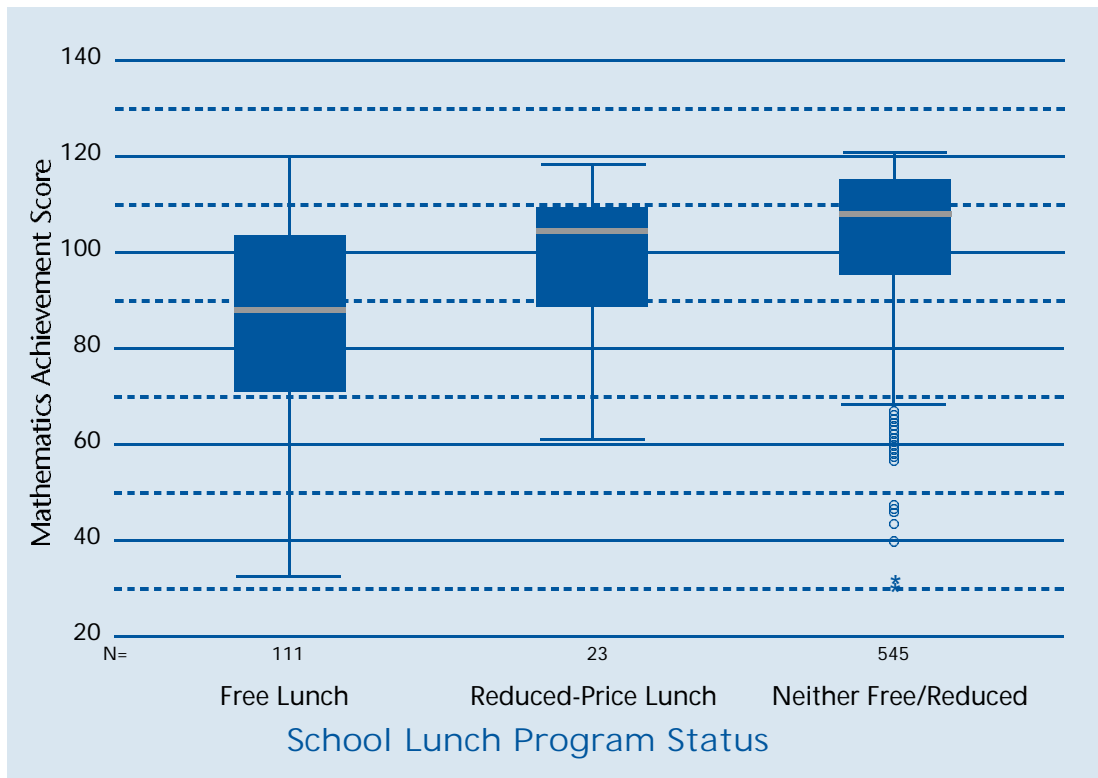
Second, although the distributions of scores for African American and Hispanic fourth-graders overlapped somewhat, the Committee saw that the 75th percentile of the distributions of scores for African American students was barely higher than the median of the distribution for Hispanic students. Some African American students scored as high as students in any other group, but as a whole, the mathematics distribution for African American students was lower than that of other groups. In fact, 75 percent of the African American students scored below the 25th percentile of the distribution for white, non-Hispanic students.

Dr. Clinton also pointed out that the box-and-whisker charts provided information about the shapes of the distributions of scores for the various racial and ethnic groups. For example, the medians of the distributions for white, non-Hispanic students and African American students did not fall in the middle of their respective boxes. In both cases, the median was closer to the top of the box than the bottom. This indicates that the lowest half of each distribution was stretched out — a condition known as a “*negatively skewed*” distribution. The negative skewness also is indicated by the long lower whiskers of these distributions, compared to the lengths of the upper whiskers, and, in the case of the distribution for white, non-Hispanic students, by the relatively large number of extreme scores in the lower part of the distribution. For each of these distributions of scores then, the box-and-whisker charts show that the upper 50 percent of the students scored pretty well, but many students in the lower halves of these distributions had a good bit of trouble with the mathematics test.

This finding suggests that good or poor performance on the state’s Grade 4 mathematics test is associated with factors other than race and ethnicity. Some African American and white, non-Hispanic fourth-graders did quite well, while others showed very poor performance. A search for other correlates of good fourth-grade mathematics performance among Lincoln County students would therefore be warranted.

Dr. Clinton asked the Committee to review one more box-and-whisker chart — a chart that was analogous to Figure 4, indicating the mathematics test performances of fourth-graders in different school lunch programs. Figure 10 contains these results.

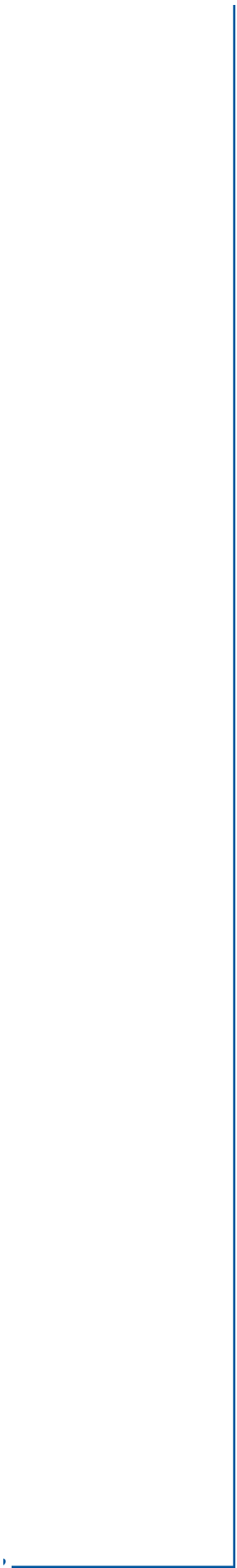
Figure 10. Mathematics test score distributions for Lincoln County Schools' fourth-graders, by school lunch program status (1995-96 school year).



The results shown in Figure 10 indicate that fourth-graders' mathematics test performance is strongly associated with their participation in a free or reduced-price school lunch program and, to the degree that such participation indicates socio-economic status, their mathematics test performance and their socio-economic status are strongly related. Dr. Clinton noted that, although all three distributions shown in Figure 10 overlapped somewhat, it is important to realize that 75 percent of the fourth-graders who were eligible for free lunch earned mathematics scores below the median of the distribution of scores of students eligible for reduced-price lunch. In turn, 75 percent of these students earned mathematics scores that were below the median of the distribution of scores of fourth-graders who were not eligible for free or reduced-price lunch. So once again, the disaggregated test scores suggest an alternative to a causal inference that minority racial or ethnic-group membership is a factor in producing low test scores. It could be the case that Lincoln County Schools hasn't found an effective remedy for the educational deprivation that children of the poor often bring to school.

The issues raised in this vignette — How can test results be disaggregated to better understand the factors associated with students' test performances? How can students' test results be displayed so as to convey interesting features of their test performances? How large must a group of students be to support reporting their collective test performances? How trustworthy are differences among the average test performances of students in various groups? What conclusions can one reach concerning the causes of students' test results? — arise in virtually all analyses of achievement test data, whether for a school, a school district, or an entire state. Although the graphs and data tables shown in this vignette did not address all of the questions raised by Lincoln County's teachers and principals, a number of important issues were considered. Other equally important questions will be addressed in the vignettes that follow.

...once again, the disaggregated test scores suggest an alternative to a causal inference that minority racial or ethnic-group membership is a factor in producing low test scores.



Vignette 2: The Harold Howe II High School

Important Questions Addressed in this Vignette:

- How can test results be disaggregated to better understand the factors associated with students' collective progress across two school years?
- How can students' test results be compared across two school years, in order to illuminate a school's progress in achieving statewide achievement goals?
- How large must a group of students be to support reporting their collective test performances?
- How trustworthy are differences between the average test performances of students in successive school years?

The Harold Howe II High School

The Harold Howe II High School is located in a medium-sized school district in a midwestern state. The district is in an area that is the ancestral home of a Native American tribe that, unfortunately, is not officially recognized by the federal government. Nonetheless, many members of the tribe still live in the area. The district includes a small city with an economy grounded in manufacturing and a surrounding rural area that produces grain crops. The local Camp Vigilance Marine Base employs a number of civilians. The children of Marine officers and enlisted personnel are enrolled in the district's public schools.

Howe is a large school that serves students from the western portion of the district. The school is in a state that has been bitten hard by the "accountability bug." As part of its accountability program for schools and school districts, the state has a required end-of-grade testing program in reading, mathematics, writing, science and social studies that tests all students in Grades 3 through 8, in addition to an end-of-course testing program for all high school students who enroll in designated basic courses, which include Algebra I, Biology, English I, U.S. History and Social Studies. The Social Studies test is called the "Economic, Legal and Political Systems Test."

During the 1995-96 school year, 149 students at Howe completed Algebra I and, therefore, took the state's end-of-course test in Algebra I. In 1996, the state added Algebra I to its list of required courses for the graduating class of 1999. As a result, many more students enrolled in Algebra I in all of the state's high schools, including Howe, during the 1996-97 school year. All of those students took the state's end-of-course test in Algebra I.

In 1995-96, 79 percent of the students at Howe who completed the end-of-course test in Algebra I identified themselves as Native American, 19 percent identified themselves as African American, and of the three remaining students who were tested in Algebra I that year, one each identified themselves as Hispanic; as white, non-Hispanic; and as of "Other" racial/ethnic background.

The following year, in 1996-97, 83 percent of the Howe students who completed the Algebra I test identified themselves as Native American, 15 percent identified themselves as African American, four identified themselves as of "Mixed Race" (a category newly added to the state's student background questionnaire that year) and one was identified as white, non-Hispanic.

The State's Testing Program

Apart from its writing test, which contains a single essay question, the state's end-of-grade and end-of-course tests are composed of multiple-choice questions. Tests are administered in the spring of each school year and are mandatory for all students who are enrolled in regular courses and whose Individualized Education Program does not preclude their being tested under the state's accountability program. Students whose native language is not English and who have been

The school is in a state that has been bitten hard by the "accountability bug."

enrolled in U.S. schools for less than two years also are excused from testing since the state does not yet offer foreign-language versions of its tests.

In keeping with its accountability-based purposes, results for the state's end-of-grade testing program are reported on a developmental scale that is common across Grades 3 through 8, as well as on a four-point achievement-level scale. The achievement levels are defined as follows:

- Level I Fails to achieve at a basic level. Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.
- Level II Achieves at a basic level. Students performing at this level demonstrate inconsistent mastery of knowledge and skills that are fundamental in this area and that are minimally sufficient to be successful at the next grade level.
- Level III Achieves at a proficient level. Students performing at this level consistently demonstrate mastery of grade-level subject matter and skills and are well prepared for the next grade level.
- Level IV Achieves at an advanced level. Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade-level work.

The state's end-of-course tests produce scores that also are reported in two ways: First, scores are reported on a scale with a mean in the low 50's and a standard deviation just smaller than 10. Second, performances are reported on a four-point achievement-level scale. The achievement levels for the end-of-course tests are defined as follows:

- Level I Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at a more advanced level in the content area.
- Level II Students performing at this level demonstrate inconsistent mastery of knowledge and skills that are fundamental in the subject area and are minimally prepared to be successful at a more advanced level in the content area.
- Level III Students performing at this level consistently demonstrate mastery of the subject matter and skills and are well prepared for a more advanced level in the content area.
- Level IV Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the subject area and are very well prepared for a more advanced level in the content area.

Although the state's end-of-course tests have not been scaled identically, their average scores and the standard deviations of their scores are similar, as are the general shapes of their score distributions. Therefore, the same scaled scores on different end-of-course tests can be regarded as being roughly equivalent in terms of students' normative statewide performance.

The School's Mathematics Curriculum Task Force

A committee of mathematics teachers at Howe High School, appointed by the Chair of the Mathematics Faculty with the approval of the school's principal, was particularly interested in students' performances on the state's Algebra I test. During the 1995-96 school year, anticipating the state's decision to require Algebra I of all students, Howe's mathematics teachers were concerned that all students be prepared for the course, enroll in the course early in their high school careers and succeed in the course. They considered the state's Algebra I end-of-course test to be a useful indicator of students' success, without the kinds of variation in standards that typically plague course grades. The teachers had worked hard during the 1995-96 school year to implement an effective pre-algebra course and to redesign the Algebra I course with strict attention to the learning objectives specified in the approved state curriculum guide.

The Mathematics Committee included Mr. James Gauss, Chair of the Mathematics Faculty at Howe; Ms. Jennifer Reimann, who taught Algebra I and various computer applications courses; and Mr. Steven Fermat, who taught Algebra II, Statistics I, and Calculus. Because of their knowledge of computer-based data analysis and statistics, these mathematics teachers received permission from Mr. Ralph Tyler, Howe's principal, to analyze the scores earned by Howe's students on the Algebra I test during the 1995-96 and 1996-97 school years. Mr. Tyler asked the Committee to consult the recent memorandum from the District Office regarding Title I disaggregation requirements in planning the analyses. Mr. Tyler obtained the scores on a diskette from the school district's Testing Coordinator, together with necessary documentation indicating the way the test data had been arranged.

The Committee's First Meeting

At the Committee's first meeting, the three teachers decided that they needed to identify the questions they wanted to answer with the Algebra I test data, to develop a plan for analyzing the data, and to assign responsibilities for conducting the necessary analyses. Mr. Gauss began by asking for suggestions on the issues that should be addressed using the Algebra I data. The Committee's suggestions were as follows:

- (1) How many students were tested during the 1996-97 school year, compared to the 1995-96 school year?
- (2) Were the 1995-96 and 1996-97 test takers similar in background?
- (3) How did the students tested during the 1996-97 school year and the 1995-96 school year compare in terms of overall performance?
- (4) How did the students with similar backgrounds who were tested during the 1996-97 school year and the 1995-96 school year compare in terms of performance?
- (5) Did the average scores of some kinds of students change more than the average scores of other kinds of students, from the 1995-96 school year to the 1996-97 school year?
- (6) Were differences between the average performances of various kinds of students statistically reliable?

Upon reviewing the questions, Mr. Fermat suggested that the Committee compute the following statistics:

- Frequency distributions of students tested, by school year and background characteristic, for every characteristic available on the diskette file provided by the state;
- Average scores and standard deviations of scores on the Algebra I test, for all students, for students categorized by background characteristic, and by school year;
- Distributions of percentages of students at each of the state's four achievement levels on the Algebra I test, overall by year, and by background characteristic and year;
- Confidence intervals around the average scores on the Algebra I test students earned in each year, overall and by background characteristic of student, or hypothesis tests of the differences between students' average scores across the two school years.

Committee members also suggested that the Algebra I scores be summarized graphically, so that changes in students' performances across school years and differences between the performances of students with various background characteristics could be seen clearly. Simple graphs were preferred over complex ones, so that results could be readily and clearly understood.

The teachers decided to meet weekly to complete the necessary data analyses. Ms. Reimann agreed to take the lead on computerized analyses of data, provided the other teachers, particularly Mr. Fermat with his solid background in statistics, would work with her. She agreed to review the data format on the diskette before the Committee's next meeting the following week.

.it is always a good idea to conduct data analyses with a working team because the results of some analyses inevitably suggest others, and new questions can be explored on the spot.

The Committee's Second Meeting

The mathematics teachers held their second meeting in Howe's computer lab. Mr. Fermat had secured a number of microcomputer-based programs for his statistics courses, including SPSS for Macintosh, SAS for Windows, and the Microsoft Excel spreadsheet program. Using these programs in combination, he felt certain that they could produce all of the statistics specified during the first meeting.

The teachers analyzed the school's Algebra I data as they talked. The teachers made a very clever decision. When experts in using packaged statistical analysis programs are available, it is always a good idea to conduct data analyses with a working team because the results of some analyses inevitably suggest others, and new questions can be explored on the spot. We are not suggesting that statistical work be done in the absence of a detailed plan like the one developed by the Mathematics Committee at Howe. Sound planning ensures that the right questions will be investigated, and it is therefore essential. However, it is virtually impossible to anticipate every interesting and potentially fruitful analysis at the outset of a study, and having the option of adding new issues along the way is almost always helpful.

The first statistical question posed by the Mathematics Committee concerned the number of Algebra I students tested during each year (1995-96 and 1996-97), overall and by background characteristic. To answer this question, Mr. Fermat used the SPSS for Macintosh computer program, analyzing both the 1995-96 and 1996-97 data files. He called for frequency distributions using each available background variable for categorization of students within each year. He obtained the following results:

Table 4. Numbers of students at Harold Howe II High School who completed Algebra I and the State End-of-Course Test in Algebra I, by school year

	1995-96	1996-97
	149	284

The data in Table 4 indicate that the enrollment of Howe High School students in Algebra I increased substantially from the 1995-96 school year to the 1996-97 school year, just as the teachers had hoped. In fact, the increase was more than 90 percent. The Committee members were very pleased with this result since the total enrollment at Howe was almost the same during the two school years.

To see whether the increase in students enrolling in Algebra I was restricted to particular types of students or was more general, Mr. Fermat produced the following tables:

Table 5. Numbers of students at Harold Howe II High School who completed Algebra I and the State End-of-Course Test in Algebra I, by racial/ethnic group and school year

Racial/Ethnic Group	1995-96 School Year	1996-97 School Year	Percent Change
African American	28	43	53.6
Hispanic	1	0	-100.0
Native American	118	236	100.0
Mixed Race	—	4	—
White, non-Hispanic	1	1	0.0
Other	1	—	—

Table 6. Numbers of students at Harold Howe II High School who completed Algebra I and the State End-of-Course Test in Algebra I, by exceptionality classification and school year

Exceptionality	1995-96 School Year	1996-97 School Year	Percent Change
Not Classified as Exceptional	147	281	91.2
Academically Gifted	—	2	—
Deaf-Blind	1	1	0.0
Speech-Language Impaired	1	0	-100.0

Table 7. Numbers of students at Harold Howe II High School who completed Algebra I and the State End-of-Course Test in Algebra I, by parental education level and school year

Parental Education Level	1995-96 School Year	1996-97 School Year	Percent Change
Not High School	28	34	21.4
High School	63	139	120.6
Trade/Business School	4	5	25.0
Community College	26	50	92.3
4-year College	24	50	108.3
Graduate School	4	6	50.0

The results shown in Tables 5 through 7 were interpreted by the Mathematics Committee as indicating a general increase in Algebra I enrollment from the 1995-96 school year to the 1996-97 school year. Although the increase in enrollment was numerically and proportionately larger for Native American than for African American students, a 54 percent increase for the latter group was taken as a positive sign. The Committee was particularly interested in the parental education variable, since it is positively related to economic status. While there was some variation in enrollment growth across students whose parents had different levels of formal education, there was no indication that Algebra I enrollment had grown exclusively among students with extensively-educated parents. The state will begin including students' free and reduced-price lunch status in its data collection in 1997-98, which will be a better indicator of students' economic status and will help to satisfy Title I reporting requirements.

The data in Tables 5 through 7 also demonstrated that virtually all Howe students who enrolled in Algebra I during the two school years had not been classified as having a disability. This concerned the Committee, and the teachers decided to investigate why students with disabilities were not enrolling in the course. In addition, only a few students who enrolled in Algebra I during the two school years listed their racial or ethnic-group membership as something other than Native American or African American. These latter results were particularly useful in deciding how to analyze students' Algebra I test performances in relation to their background characteristics. The Committee wanted to protect students' privacy rights in all of its analyses and reports. The teachers therefore decided to limit reports of their analyses to categories of students with sample sizes of 10 or more.

The teachers were now ready to pursue some of their interests in students' Algebra I test performances across the 1995-96 and 1996-97 school years.

They computed both graphs and tables because they wanted to see the fine detail that is most readily gained by reading a table in this case, actual averages by background characteristic for each school year), as well as the general pattern across school years, information that is most readily gained by looking at a graph.

The Committee's Third Meeting

The third set of statistics the Mathematics Committee teachers wanted to produce involved students' average scores on the state's Algebra I test, for all students and for students classified by background characteristic, during the 1995-96 and 1996-97 school years. These statistics were readily produced by Mr. Fermat and his colleagues, again using the SPSS computer program and the Microsoft Excel program on the school's Macintosh desktop computer. They computed both graphs and tables because they wanted to see the fine detail that is most readily gained by reading a table (in this case, actual averages by background characteristic for each school year), as well as the general pattern across school years, information that is most readily gained by looking at a graph. The graphs and tables that the teachers produced were as follows:

Figure 11. Average scores of students at Harold Howe II High School who completed the Algebra I test, by racial/ethnic group membership and school year

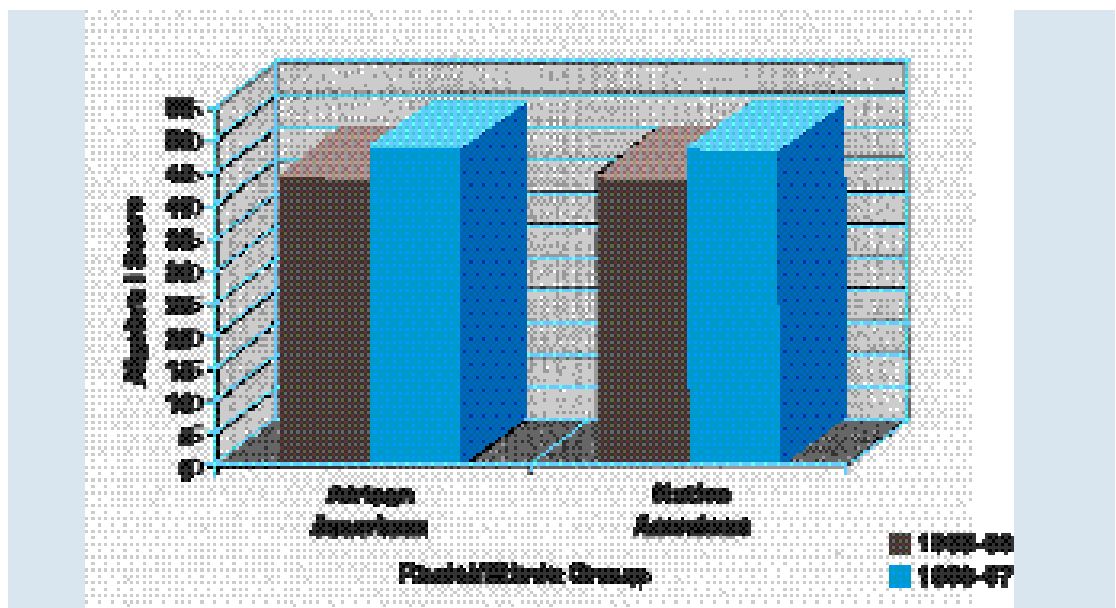


Table 8. Average scores of students at Harold Howe II High School who completed the Algebra I test, by racial/ethnic group membership and school year

School Year	African American	Native American
1995-96	44.4	44.9
1996-97	48.7	48.2

Figure 12. Average scores of students at Harold Howe II High School who completed the Algebra I Test, by gender and school year

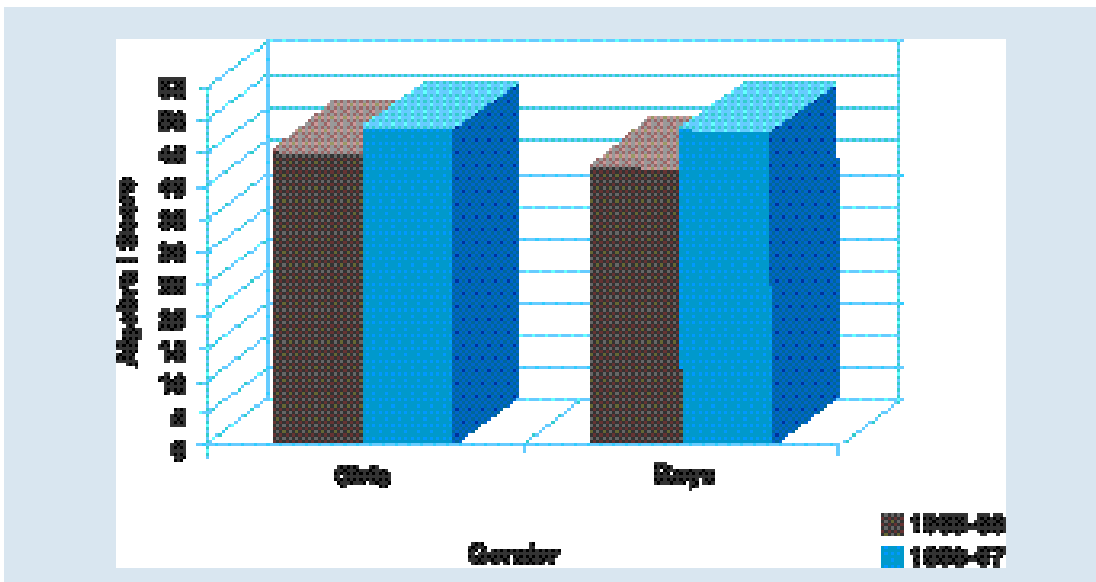
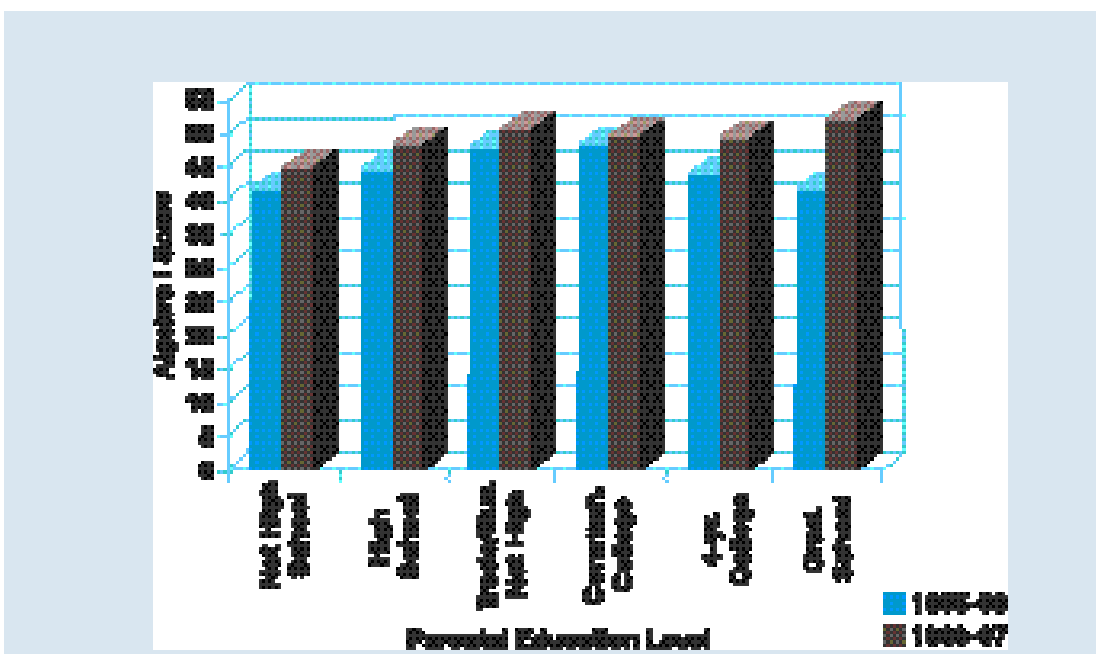


Table 9. Average scores of students at Harold Howe II High School who completed the Algebra I Test, by gender and school year

School Year	Girls	Boys
1995-96	46.1	43.5
1996-97	48.8	47.7

Figure 13. Average scores of students at Harold Howe II High School who completed the Algebra I Test, by parental education level and school year



It appeared that movement was consistently in the desired direction and that no group of students was being left behind by the innovative pre-algebra preparation program the teachers had implemented in 1995.

Table 10. Average scores of students at Harold Howe II High School who completed the Algebra I Test, by parental education level and school year

Parental Education	1995-96	1996-97
Not High School	41.6	44.8
High School	44.9	48.4
Trade/Bus. School	47.8	50.4
Commun. College	48.2	49.2
4-yr. College	44.1	49.2
Grad. School	41.8	51.8

The Mathematics Committee was very pleased with the results shown in Figures 11 through 13 and Tables 8 through 10. All groups for which Algebra I averages were analyzed — boys and girls, Native American students and African American Students, and students grouped by parental education level — showed an increase in average Algebra I test scores from the 1995-96 school year to the 1996-97 school year. It appeared that movement was consistently in the desired direction and that no group of students was being left behind by the innovative pre-algebra preparation program the teachers had implemented in 1995.

Ms. Reimann was pleased to see consistent movement in the desired direction, but she wondered whether the changes in Algebra I averages shown in these tables and graphs could be attributed solely to differences between the samples of students who happened to enroll in Algebra I during the two school years. In other words, she asked, could these differences be solely due to random sampling fluctuation? The other teachers thought this was an excellent question that should be pursued through appropriate analyses of the test data. Fortunately, the SPSS computer program produces the statistics necessary to answer the question, and the teachers quickly got back to the data.

Using the hypothesis testing procedures described in detail in Appendix D, the teachers tested the null hypothesis that the average Algebra I scores earned by each group of students during the 1995-96 and 1996-97 school years were identical, in the populations from which these students had been sampled. Because they tested a number of hypotheses at the same time, the teachers used a value larger than 1.96 to determine whether their t-statistics indicated statistical significance (see Appendix E for details).

The t-statistic (see Appendix D) corresponding to the difference between the Algebra I averages earned by Native American students during the 1995-96 and 1996-97 school years was 4.08, based on a total sample size of about 350. The critical value against which this t-statistic was compared to determine its statistical significance was equal to 2.28 according to the Table in Appendix E. (To be conservative, the tabled value for an overall sample size of 100 was used.) Since 4.08 is larger than the critical value of 2.28, the teachers concluded that the difference between the average Algebra I scores earned by Native American students during the 1995-96 and 1996-97 school years was unlikely to be a chance phenomenon, and that real improvement had occurred. The same statistical comparison for African American students produced a t-statistic of 2.29, which was just larger than the critical value of 2.28. Again, the teachers concluded that the difference between the students' Algebra I averages during the two school years was statistically significant. For students of both racial/ethnic groups, the Mathematics Committee concluded that Algebra I averages had improved by more than a chance amount from 1995-96 to 1996-97.

When the Mathematics Committee teachers compared the Algebra I averages earned by boys during the two school years, they likewise concluded that the change was statistically significant, since the corresponding t-statistic was 3.97. For girls, the t-statistic was 2.68 — again indicating a difference in average scores across the two school years that was statistically significant at the 0.05 level. For each of these comparisons, the sample size was well over 100, so the critical value of 2.28 from Appendix E was again used.

When they looked at changes in Algebra I averages for students whose parents had differing amounts of formal education, the teachers spent a bit more time computing, since six different pairs of averages had to be tested for statistical significance. Again, the computer programs made the computations feasible. The teachers found the following results.

Table 11. Results of t-test to test statistical significance of the change in Algebra I scores from 1995-96 to 1996-97 for students with parents of different educational levels.

Parental Education Level	t-statistic	Critical Value	Statistically Significant?
Not High School	1.70	2.73	No
High School	3.13	2.68	Yes
Trade/Bus. School	0.44	3.64	No
Commun. College	0.64	2.71	No
4-yr. College	2.96	2.71	Yes
Grad. School	2.19	3.48	No

The teachers were at first disappointed by these mixed results, since they indicated that, although the average Algebra I score was higher in 1996-97 than in 1995-96 for all groups of students, some of the differences were so small that they could be attributed entirely to random sampling variation. In other words, when the answer to the question “Statistically Significant?” is “No”⁷, the difference cannot be considered statistically reliable. But then the teachers looked at the numbers of Algebra I students whose parents had various levels of formal education and found that, for the larger groups, the differences in Algebra I averages tended to be statistically significant across the two school years. They dismissed the non-significant results for the students whose parents’ highest level of education was “Trade or Business School” because the numbers of such students in the two school years were only four and five, respectively. For the same reason, they dismissed the results for students whose parents’ highest level of education was “Graduate School” since those respective sample sizes were four and six students in the two school years. Of the four remaining groups, the largest by far was students whose parents completed high school (45 percent of all students in 1996-97), and for that group, the difference between the 1995-96 and 1996-97 Algebra I test averages was statistically significant. However, the teachers had to conclude that the gains shown by some groups, namely those of students whose parents’ highest levels of education were “Less Than High School” or “Community College,” might be illusory.

At this point, the teachers felt that they understood the magnitude and extent of changes in the average Algebra I test scores earned by Howe’s students between the 1995-96 and 1996-97 school years. The overall average in 1995-96 was 44.7 and in 1996-97 it was 48.3, a gain of about three and a half points. Both boys and girls had shown increases in their average scores on the Algebra I test, as had African American and Native American students. Students whose parents had completed high school and students whose parents had completed four-year colleges, the largest groups tested, also had shown statistically reliable gains, but the gains of student groups whose parents had other levels of formal education were not statistically reliable.

Mr. Gauss then raised an interesting question. Whether or not a difference between averages is statistically reliable depends, in part, on the sizes of the samples used to compute the averages. A given difference between averages might be statistically significant if the sample sizes are large but not statistically significant if the sample sizes are small. Also, knowing that an observed difference is unlikely to be equal to zero in the populations underlying the observed samples offers no more than a small degree of comfort and tells nothing about whether the observed difference is large or small in a substantive sense. Mr. Gauss wondered whether there was some statistic that would indicate whether the average Algebra I scores earned by Howe High’s students during the 1995-96 and 1996-97 school years differed by a small amount or a substantial amount. Mr. Gauss

Knowing that an observed difference is unlikely to be equal to zero in the populations underlying the observed samples offers no more than a small degree of comfort and tells nothing about whether the observed difference is large or small in a substantive sense

⁷ Some of the critical values shown in the table above were read from a more complete listing than is provided in Appendix D. Do not, therefore, be surprised if you do not find all of these critical values in the abbreviated table contained in Appendix D. A complete table of critical values for the Bonferroni hypothesis testing procedure can be found in an article by J. R. Bailey (1977), *Journal of the American Statistical Association*, 72, 469-478.

was asking whether the difference in the students' average performance across the two school years was *substantively significant*.

Mr. Fermat said that he often raised similar questions in his statistics courses. He didn't want his students to use statistical significance as a "holy grail" indicator of importance — imbuing it with greater meaning than it deserved. He emphasized repeatedly the dependence of statistical significance on sample size and its true meaning when applied to the difference between sample averages: Statistical significance indicates that there is likely to be some non-zero difference, but it says nothing about the magnitude of that difference. An alternative statistic, called the *effect size*, indicates the size of the observed difference between sample means on a scale with universal interpretation regardless of the units used to measure the dependent variable (in this case, number of items answered correctly on the Algebra I test).

An estimated effect size can be computed quite easily and is readily interpreted: Just calculate the difference between the sample averages and divide the difference by the pooled standard deviation of the scores in the two samples. The steps to be followed are described in Appendix E. The effect size indicates how much the sample averages differ in terms of number of standard deviation units. Jacob Cohen, who first proposed using effect size as an indicator, described an effect size of 0.2 as small, an effect size of 0.5 as moderate, and an effect size of 0.8 as large. In other words, if the difference between sample averages is no more than two-tenths of a standard deviation, the difference should be regarded as small; a difference of half a standard deviation should be regarded as moderate; and a difference of eight-tenths of a standard deviation or larger should be regarded as a large difference.

When the Mathematics Committee calculated effect sizes corresponding to the differences between the 1996-97 and 1995-96 Algebra I test averages for the various groups, the findings were as follows:

Table 12. Estimated effect size of the change in Algebra I scores from 1995-96 to 1996-97, by gender, racial/ethnic group, and parents' educational level

Group	Estimated Effect Size	Magnitude
Gender		
Girls	0.38	Small to Moderate
Boys	0.55	Moderate
Racial/Ethnic Group		
African American	0.53	Moderate
Native American	0.45	Moderate
Parental Education		
Not High School	0.42	Small to Moderate
High School	0.46	Small to Moderate
Trade/Bus. School	0.30	Small
Community College	0.15	Small
4-yr. College	0.73	Moderate to Large
Grad. School	1.40	Large

Note that effect sizes have been computed for the various groups regardless of whether the groups' differences in average scores across school years were statistically significant. This is because issues of statistical significance and substantive significance should be investigated separately, even though both should be considered when interpreting results. That is, one should consider a statistical difference to be trustworthy and of some importance only if it is both statistically significant and if the magnitude of its estimated effect size is large enough to be of interest. The hedging that appears in the last portion of the preceding sentence is intentional. For some purposes, researchers and policymakers will consider even small effect sizes to be important (an example would be a reduction in the average incidence of heart disease, following the administration of a new medication), while for other purposes, only moderate-to-large effect sizes would

One should consider a statistical difference to be trustworthy and of some importance only if it is both statistically significant and if the magnitude of its estimated effect size is large enough to be of interest.

demand attention. One might also consider the magnitude of an effect size in the context of those associated with alternative courses of action; achieving a larger effect size while minimizing costs would be regarded favorably, even though the effect size achieved was classified as small.

Having thoroughly explored the differences between students' average scores on the Algebra I test during the 1995-96 and 1996-97 school years, the Mathematics Committee decided to devote the next week's meeting to examining their questions concerning students' achievement-level performances during the two school years.

The Committee's Fourth Meeting

The Mathematics Committee again met in Howe's computer lab so that the teachers could pursue their analyses. Ms. Reimann suggested that they construct distributions of the percentage of students who had scored at each of the four achievement levels defined by the state on the Algebra I test. She particularly wanted to see how the distributions compared across the 1995-96 and 1996-97 school years. Mr. Fermat said that the SPSS program would readily construct the frequency distributions for each year, and that those results could be input to the Microsoft Excel program to produce graphs that illustrated students' comparative performances across the two school years. The teachers produced the following graphs for the various groups of students:

Figure 14. Distribution of Algebra I achievement levels for girls and boys, by school year (1995-96 and 1996-97)

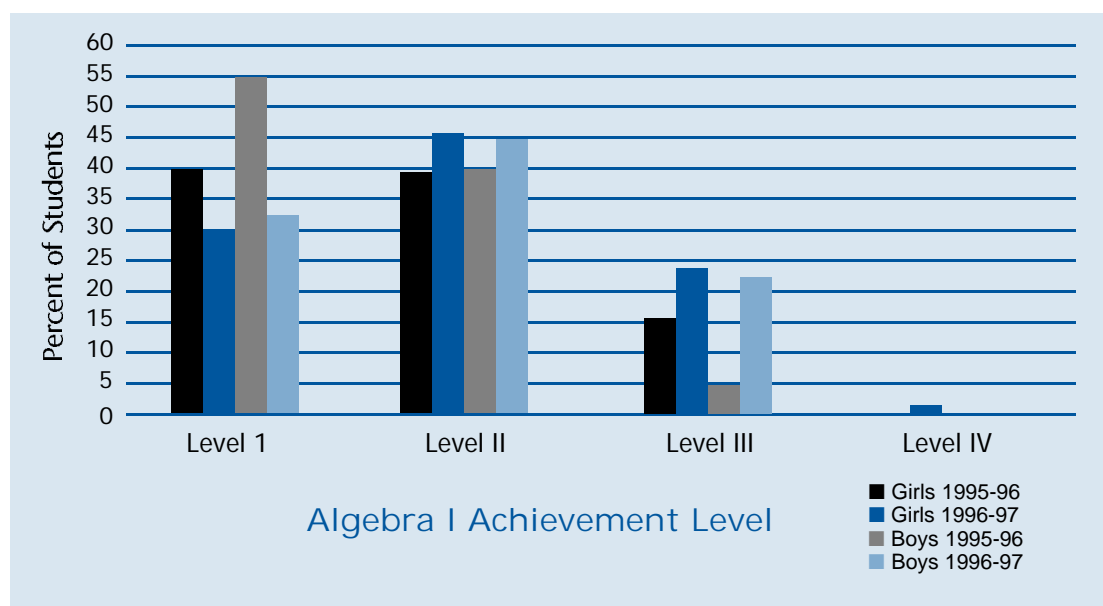


Figure 15. Distribution of Algebra I achievement levels for girls and boys, by school year (1995-96 and 1996-97)

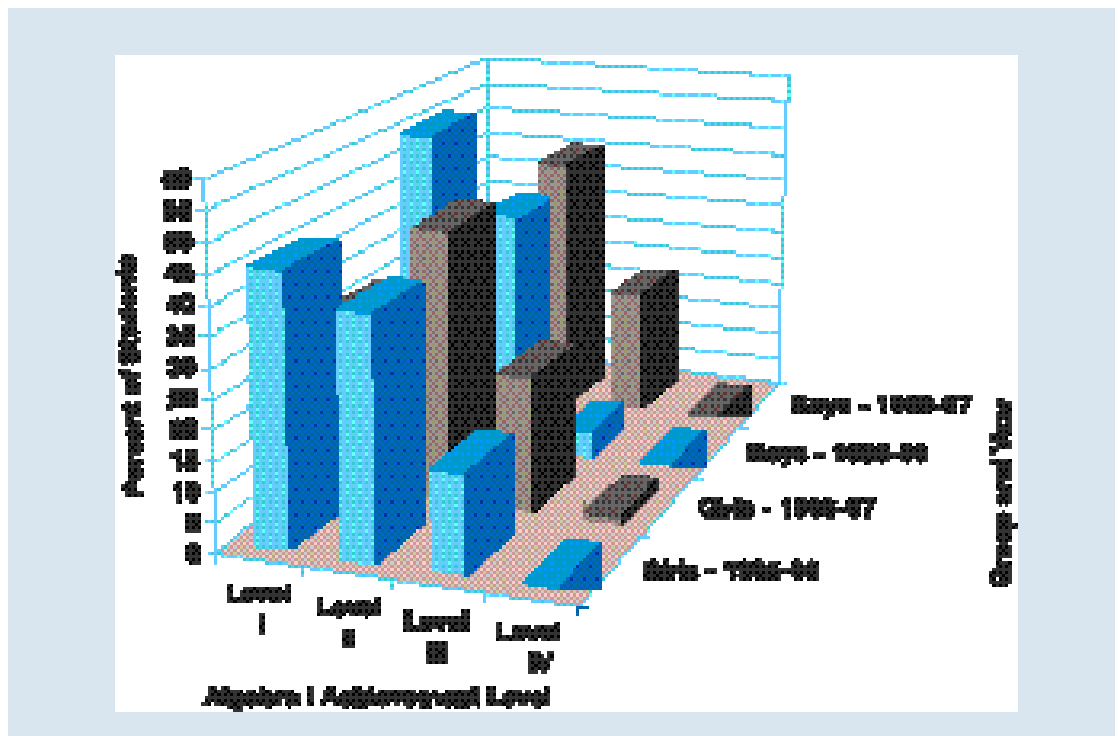


Figure 16. Distribution of Algebra I achievement levels for African American and Native American students, by school year (1995-96 and 1996-97)

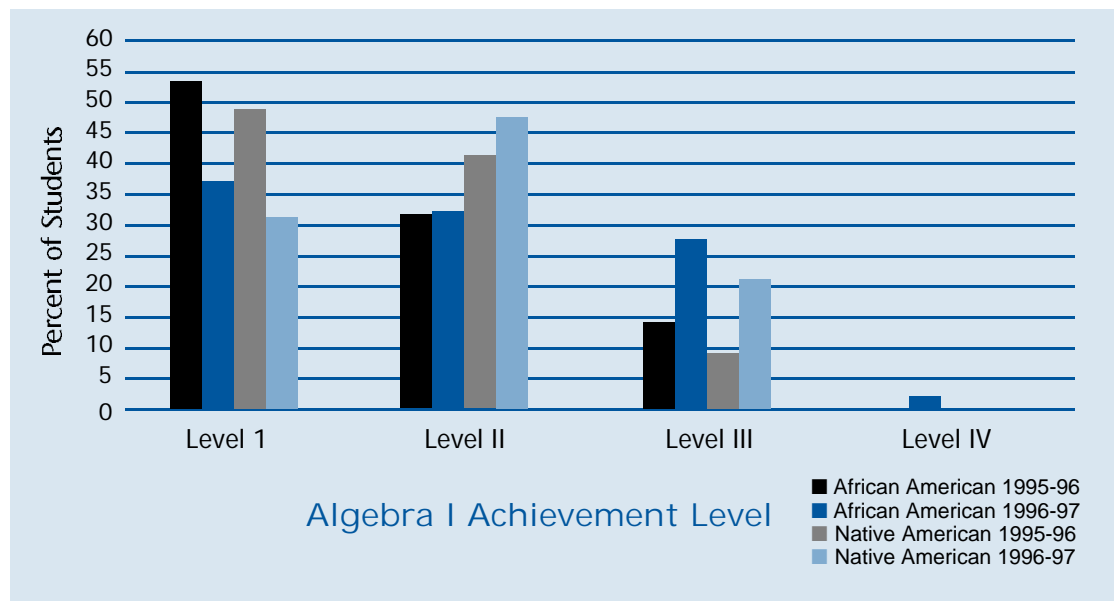


Figure 17. Distribution of Algebra I achievement levels for African American and Native American students, by school year (1995-96 and 1996-97)

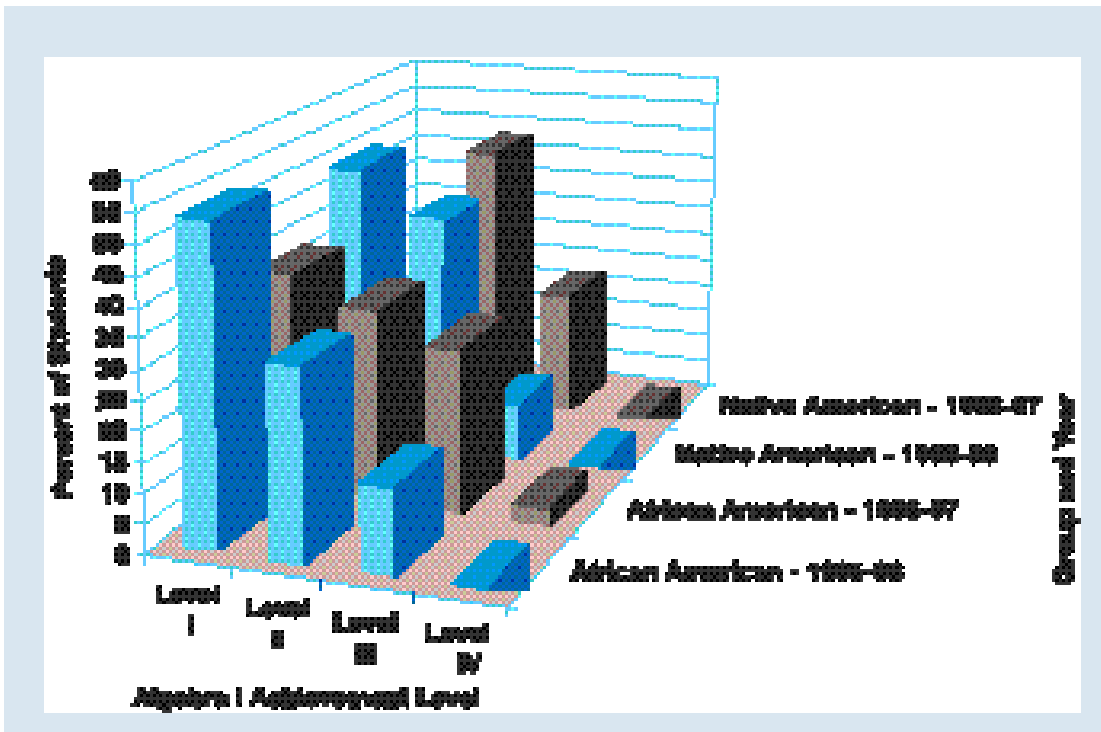
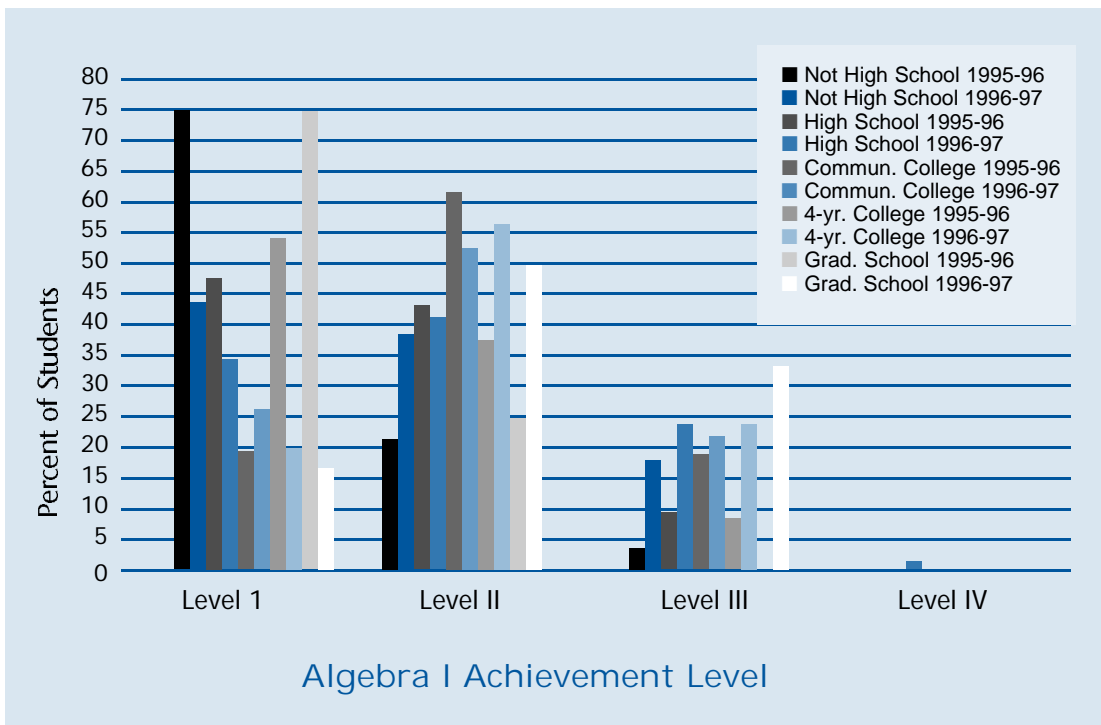
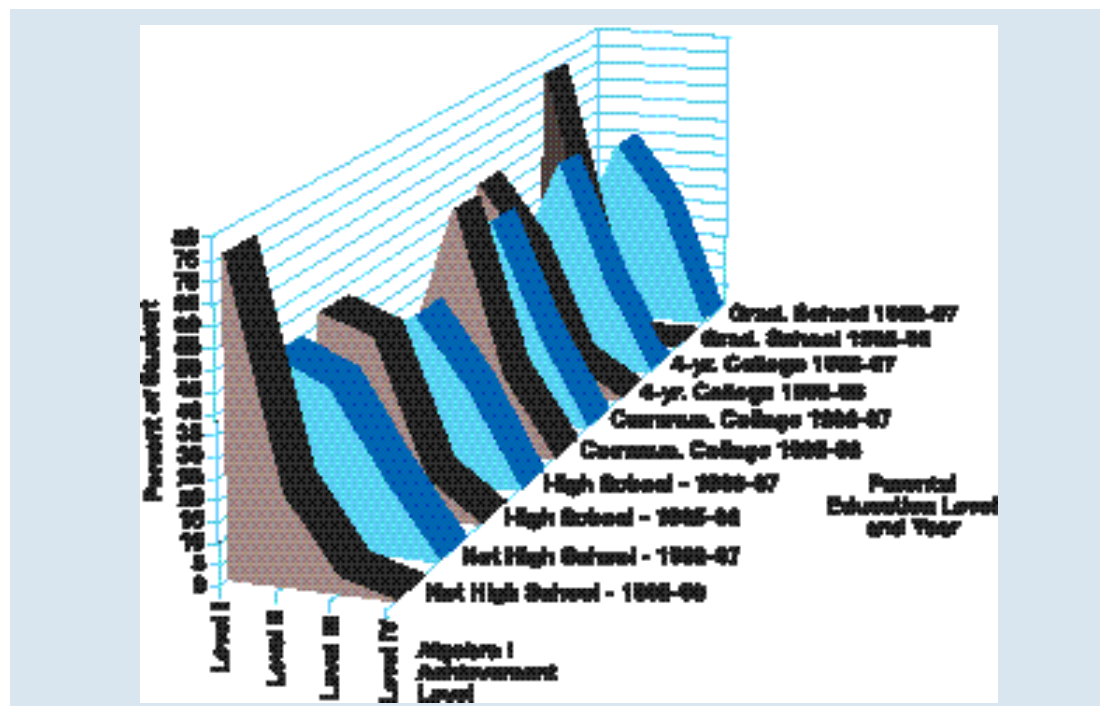


Figure 18. Distribution of Algebra I achievement levels by students' parental education level and school year (1995-96 and 1996-97)



Although about one in five of the students in most of the groups had scored at Level III (indicating proficient performance) during the 1996-97 school year (generally a marked increase from the 1995-96 school year), large percentages of students were still performing at Levels I and II.

Figure 19. Distribution of Algebra I achievement levels by students' parental education level and school year (1995-96 and 1996-97)



When the teachers reviewed the distributions shown in Figures 14 through 19, they felt that the achievement-level results were consistent with their earlier findings regarding changes in Algebra I averages across the 1995-96 and 1996-97 school years. For each group of students, the percentage who scored at Level I went down from the first school year to the next, and the percentages of students who scored at higher achievement levels went up. However, these distributions were revealing in several other ways. First, hardly any students scored at Level IV (consistently superior). Second, although about one in five of the students in most of the groups had scored at Level III (indicating proficient performance) during the 1996-97 school year (generally a marked increase from the 1995-96 school year), large percentages of students were still performing at Levels I and II, indicating failure to achieve sufficient mastery of the Algebra I material or inconsistent mastery of the Algebra I material.

Mr. Fermat and his colleagues produced a pair of graphs — Figures 14 and 15 — that illustrated the distributions of Algebra I achievement by gender and school year, another pair — Figures 16 and 17 — that illustrated the distributions of Algebra I achievement by racial/ethnic group and school year, and yet another pair — Figures 18 and 19 — that illustrated distributions of Algebra I achievement by parental education level and school year. The teachers thought that the three-dimensional graphs — Figures 15, 17 and 19 — more clearly illustrated the general trend in Algebra I achievement within each group across the two school years, but the two-dimensional graphs — Figures 14, 16 and 18 — made it easier to determine the actual percentages of students in each group who achieved at particular levels. Different kinds of graphs are more suited to different purposes and it is not necessary to restrict analysis to a single style of graph.

The Mathematics Committee members reached the following final conclusions. First, their attempts to recruit more students for Algebra I appeared to be working, since enrollments increased markedly across the two school years. Second, they were reasonably successful in reaching all groups of Howe students, since enrollment increases were found for all of the groups they examined. Third, the pre-Algebra program appeared to be helping, since students' average Algebra I test performances increased from 1995-96 to 1996-97 in the face of a recruitment program that had enrolled students who otherwise might not have elected to take the course. Fourth, and most important, there was no room for complacency, since the vast majority of Howe students who

different kinds of graphs are more suited to different purposes and it is not necessary to restrict analysis to a single style of graph.

enrolled in Algebra I were still scoring at Level I or Level II on the state's test, and these levels of performance indicated less-than-proficient performance.

In response to these results, the Mathematics Committee teachers proposed several strategies for improvement. First, the content of the Algebra I course should be carefully reviewed against the state's content blueprint for the Algebra I test to see whether the course content was inconsistent with what the state expected Algebra I students to have learned. The test blueprint was readily available in a technical manual published by the State Department of Education. The teachers resolved to ask their principal for a copy of the technical manual. Second, the teachers decided to review students' performances on unit classroom tests in Algebra that had been given during the 1996-97 school year to see whether there were particular concepts or content areas covered by the state test that were giving students trouble. Additional or modified instruction on these content areas would then be prepared for the 1997-98 school year and, perhaps, post-classroom-test tutorials for students having difficulties in these areas would be developed and offered during the 1997-98 school year.

The Howe mathematics teachers had learned a great deal from their disaggregated analyses of their students' Algebra I test performances and they had developed a positive strategy for increasing students' Algebra knowledge and skills and their test performances in future school years. Mr. Tyler, Howe's principal, was understandably both pleased and appreciative of the Committee's results.

The Howe mathematics teachers had learned a great deal from their disaggregated analyses of their students' Algebra I test performances and they had developed a positive strategy.

Vignette 3: The State of Euphoria

Important Questions Addressed in this Vignette:

- How can students' test results be analyzed across four years, so as to illuminate medium-term trends?
- How can test results be disaggregated to better understand the factors associated with students' collective progress across four school years?
- When are differences in average achievement across years statistically reliable and educationally meaningful?
- What factors can contribute to cross-year trends in student achievement?
- How can trends in students' collective test results be modeled?

The State of Euphoria

Euphoria is a western state that tests all of its sixth-graders in reading, mathematics and writing. The state is modest in size but has rich natural resources and varied terrain. There is only one large city in Euphoria; the rest of the state is divided among smaller cities, suburbs and rural areas. Much of the eastern portion of Euphoria is rugged and mountainous, with low population density.

Economically, Euphoria is neither prosperous nor poverty-stricken. Its diverse population includes families at all levels of the economic spectrum, with almost a fourth of its sixth-graders eligible to receive either free or reduced-price school lunches. Euphoria is racially and ethnically diverse. During the 1996-97 school year, its sixth-graders included about 10 percent who identified themselves as African American, almost 10 percent who identified themselves as Hispanic, two percent who identified themselves as Asian American and half of one percent who identified themselves as Native American. Most of the rest, more than 72 percent, identified themselves as white, non-Hispanic.

Euphoria's major industries are lumber production and tourism, although recent years have seen growth in technological research and production and in a variety of service industries as residents of more-populous western states look to Euphoria as a state they would like to emulate. Euphoria's future is bright, as long as its growing pharmaceutical industry continues to enjoy increased marketing success.

Euphoria's Statewide Testing Program

Consistent with the West's tradition of rugged individualism, Euphoria has delegated control of public education largely to locally-elected school boards. Only since the 1993-94 school year has Euphoria had a statewide testing program, and then only in three core subjects — reading, mathematics and writing — and only in a single grade (the sixth). Whether Euphoria's testing program will be expanded in future years is an open question. Certainly, state legislators and other policymakers would have to be convinced that the benefits of universal achievement testing of the state's students outweigh its costs and lead to improved education of the state's students.

Euphoria's embryonic experiment in statewide test-based school accountability included the establishment of performance standards for sixth-graders at the inception of its new testing program in 1993. Three achievement levels were established for the state's sixth-grade tests in reading and writing that year, and four achievement levels were established for the state's sixth-grade test in mathematics. When the Title I program was reauthorized, the State Board of Education realized that the proficiency levels and assessments it had used would have to be modified to measure a standard of performance that exceeded the "competent" level, and provisions would have to be made for assessing students in elementary and secondary schools that received Title I

Euphoria's embryonic experiment in statewide test-based school accountability included the establishment of performance standards for sixth-graders.

funds. The Board decided to maintain its current state testing program and to require districts to assess students in elementary and secondary schools until the Board had worked with the districts to determine the best way to design an expanded assessment system. The definitions of the state's current achievement levels are as follows:

For Grade Six Reading:

Not Proficient (Scores of 49 or below)

These scores are well below the statewide goal for reading. Generally, students who score at this level can comprehend, with some difficulty, materials written below a sixth-grade level.

Almost Proficient (Scores of 50 to 58):

These scores are below the statewide goal for reading. Generally, students who score at this level can comprehend, with some teacher assistance, textbooks and other materials typically used at grade six or below.

Proficient (Scores of 59 or above)

These scores are at or above the statewide goal for reading. Students who score at this level possess the knowledge and skills necessary to successfully perform the tasks and assignments appropriately expected of a student at this grade level with minimal teacher assistance. Generally, students who score at this level can comprehend textbooks and other materials typically used at grade six or above.

For Grade Six Writing:

Not Proficient (Scores of 2 to 5)

These scores are well below the statewide goal for writing. Generally, students who score at this level produce papers that are very weakly developed with few and/or vague details. These papers are too brief to indicate organization and are awkward and confusing with almost no awareness of audience.

Almost Proficient (Scores of 6 or 7)

These scores are slightly below the statewide goal for writing. Generally, students who score at this level produce papers that are weakly developed. These papers include some expansion of ideas. Usually, general and specific details are presented to show evidence of organization and/or sequencing.

Proficient (Scores of 8 to 12)

These scores are at or above the statewide goal for writing. Generally, students who score at this level produce fluent, well-developed papers with expansion on most or all key ideas. Papers are adequately or fully elaborated with at least a mixture of general and specific details. Satisfactory to strong organizational strategy and/or sequencing is evident in these papers.

For Grade Six Mathematics:

Very Low (Scores below 79)

Students who score in this range are achieving well below the statewide goal for mathematics. Generally, students who score at this level demonstrate very limited computational skills, conceptual understandings, and problem-solving abilities.

Not Proficient (Scores of 80 to 98)

Students who score in this range are achieving below the statewide goal for mathematics. Generally, students who score at this level demonstrate partially developed problem-solving abilities but have limited computational skills and conceptual understandings.

Almost Proficient (Scores of 99 to 121)

Students who score in this range are achieving slightly below the statewide goal for mathematics. Generally, students who score at this level demonstrate only partially developed computational skills, conceptual understandings and problem-solving abilities.

Proficient (Scores of 122 or above)

Students who score in this range are achieving at or above the statewide goal for mathematics. Students who score at this level possess the knowledge and skills necessary to perform the tasks and assignments expected of sixth-graders with minimal teacher assistance. Generally, these students demonstrate well-developed computational skills, conceptual understandings and problem-solving abilities.

Euphoria's sixth-grade reading test contains 77 multiple-choice items that address characteristics of 11 reading passages. It is focused on students' comprehension of non-fiction English prose. The state's sixth-grade mathematics test includes items in multiple-choice, grid-in and open-ended (short answer) format. Each item is scored on a (0,1) scale, with one point awarded for a correct answer. The test covers mathematics concepts, mathematics facts and computation, problem-solving and applications. Euphoria's sixth-grade writing test requires students to write an essay in response to a single prompt. Students are given 45 minutes to plan and compose a first-draft essay that is scored independently by two readers on a six-point scale. The scores awarded by the two readers are summed, producing scores on a 2-point to 12-point scale.

Euphoria's State Testing Office and Its Charge

The State Department of Education in Euphoria includes an Office of Testing and Assessment that employs three professionals and four support staff members. The staff has responsibility for planning, operating and reporting the results of Euphoria's sixth-grade testing program. Actual test development and distribution, receipt and scoring of test materials are completed under contract by a commercial test publisher. Office of Testing and Assessment staff members develop requests for proposals; handle the contracting process; oversee contractor activities; provide information to local school districts and the state legislature; and prepare all reports to the State Superintendent of Schools, the State Board of Education and the legislature on the results of testing the state's sixth-graders.

Dr. Milicent Stanford is Director of Assessment for Euphoria. She holds a Ph.D. with a specialty in testing and measurement from the University of Iowa. Her two professional colleagues, Dr. Steve Hargraves and Dr. Linda Gasper, hold Ph.D.'s in educational psychology with minors in statistics from Oregon State University. All three professionals have substantial experience in the field of testing and measurement at the school district level. Dr. Gasper worked for a commercial test publisher for almost a decade prior to joining Euphoria's Department of Education.

In response to a recent legislative act, the State Superintendent of Schools, Mr. Roger Clawson, has directed Dr. Stanford to produce a report on trends in sixth-graders' achievement-test performances over the four years that the state's testing program has been in place. Legislators want to know whether there is any evidence of a material change in student achievement over the time period and whether there is evidence of differences in tested achievement or achievement growth 1) between girls and boys; 2) among racial/ethnic groups; 3) between students whose parents have less education and students whose parents have more education; and 4) among other student groups of interest.

In particular, the legislature wanted to know whether the state was making progress in increasing the number of sixth-graders whose school achievement was "Proficient," as defined by the state's achievement test standards, and whether any group appeared to be lagging behind in progress toward the goal of proficiency. Mr. Clawson asked that the report produced by the Office of Testing and Assessment, to the degree possible, also provide the test results required under Title I of the Improving America's Schools Act.

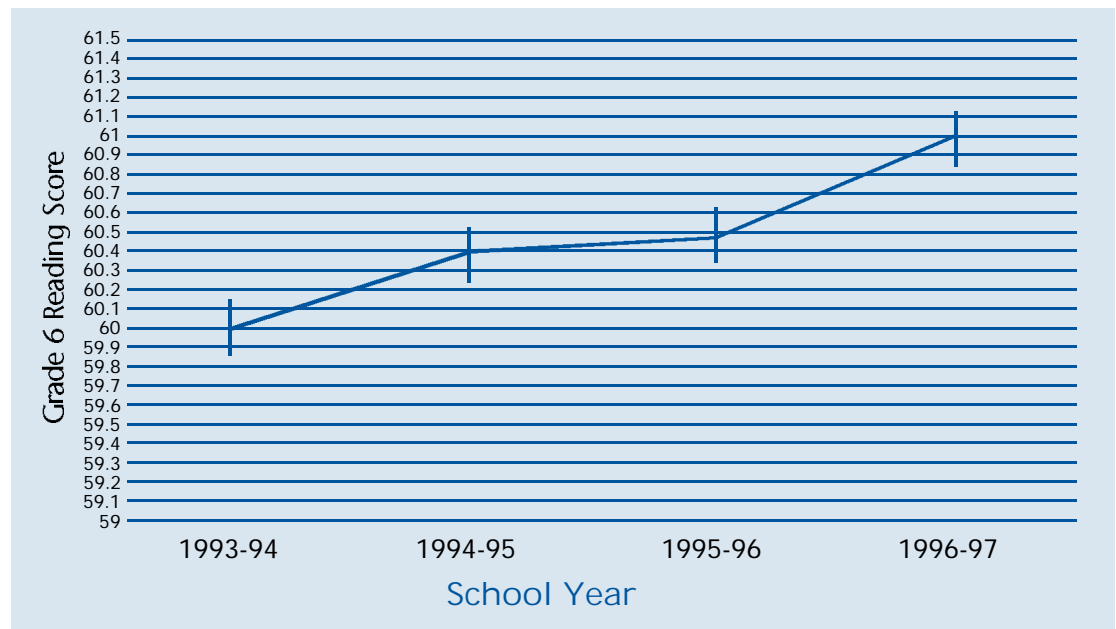
In particular, the legislature wanted to know whether the state was making progress in increasing the number of sixth-graders whose school achievement was "Proficient."

To produce the report requested, Dr. Stanford asked Drs. Hargraves and Gasper to retrieve files of data for tests administered in spring in 1994, 1995, 1996 and 1997 and to begin compiling the necessary statistics and graphs. The first set of graphs they produced illustrated the progression of students' average scores on the reading, mathematics and writing tests across the four school years 1993-94 to 1996-97, together with confidence intervals around those averages⁸. The confidence intervals were computed by following the steps described in Appendix C. These graphs are shown in Figures 20 through 22 for tests in the three subject areas.

From Figure 20, it appears that the average reading score of Euphoria's sixth-graders increased every year from 1993-94 through 1996-97. Although this is true, notice the overlap in the vertical lines surrounding the averages for 1994-95 and 1995-96. These vertical lines represent 95 percent confidence intervals around corresponding averages, and the overlap indicates that the two averages differ so little that we would retain the hypothesis that the underlying population averages were equal. In other words, the apparent difference between the 1994-95 and 1995-96 average reading scores is not statistically reliable.

Reliable gains in average mathematics scores have occurred for Euphoria's sixth-graders only since the 1994-95 school year (see Figure 21), and in writing only since the 1995-96 school year (see Figure 22). Here again, in the absence of the confidence intervals surrounding each year's average, one might incorrectly conclude that mathematics scores have increased steadily across school years. It is obvious that the pattern of writing test averages fluctuated substantially during the 1993-94 through 1995-96 school years, even without examining the confidence intervals surrounding the sample averages⁹.

Figure 20. State of Euphoria average reading test score for sixth-graders, by school year, with 95 percent confidence intervals around sample averages



⁸ Please see pages 22 for a discussion on the interpretation of confidence intervals when all students in a school, school district or state are tested in a given school year.

⁹ Whether this fluctuation is attributable to differences among the distributions of writing skill of students tested in different school years or to errors of measurement associated with essay prompts that differed in difficulty across school years is not clear. See the footnote on page 27 for a discussion of the latter issue.

Figure 21. State of Euphoria average mathematics test score for sixth-graders, by school year, with 95 percent confidence intervals around sample averages

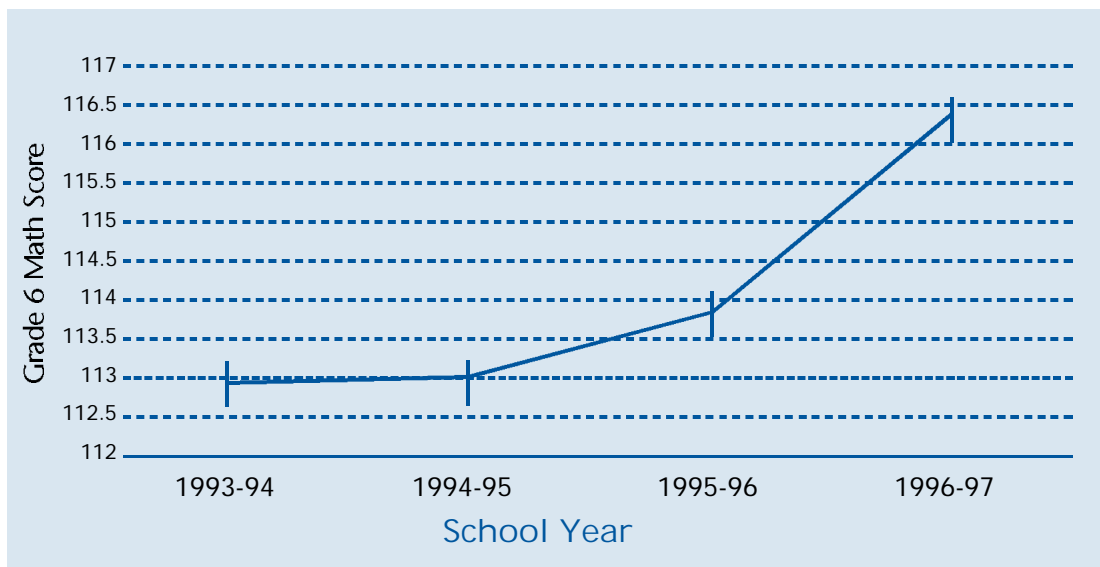
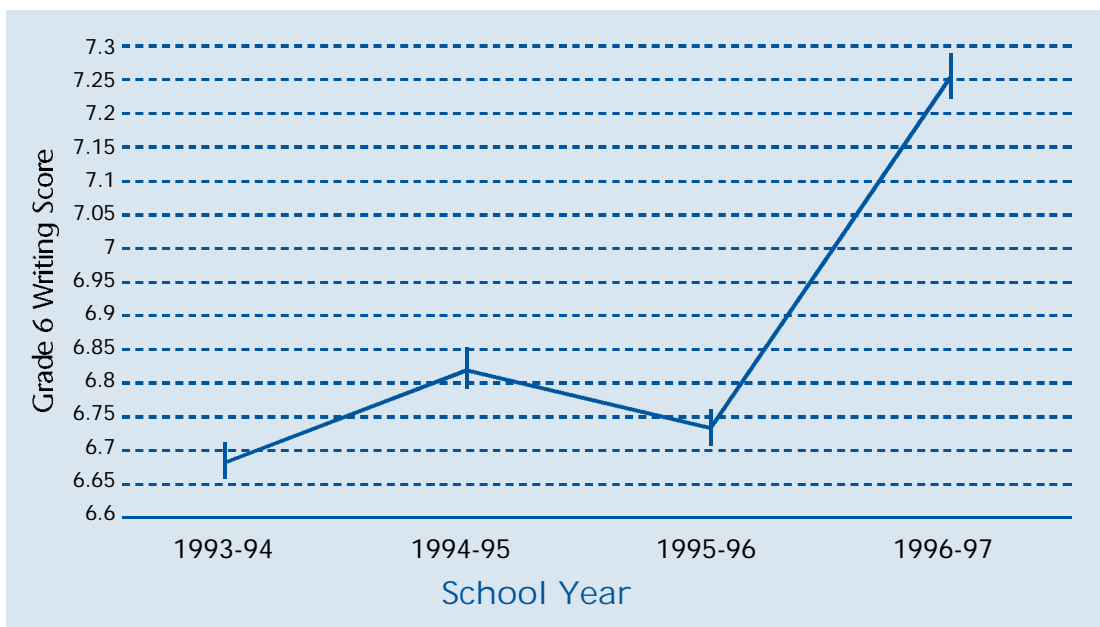


Figure 22. State of Euphoria average writing test score for sixth-graders, by school year, with 95 percent confidence intervals around sample averages



A second set of analyses conducted by Drs. Hargraves and Gasper resulted in Figures 23 through 25, showing average scores on tests in reading, mathematics, and writing for populations of sixth-graders classified by racial or ethnic group. Until the 1995-96 school year, Euphoria did not report test results separately for Asian American and Native American students, nor did the state provide these students with the opportunity to classify themselves as anything but “Other” on a questionnaire that listed specific racial and ethnic groups only as “African American,” “Hispanic,” and “white, non-Hispanic.” The appropriate change made in 1995 renders the “Other” category useless for plotting long-term trends. Notice that on Figures 23 through 25, the average score for “Other” appears to drop precipitously between the 1994-95 and the 1995-96 school years. This drop is entirely an artifact of the modified definition of “Other.” Since the category did not include Asian American and Native American students during the 1995-96 and

This drop is entirely an artifact of the modified definition of “Other.”

When test results are disaggregated, the definitions of subpopulations for which test results are separately reported are critical.

1996-97 school years, it is not surprising that the average test scores earned by students in this category appeared to fall; Asian American students often have the highest average test scores when their results are reported separately. It was largely their contribution that resulted in high average scores for "Other" during the early school years.

The important more-general point to be made here is that when test results are disaggregated, the definitions of subpopulations for which test results are separately reported are critical. More to the point, unrecognized changes in definitions across reporting periods can lead to highly erroneous reporting and interpretation of students' test results.

Figure 23. State of Euphoria average reading test score for sixth-graders, by racial/ethnic group and school year

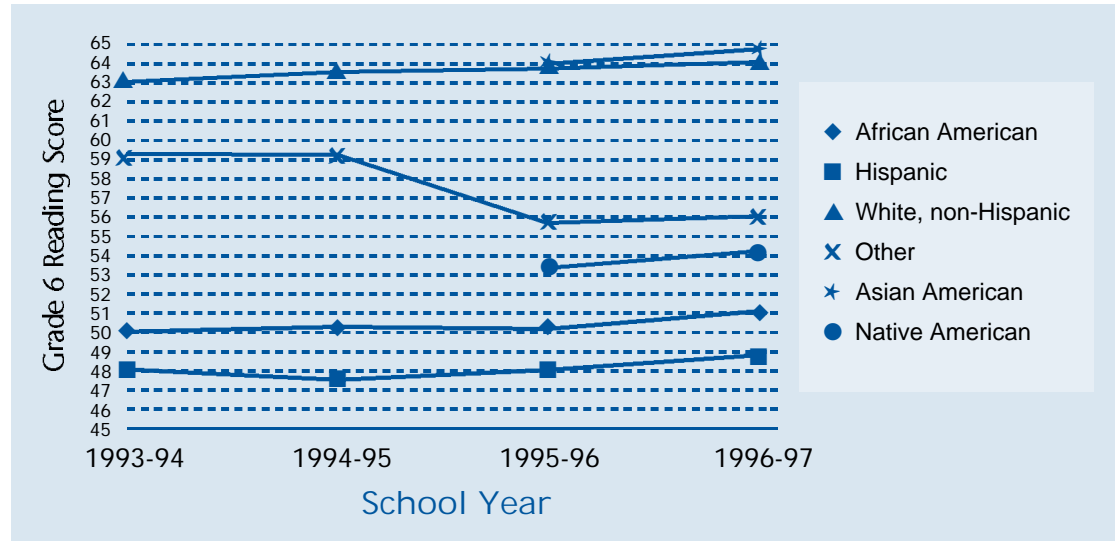


Figure 24. State of Euphoria average mathematics test score for sixth-graders, by racial/ethnic group and school year

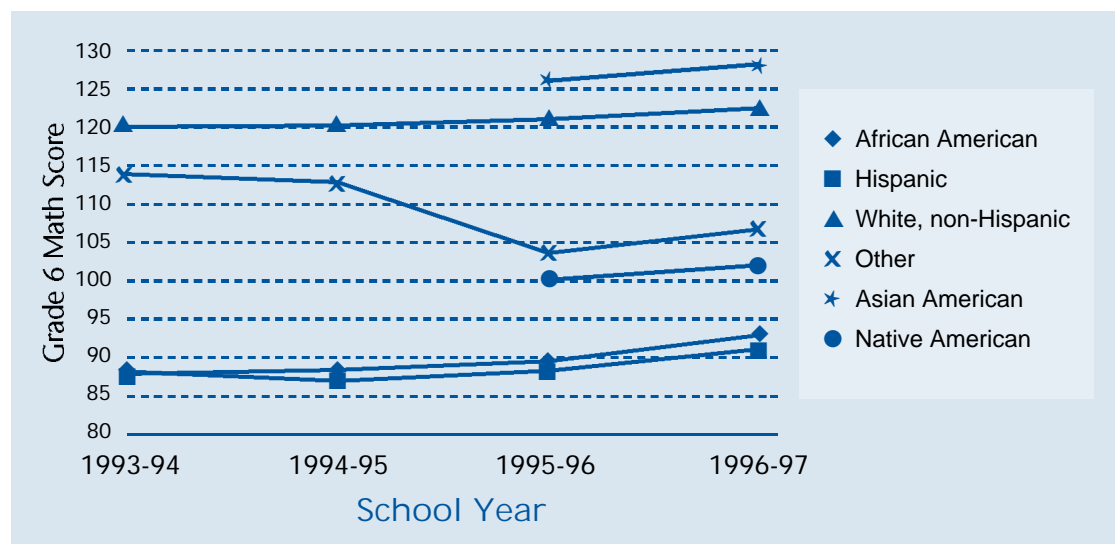
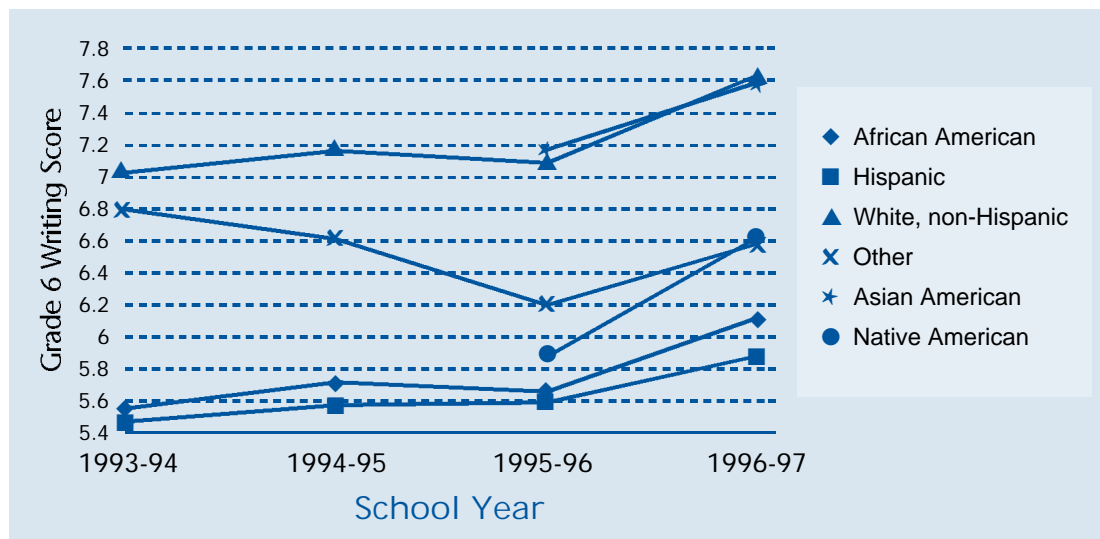


Figure 25. State of Euphoria average writing test score for sixth-graders, by racial/ethnic group and school year



Dr. Hargraves and Dr. Gasper reached two conclusions when they reviewed the achievement test averages shown in Figures 23 through 25. First, observed trends in achievement test averages across years were quite similar for students in every self-reported racial and ethnic group for which results were computed. That is, increases and decreases across school years were of similar magnitude for each group. Second — and of greater importance for policy purposes — in all three subjects, the average scores of white, non-Hispanic students and Asian American students were substantially higher than those of African American, Native American, and Hispanic students. Euphoria’s schools have not found a way to help these lower-scoring groups overcome their relative deficit in average achievement, even though that deficit appears to be substantial and consistent.

Average scores on tests are difficult to interpret, except in a relative sense, either over time or between subpopulations of students. This is because number-right score scales (often termed “raw” scores) rarely carry implicit meaning. For example, it would be difficult to explain the meaning of an average statewide mathematics test score of 113 earned by sixth-graders during the 1994-95 school year, apart from noting that the lowest average score students could have earned was zero and the highest was 155. Perhaps that is why the Euphoria Department of Education defined achievement levels for its sixth-grade tests and the state legislature asked for a report on trends in the percent of students whose test performance warranted labeling them as “Proficient.”

Graphs that illustrate trends in the percentages of Euphoria’s sixth-graders whose reading, mathematics and writing test scores placed them in the “Proficient” category are shown in Figures 26 through 31. For each subject tested, Drs. Hargraves and Gasper prepared a graph that showed the trend across four school years for all tested sixth-graders in the state and another that showed the trend for students within a number of racial or ethnic groups. These graphs fulfill at least part of the legislative request for analyses of trends in the percentage of the state’s sixth-graders who were performing proficiently on the state’s assessments. Other graphs, not discussed in this paper, illustrated corresponding trends for female students and male students, for students eligible for various school lunch programs (as a surrogate for differences in family economic status), for students with disabilities and for those without disabilities, for students who are proficient in English and for those who are not, and for students whose parents had completed various amounts of formal schooling. These analyses not only satisfied the requirements of the Title I program for grades 6 through 9, but provided Euphoria’s legislature with the test results it had requested.

From their inspection of Figures 26, 28 and 30, Drs. Hargraves and Gasper concluded that higher percentages of Euphoria’s sixth-graders had achieved at the “Proficient” level in reading, mathematics, and writing during the 1996-97 school year than had been the case four years earlier.

Average scores on tests are difficult to interpret, except in a relative sense, either over time or between subpopulations of students.

The results shown in Figures 27, 29 and 31 illustrate quite dramatically the persistent differences in test results across sixth-graders in various racial and ethnic groups and highlight a policy issue that demands the attention of Euphoria's legislature.

er, during the 1993-94 school year. Using the procedures described in Appendix A, the two researchers computed 95 percent confidence intervals around the reported percentages for each school year. With about 35,000 sixth-graders tested statewide, the standard error of each reported percentage was about 0.25 percent, and each associated 95 percent confidence interval had an upper limit that was about 0.5 percent above the reported percentage for each year, and a lower limit that was about 0.5 percent below the reported percentage for each year. Therefore, the increases in the percentages of the state's sixth-graders whose scores were in the "Proficient" category from the 1993-94 school year to the 1996-97 school year were statistically reliable for all three tested subjects.

Upon inspecting the results shown in Figures 27, 29 and 31, Drs. Hargraves and Gasper reached several conclusions. First, from Figure 27 the general trend in reading performance for students in various racial and ethnic groups appeared similar across the four school years, with slightly greater year-to-year increases for white, non-Hispanic students. As was true for the graphs illustrating trends in mean scores, in Figures 27, 29 and 31 the apparent decrease in percent "Proficient" for students in the "Other" category is an artifact of the modified definition of this category between the 1994-95 and the 1995-96 school years, and cannot be interpreted as a real drop in proficiency. Second, from Figure 29, greater increases in the percent of students in the "Proficient" category in mathematics were observed in more recent school years than in earlier school years for students in all groups for which four-year trend data were available. Thus year-to-year improvement in six-graders' mathematics performance appeared to be increasing. Since Euphoria did not report achievement test results separately for Native American students and Asian American students until the 1995-96 school year, this conclusion does not apply to those groups. Third, the somewhat erratic trend in the percent of sixth-graders who were "Proficient" in writing shown in Figure 30 was reflected in the performances of students in all racial and ethnic groups for which four-year trend data were available (see Figure 31). Between the 1995-96 and 1996-97 school years, the improvement in writing performance of Hispanic students appeared to be somewhat smaller than that experienced by students in other racial and ethnic groups.

Again, the percentages of Asian American and white, non-Hispanic students with test scores in the "Proficient" category were dramatically higher than were corresponding percentages for students in other racial or ethnic groups. The results shown in Figures 27, 29 and 31 illustrate quite dramatically the persistent differences in test results across sixth-graders in various racial and ethnic groups and highlight a policy issue that demands the attention of Euphoria's legislature.

Figure 26. State of Euphoria, percent of sixth-graders with reading test scores in the proficient category by school year

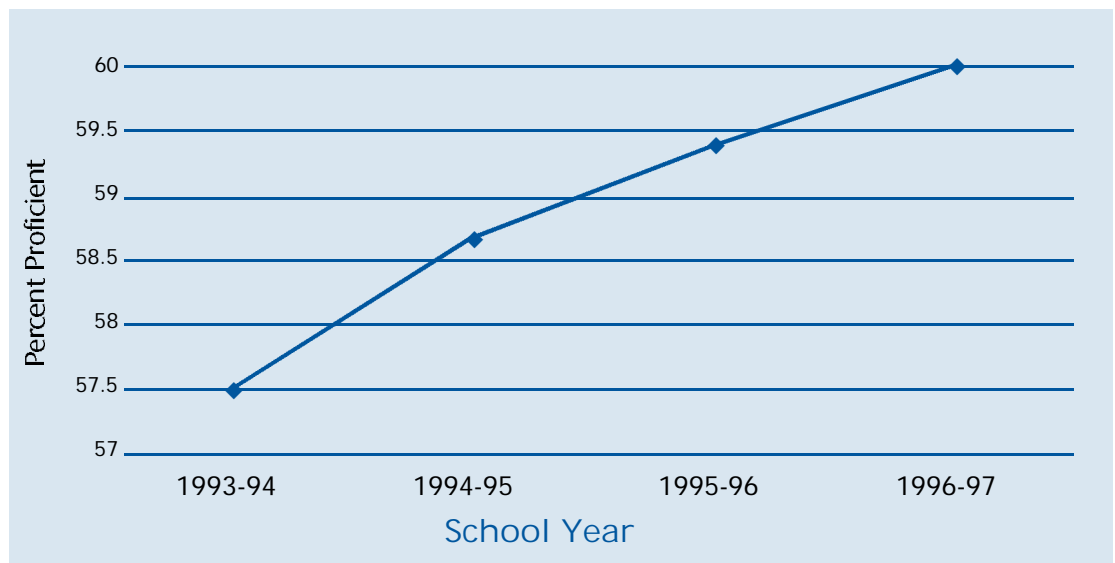


Figure 27. State of Euphoria, percent of sixth-graders with reading test scores in the proficient category, by racial/ethnic group and school year

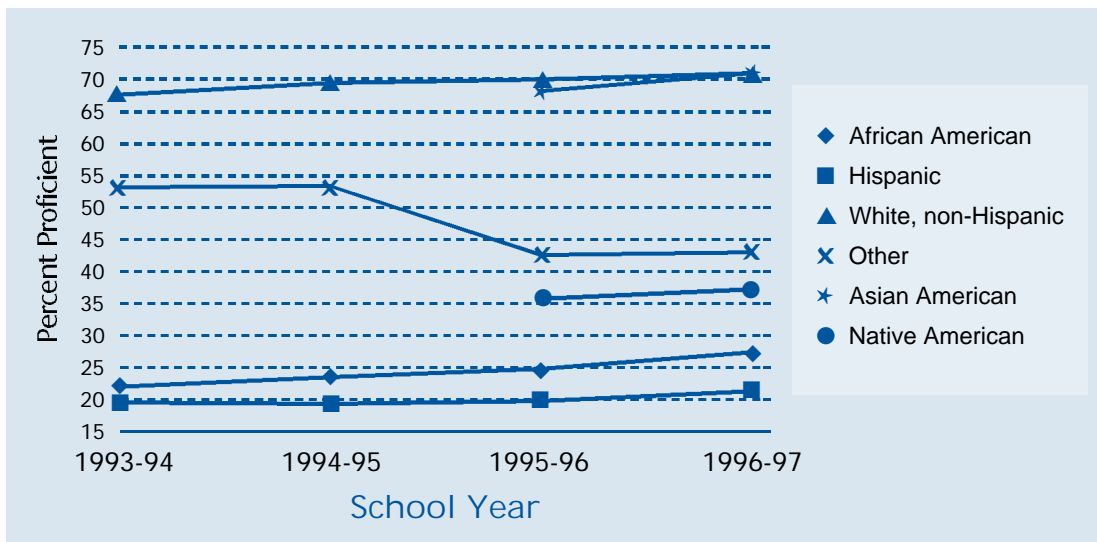


Figure 28. State of Euphoria, percent of sixth-graders with mathematics test scores in the proficient category, by school year

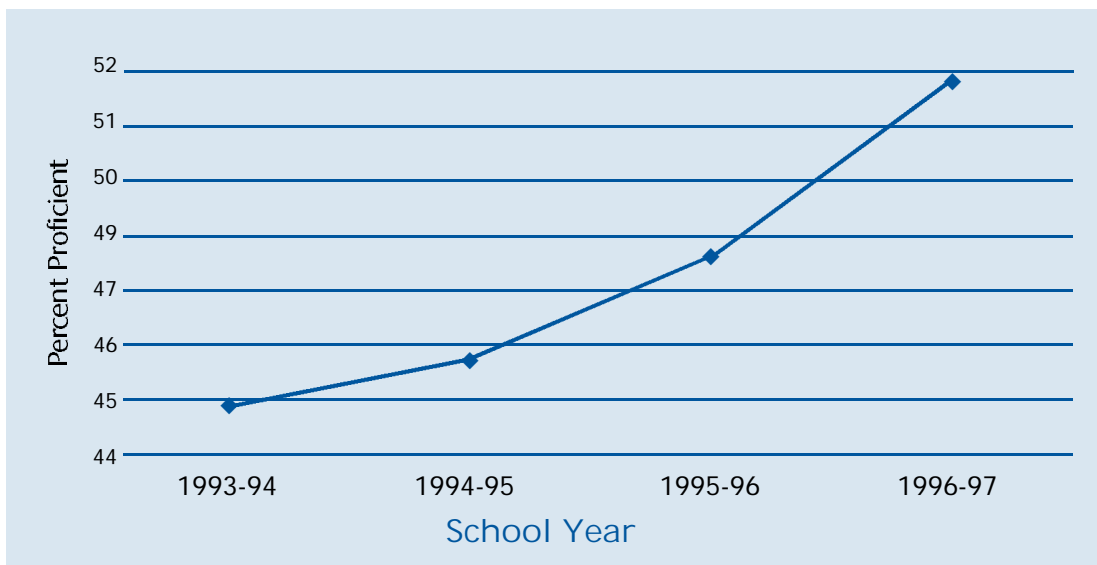


Figure 29. State of Euphoria, percent of sixth-graders with mathematics test scores in the proficient category, by racial/ethnic group and school year

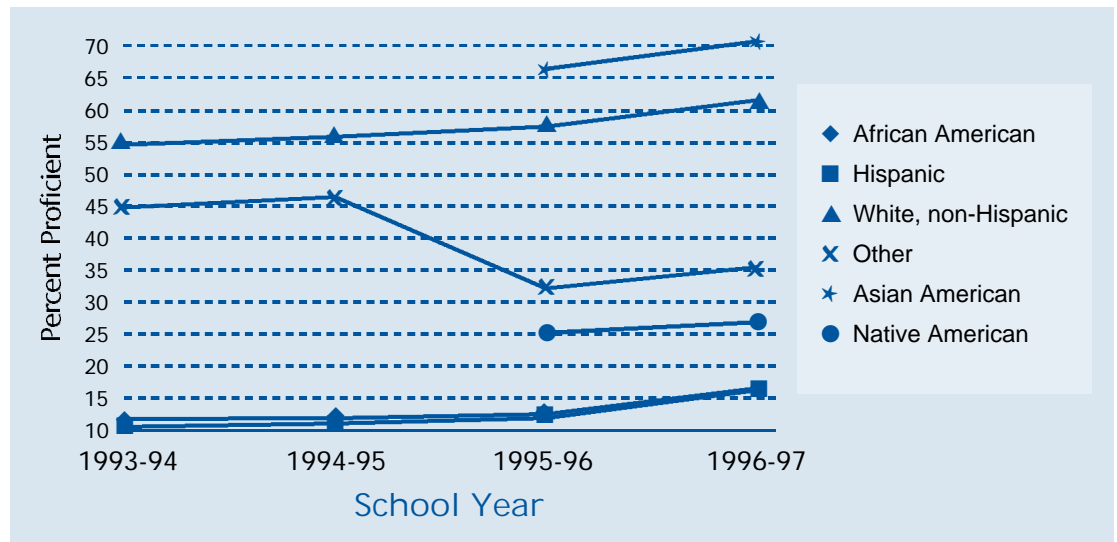


Figure 30. State of Euphoria, percent of sixth-graders with writing test scores in the proficient category, by school year

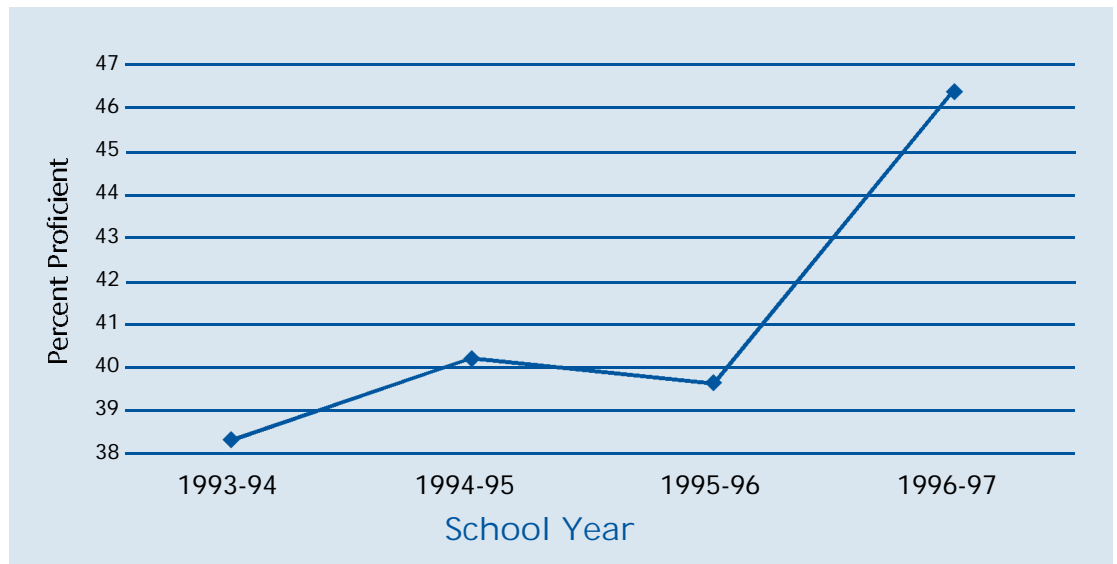
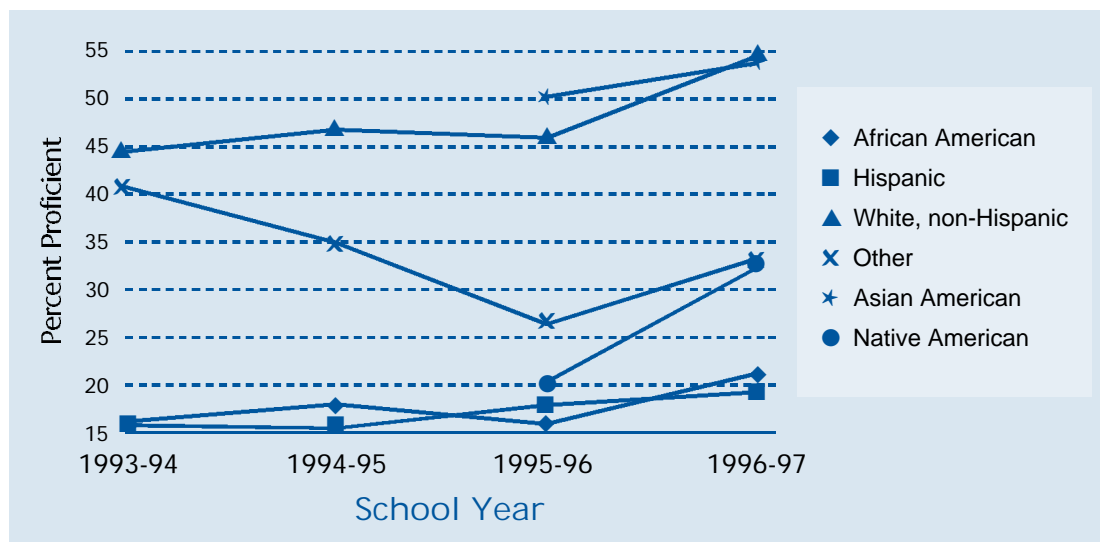


Figure 31. State of Euphoria, percent of sixth-graders with writing test scores in the proficient category, by racial/ethnic group and school year



One of the results that Drs. Hargraves and Gasper wanted to report to Euphoria’s state legislators was the average increase per school year in the percentage of students who were classified as “Proficient in reading, mathematics and writing. To calculate this figure, the researchers conducted what are called *simple linear regression analyses* using the percent of sixth-graders who were classified as Proficient in each of the subject areas as independent variables, and school year as the dependent variable. Each of the three analyses used data series that began with the 1993-94 school year. That year was coded as zero, the next school year was coded as one, and so on. Only the results for the reading test are discussed here, since the interpretation would be similar for all three subject areas. The DataDesk statistical computer program was used to complete the regression analyses, but similar results could have been obtained with the SPSS program or the Minitab program. The results for all tested sixth-graders on the reading test are shown in Figure 32.

Figure 32. Simple linear regression of percent of Euphoria’s sixth-grade students who were classified as proficient in reading on school year, from 1993-94 to 1996-97

REGRESSION OF PERCENT PROFICIENT IN READING ON YEAR OF TIME SERIES, FROM TIME 0 (1993-94)				
Dependent variable	Percent Proficient			
No Selector				
R squared = 97.2% R squared (adjusted) = 95.8%				
s = 0.2214 with 4 - 2 = 2 degrees of freedom				
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3.362	1	3.362	68.6
Residual	0.098	2	0.049	
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	57.67	0.1852	311	< 0.0001
School Year	0.82	0.099	8.28	0.0143

number of interesting conclusions can be drawn from the data shown in Figure 32.

It is dangerous to assume that the results observed during one four-year period will generalize to other time spans.

A number of interesting conclusions can be drawn from the data shown in Figure 32. First, the two numbers shown in boldface type under the heading “Coefficient” indicate, respectively, a model-based estimate of the percent of sixth-graders who were classified as “Proficient” in reading during the 1993-94 school year (shown in the row labeled “Constant”), and an estimate of the average yearly change in the percent of sixth-graders who were classified as “Proficient” in reading (shown in the row labeled “School Year”). The figure indicates that the average yearly change was 82 hundredths of a percent — just under one percent per year — with a consequent expected increase of $(4)(0.82) = 3.28$ percent across the four school years between 1993-94 and 1996-97.

Second, results in Figure 32 indicate that this average yearly change is statistically reliable beyond the 0.05 level of statistical significance. This can be determined from the value 0.0143 shown in the column headed “prob” and the row headed “School Year.” This figure indicates that, if the average yearly change in percent for the underlying population were equal to zero, the probability of observing an average yearly change of at least 0.82 would equal 0.0143. Since this probability is less than 0.05, Drs. Hargraves and Gasper concluded that the observed value was statistically significant at the 0.05 level.

The third noteworthy value in Figure 32 is the 95.8% figure just to the right of the heading “R-squared (adjusted) =.” This figure indicates that the straight-line regression model of the trend in percent of sixth-graders whose reading score is classified as “Proficient” fits the observed data very well, in that almost 96 percent of the variation in percentages across the four school years is accounted for by the model.

Two caveats must be noted when the regression analysis results are interpreted. First, since almost all of Euphoria’s sixth-graders were tested in reading each year, the distinction between observed results and results for an “underlying population” is not clear. In this case, one could argue that the sample is the entire population of interest. When one examines the statistical significance of an observed result, an underlying inference to some larger population is involved. The nature of the inference is not clear when data have been collected for the entire population. In this case, one could argue that the inference extends beyond the particular four school years for which reading data were collected. However, it is dangerous to assume that the results observed during one four-year period will generalize to other time spans. Second, when sample sizes are very large, as is the case here since about 35,000 sixth-graders were tested each year, one must be careful to distinguish between statistical significance and educational importance. It is no great comfort to conclude that a very small yearly increase in the percentage of students who were classified as “Proficient” in reading is reliably different from zero. Granted, a positive trend certainly is desirable. But a very small trend coefficient (e.g., 0.82) should not be interpreted as noteworthy just because it is statistically significant.

Drs. Hargraves and Gasper also completed simple linear regression analyses of the percent of students whose reading scores were classified as “Proficient” for all racial and ethnic groups for which four years of test data were available. The results of these analyses are shown in Figures 33 through 35 for white, non-Hispanic students, African American students and Hispanic students, respectively.

Figure 33. Simple linear regression of percent of Euphoria's white, non-Hispanic sixth-grade students who were classified as proficient in reading on school year, from 1993-94 to 1996-97

REGRESSION OF PERCENT PROFICIENT IN READING ON YEAR OF TIME SERIES, FROM TIME 0 (1993-94)				
Dependent variable	White, non-Hispanic			
No Selector				
R squared = 97.4% R squared (adjusted) = 96.1%				
s = 0.2470 with 4 - 2 = 2 degrees of freedom				
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	4.608	1	4.608	75.5
Residual	0.122	2	0.061	
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	67.91	0.2066	329	< 0.0001
School Year	0.96	0.1105	8.69	0.013

Figure 34. Simple linear regression of percent of Euphoria's African American sixth-grade students who were classified as proficient in reading on school year, from 1993-94 to 1996-97

REGRESSION OF PERCENT PROFICIENT IN READING ON YEAR OF TIME SERIES, FROM TIME 0 (1993-94)				
Dependent variable	African American			
No Selector				
R squared = 98.8% R squared (adjusted) = 98.2%				
s = 0.1449 with 4 - 2 = 2 degrees of freedom				
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	3.528	1	3.528	168
Residual	0.042	2	0.021	
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	22.89	0.1212	189	< 0.0001
School Year	0.84	0.0648	13	0.0059

Figure 35. Simple linear regression of percent of Euphoria's Hispanic sixth-grade students who were classified as proficient in reading on school year, from 1993-94 to 1996-97

REGRESSION OF PERCENT PROFICIENT IN READING ON YEAR OF TIME SERIES, FROM TIME 0 (1993-94)				
Dependent variable	Hispanic			
No Selector				
R squared = 53.0% R squared (adjusted) = 29.5%				
s = 0.4914 with 4 - 2 = 2 degrees of freedom				
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.5445	1	0.5445	2.25
Residual	0.483	2	0.2415	
Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	19.73	0.4112	48	0.0004
School Year	0.33	0.2198	1.5	0.272

Comparison of corresponding results shown in Figures 33 through 35 supports several conclusions. First, notice the dramatic difference between the boldfaced value listed in the "Coefficient" column and the "Constant" row for white, non-Hispanic students and the corresponding values for African American and Hispanic students. The linear regression model estimates that three times as many white, non-Hispanic students as African American and Hispanic students had reading scores in the "Proficient" category in 1993-94. Second, comparison of the boldfaced values listed in the "Coefficient" column and the "School Year" row shows that the estimated yearly change in percent of students with reading scores in the "Proficient" category was similar for white, non-Hispanic students and African American students (0.96 and 0.84, respectively), but was much smaller for Hispanic students (only 0.33). In fact, the yearly percent change value for Hispanic students was not reliably different from zero (note the 0.272 value in the "prob" column and the "School Year" row). Since this "prob" value is larger than 0.05, the yearly change value is not statistically significant at the 0.05 level. Thus one cannot conclude that the yearly change in the percent of Hispanic sixth-graders who were classified as Proficient in reading differed reliably from zero.

Finally, in contrast to the results for white, non-Hispanic students and African American students, the simple linear regression model does not do a good job of predicting the percent of Hispanic students whose reading scores were in the "Proficient" category during the four school years. With an R-squared (adjusted) value of only 29.5 percent, more than 70 percent of the variation in percent "Proficient" is not accounted for by the model. Either the cross-year pattern of the percent of Hispanic sixth-graders whose reading scores were in the "Proficient" category is essentially random, or something other than a straight-line growth model is needed to characterize the pattern. Four years of percentage data constitute a very small sample from which to generalize, but there does not appear to be a consistent pattern of percentages across years for Hispanic sixth-graders.

Conclusions

Title I has offered a new challenge and a new opportunity to explore students' tested achievement by disaggregating the data in new ways. The fundamental purpose of disaggregating data is to make visible the progress or lack of progress of various subgroups of students. This is an opportunity for educational decision makers to address new questions that may help them to appropriately direct attention and resources where they are most needed.

This report explores a broad sample of the types of questions that might be asked by teachers, principals, curriculum committees, school boards, legislatures, etc. It uses actual student assessment data to explore the ways in which these questions might be addressed and the types of practical and technical issues that might arise in the process. The three vignettes that form the body of the report illustrate a number of important issues in the analysis of students' achievement test data for guiding instructional planning, policy analysis and reporting to the public. Among the issues addressed are:

- Definitions of categories for disaggregation
- Minimum size of reporting groups
- Measurement error and sampling error
- Construction of confidence intervals
- Methods for displaying data: various graphs and tables
- Complexity of determining the cause of particular achievement patterns
- Appropriate interpretation of group data
- Appropriate comparisons among groups or over time

Even though the settings described in these vignettes will not mirror those of many readers, all of the issues, approaches to analysis, recommendations and cautions discussed can be generalized to any context where students' collective test performances must be interpreted to inform educational decisions.

We hope that this report will stimulate new ideas for exploring, interpreting, and using student achievement data. We hope that our readers will be empowered to analyze and discuss their data with increased curiosity and confidence that their interpretations are appropriate and fair. Ultimately, we hope that our work, in some small way, will further the Title I vision that disaggregating student achievement data and illuminating the progress of all groups of students will lead to better educational decisions that support the success of all.

Title I has offered a new challenge and a new opportunity to explore students tested achievement by disaggregating the data in new ways.

We hope that this report will stimulate new ideas for exploring, interpreting, and using student achievement data.

Appendix A: Computing a Confidence Interval Around a Sample Percentage

(Data are from Vignette 1 — Sample Percentage is 15.5; Sample Size is 679)

The standard error of a sample percentage is computed by using the following approximation:

1. Convert the percentage to a proportion by dividing it by 100:

$$15.5/100 = 0.155$$

2. Multiply the proportion by one minus the proportion:

$$0.155(1 - 0.155) = (0.155)(0.845) = 0.130975$$

3. Divide the product by the sample size used to compute the percentage:

$$0.130975/679 = 0.0001928939617083$$

This is called the *variance error of estimate*.

4. Calculate the square root of the quotient:

$$(0.0001928939617083) = 0.0139$$

This is the *standard error of estimate*.

5. To calculate a 95 percent confidence interval around the sample proportion, multiply the standard error of estimate by 1.96. Then add this product to the sample proportion to get an upper confidence limit and subtract this product from the sample proportion to get a lower confidence limit:

$$(0.0139)(1.96) = 0.027$$

$$0.155 + 0.027 = 0.182$$

This is the upper confidence limit around the sample proportion.

$$0.155 - 0.027 = 0.128$$

This is the lower confidence limit around the sample proportion.

6. Multiply the upper and lower confidence limits by 100 to convert them to percentages:

$$(0.182)(100) = 18.2 \text{ percent}$$

This is the upper confidence limit around the sample percentage.

$$(0.128)(100) = 12.8 \text{ percent}$$

This is the lower confidence limit around the sample percentage.

Appendix B: Using Confidence Intervals to Compare Two Sample Percentages and Conducting a Hypothesis Test on the Difference Between Two Percentages

(Data from Vignette 1 are Used)

Two approaches to investigating the statistical reliability of the difference between two sample percentages will be shown in this appendix. First, we'll compute a confidence interval around each sample percentage and see if the confidence intervals overlap. Second, we'll test the null hypothesis that, in populations from which the samples of students were selected, corresponding population proportions are equal. To begin with, using the approach described in Appendix A, we will compute a 95 percent confidence interval around the sample proportions of fourth-graders whose mathematics scores placed them in the Proficient or Advanced categories. Then we'll compare those confidence intervals.

The proportion of boys with scores in the Proficient or Advanced categories was $56.1/100 = 0.561$. The number of fourth-grade boys tested was 337. The standard error of this proportion is therefore $(0.561)(1 - 0.561)/337 = 0.027$. The 95 percent confidence interval around the percentage of boys with mathematics scores in the Proficient or Advanced categories is therefore $100[0.561 + (1.96)(0.027)] = 100(0.614) = 61.4$ percent for an upper confidence limit, to $100[0.561 - (1.96)(0.027)] = 100(0.508) = 50.8$ percent for a lower confidence limit. We are 95 percent confident that the percentage of fourth-grade boys with mathematics scores in the Proficient or Advanced categories is between 50.8 percent and 61.4 percent.

By the same logic, the proportion for girls was $49.4/100 = 0.494$. The number of fourth-grade girls tested was 342. The standard error of this proportion was therefore $(0.494)(1 - 0.494)/342 = 0.027$. The 95 percent confidence interval around the percentage of girls with mathematics scores in the Proficient or Advanced categories is therefore $100[0.494 + (1.96)(0.027)] = 100(0.547) = 54.7$ percent for an upper confidence limit, to $100[0.494 - (1.96)(0.027)] = 100(0.441) = 44.1$ percent for a lower confidence limit. We are 95 percent confident that the percentage of fourth-grade girls with mathematics scores in the Proficient or Advanced categories is between 44.1 percent and 54.7 percent.

Notice that these two confidence intervals overlap substantially. Thus it could well be the case that the population percentage of girls with mathematics scores in the Proficient or Advanced categories was actually higher than the population percentage of boys with mathematics scores in the Proficient or Advanced categories. We must conclude that the difference between the two sample percentages is not statistically reliable.

The hypothesis testing approach begins with the proposition that the population percentages of fourth-grade boys and girls with mathematics scores in the Advanced category is the same. We then analyze the sample data to determine whether the observed sample difference between the percentages for girls and boys is so large that the proposition is highly unlikely. If we conclude in the affirmative, we "*reject the hypothesis*" that the population percentages are equal.

To carry out the hypothesis test, we compute something called a *test statistic*. In this case, the test statistic is found by subtracting the observed percentage for girls from the observed percentage for boys, and then dividing that difference by its standard error. The standard error of the difference is merely the square root of the sum of the error variance of the percentage for boys and the error variance of the percentage for girls. The computational steps are as follows.

1. Calculate the difference between the sample proportions for boys and girls:

$$\text{Difference} = 0.561 - 0.494 = 0.067$$

2. Calculate the error variance of the proportion for boys:

$$\text{Error var}_{\text{boys}} = (0.561)(1 - 0.561)/337 = 0.0007307$$

3. Calculate the error variance of the proportion for girls:

$$\text{Error var}_{\text{girls}} = (0.494)(1 - 0.494)/342 = 0.0007309$$

4. Sum the two error variances and take the square root:

$$(0.0007307 + 0.0007309) = 0.038231$$

This is the standard error of the difference.

5. Calculate a test-statistic as the difference between the sample proportions divided by the standard error of the difference:

$$t = 0.067 / 0.038231 = 1.753$$

6. Compare this difference to a critical value from a table of the normal distribution which, in this case, for a *level of significance*¹⁰ of 0.05 will be a value of 1.96. Since the observed value of the test statistic is only 1.753, we conclude that the sample proportions for boys and girls do not differ at the 0.05 level of statistical significance. In other words, the corresponding sample percentages do not differ reliably.

¹⁰ The level of significance is, in this case, the probability of finding a difference between sample proportions equal to or greater than the observed difference, when the two population proportions are equal. For a level of significance of 0.05 in the current example, the test statistic would have to be at least 1.96. Since it is smaller, we retain the hypothesis that the two population proportions (percentages) are equal.

Appendix C: Constructing a Confidence Interval Around a Sample Average

(Data from Vignette 1 are Used)

To construct a 95 percent confidence interval around a sample average (also called a sample *mean*), complete the following steps:

1. Look up or compute the standard deviation of scores for the group that gave rise to the average. The easiest way to compute a standard deviation is to use a packaged computer program, such as SPSS, SAS, Minitab, or DataDesk. If you are going to compute a standard deviation by hand (a lot of unnecessary work in this day of desktop computers), you would (a) calculate the average of the scores, (b) subtract each score from the average, (c) square the difference between each score and the average, (d) add up the squared differences, (e) divide the sum by the sample size minus one, and (f) calculate the square root of the resulting quotient.

In the current example, Dr. Clinton used the SPSS program to find the standard deviation of mathematics test scores for each group of fourth-graders. Her results were as shown in the following table:

Group	Sample Size	Standard Deviation of Grade 4 Mathematics Scores
African American	146	21.5467
Hispanic	88	15.6731
White, non-Hispanic	335	12.7672
Other	41	17.5876
Asian American	16	15.6072
Native American	2	11.3137

2. Divide each standard deviation by the square root of its respective sample size to form *standard errors of the mean*:

For example, for African American fourth-graders, the standard error of the mean is given by $21.5467 / \sqrt{146} = 1.7832$.

3. Multiply each standard error of the mean by 1.96, to determine half the width of the confidence interval:

For example, for African American fourth-graders, the product would be $(1.7832)(1.96) = 3.4951$

4. Add this product to the average to form the upper confidence limit and subtract the product from the average to form the lower confidence limit:

For example, for African American fourth-graders, the upper and lower confidence limits around the sample average would be $86 + 3.495 = 89.495$ and $86 - 3.495 = 82.505$. We would therefore say that we are 95 percent confident that the interval 82.5 to 89.5 (rounding appropriately) includes the population mean mathematics score for Lincoln Public Schools' African American fourth-graders.

Appendix D: Testing the Null Hypothesis that Two Population Averages are Equal

To test the hypothesis that two population averages (also called population *means*) are equal (an example of a *null hypothesis*), complete the following steps:

1. Look up or compute the standard deviation of scores for the group that gave rise to the first sample average. The easiest way to compute a standard deviation is to use a packaged computer program, such as SPSS, SAS, Minitab, or DataDesk. If you are going to compute a standard deviation by hand (a lot of unnecessary work in this day of desktop computers), you would (a) calculate the average of the scores, (b) subtract each score from the average, (c) square the difference between each score and the average, (d) add up the squared differences, (e) divide the sum by the sample size minus one, and (f) calculate the square root of the resulting quotient.
2. Look up or compute the standard deviation of scores for the group that gave rise to the second sample average.
3. Square the standard deviation of scores for the first sample by multiplying that standard deviation by itself. This produces what is called a *sample variance*.
4. Square the standard deviation of scores for the second sample. This results in the sample variance for the second sample.
5. Multiply the first sample variance by the size of the first sample, minus one. For example, if the size of the first sample was 25, you would multiply that first sample variance by 24.
6. Multiply the sample variance for the second sample by the size of the second sample, minus one.
7. Add the values calculated in steps 5 and 6 together, and then divide the total by the sum of the two multipliers. For example, if you had data that looked like those in the following table, you would do the following, to this point:

Statistic/ Sample	Sample Size	Average	Standard Deviation
1995-96	78	43.5	6.88
1996-97	120	47.7	8.08

The variance of the 1995-96 sample was $(6.88)^2 = 47.33$.

The variance of the 1996-97 sample was $(8.08)^2 = 65.29$.

For the first sample, the variance multiplied by the sample size minus one yields $(47.33)(78 - 1) = 3644.41$.

For the second sample, the variance multiplied by the sample size minus one yields $(65.29)(120 - 1) = 7769.51$.

The sum of these products is $3644.41 + 7769.51 = 11,413.92$.

This sum, divided by the sum of the two multipliers is $11,413.92 / [(78 - 1) + (120 - 1)] = 11,413.92 / 196 = 58.23$.

8. Calculate the square root of the value computed in Step 7 (this is called the *standard error of the mean difference*) = $(58.23) = 7.63$.

9. Calculate the difference between the two sample averages = $47.7 - 43.5 = 4.2$
10. Divide the difference between sample averages by the standard error of the mean difference = $4.2/7.63 = 0.55$. This fraction is called a *test statistic*, and in this case is denoted by the lower-case letter "t."
11. The t-statistic is compared to a critical value that is normally read from a table of Student's t-distribution. Since the sum of the two sample sizes is so large (in this case far greater than 50), the value 1.96 can be used as a critical value. Since the computed t-value (0.55) is less than 1.96, we would retain (not reject) the null hypothesis that the population averages for 1995-96 and 1996-97 are equal. Only when a test statistic equals or exceeds a critical value do we reject a null hypothesis of equality of two population averages.

Appendix E: Testing Simultaneous Null Hypotheses on Differences Between Pairs of Population Means

When the difference between one pair of sample averages is tested for statistical significance and the sample of cases that produced each mean is, say, larger than 50, one is reasonably safe using the value 1.96 as a criterion (critical value) for declaring the resulting t-statistic to be “statistically significant at the 0.05 level” (see Appendix B for details). However, when a number of pairs of averages are tested for statistical significance, the criterion for significance must be adjusted upward, depending on the number of averages that are being compared.

One approach to handling this problem makes use of a mathematical result known as Bonferroni’s inequality and statistical tables produced by J. R. Bailey and published in the 1977 *Journal of the American Statistical Association* (Volume 72, pp. 469-478). These tables contain the appropriate criterion values (critical values) for use in judging the statistical significance of t-statistics at the 0.05 level of significance, depending on the number of pairs of averages that are being compared and the sizes of the samples used to compute the averages. Only a few of the criterion values are shown here for illustrative purposes. The table shows that, as the total sample size (meaning the sum of the sample sizes used to compute each pair of averages) gets larger, the criterion value gets smaller and, as the number of pairs of averages being compared is increased, the criterion value gets larger. Numbers of pairs of averages between 6 and 10 are not shown because the criterion value doesn’t increase much in that range. (If you are comparing seven pairs of averages, for example, you would be safe using the criterion value for 10 pairs.) Sample sizes larger than 100 are not shown because the criterion (critical) values do not get much smaller for larger overall sample sizes. (For example, for 10 comparisons between pairs of sample averages and a total sample size of 100, the critical value is 2.87, compared to a critical value of 2.81 when the total sample size is 1000.)

Table E-1. Approximate criterion values (critical values) to use when testing the statistical significance at the 0.05 level between pairs of sample averages.

Number of Pairs of Averages Compared	Total Sample Size for a Pair of Averages	Approximate Criterion Value
1	25	2.07
2	25	2.40
3	25	2.58
4	25	2.71
5	25	2.81
6	25	2.89
10	25	3.10
1	50	2.01
2	50	2.31
3	50	2.49
4	50	2.60
5	50	2.69
6	50	2.76
10	50	2.95
1	100	1.98
2	100	2.28
3	100	2.43
4	100	2.54
5	100	2.63
6	100	2.69
10	100	2.87

Appendix F: Computing the Effect Size of the Difference Between Two Sample Averages

To compute an effect size corresponding to the difference between two sample averages, complete the following steps:

1. Subtract the smaller sample average from the larger one.

When comparing the average Algebra I score for boys in 1996-97 with the average in 1995-96, the difference was found to be $47.7 - 43.5 = 4.2$ points.

2. Calculate the standard deviation of each sample average. Note the size of the sample used to compute each sample average.

The 1995-96 standard deviation was 6.88, based on a sample size of 78 boys who completed the Algebra I test. The 1996-97 standard deviation was 8.08, based on a sample of 120 boys.

3. Square the first standard deviation to form the variance of the first sample.

The variance of the 1995-96 sample was $(6.88)^2 = 47.34$.

4. Square the second standard deviation to form the variance of the second sample.

The variance of the 1996-97 sample was $(8.08)^2 = 65.28$.

5. Multiply the first sample variance by its sample size minus 1, and do the same thing to the second sample variance.

For the first sample: $(47.34)(78 - 1) = 3645$

For the second sample: $(65.28)(120 - 1) = 7768$

6. Sum the two products and divide by the sum of the sample sizes minus two. This forms a weighted average of the two variances, weighted by their approximate sample sizes.

$(3645 + 7768)/(78 + 120 - 2) = 58.23$

7. Calculate the square root of the variance calculated in Step 6. This is the pooled standard deviation of the two samples.

$(58.23) = 7.63$

8. Divide the difference between the sample averages that you calculated in Step 1 by the pooled standard deviation that you calculated in Step 7. The result will be the estimated effect size.

The effect size was $4.2/7.63 = 0.55$. This is, according to Cohen, a moderate effect size.