NAEP Trends: Main NAEP vs. Long-Term Trend

Albert E. Beaton
Boston College

James R. Chromy
Research Triangle Institute

December 2010
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics (NCES).

This report is based on work that was jointly supported by NCES (contract ED-01-CO-0026-005) and the National Science Foundation (grant 454755). Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of NCES or the National Science Foundation.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/ 843-8100
Fax: 650/ 843-8200

Acknowledgments

We wish to acknowledge the helpful encouragement and suggestions of our discussants, Richard Duran, David Grissmer, Larry Hedges, and Ina Mullis as well as the other members of the NVS Panel.

We are also indebted AIR editors Susan Bratten, Holly Baker, and Heather Banks and desktop publishers Karen Ward and Patricia Louthian.

We further wish to acknowledge the help, support, and encouragement of Fran Stancavage, Project Director, and George Bohrnstedt, Chair of the NVS Panel.

CONTENTS

List of Tables

List of Figures

1. Introduction

This study addresses the measurement of mathematics and reading trends by two separate National Assessment of Educational Progress (NAEP) statistical series. The first series is based on the Long-Term Trend assessment, which dates back to 1969 and assesses students in three age groups: 9-year-olds, 13-year-olds, and 17-year-olds. The other series, now described as Main NAEP, assesses students by grade: grade 4, grade 8, and grade 12. This report compares Long-Term Trend and Main NAEP results and is limited to mathematics since 1990 and reading since 1992—the period that corresponds to the implementation of Main NAEP in its current form. The focus is on average levels of change over the period since 1990 and on a comparison of the two data series over that time period. The Nation's Report Card reports and other National Center for Education Statistics (NCES) reports provide other summary statistics and direct year-to-year comparisons to measure short-term trends within each series.

The estimated trends from the two assessments are published separately by NCES, and readers are advised not to compare the two. However, because the trend lines may be used for policy formation, differences in trend lines can become important issues. For example, *The New York Times* (Dillon 2009a) carried a discussion of the 2008 study of trends in academic progress, based on Long-Term Trend (NCES 2009a).

A second *article in The New York Times* (Dillon 2009b) focused on regional and state differences in the racial achievement gap, based on Main NAEP. The source for this article was an NCES report that studied achievement gaps using data from both series, but restricted to public school students in grades 4 and 8 for Main NAEP and to ages 9 and 13 for Long-Term Trend (Vanneman et al. 2009).

Earlier, Hauser, Brown and Prosser (1997, pp. 220–225) discussed the two trend measures and some of the general weaknesses of NAEP data. Of particular interest is their assertion that real trends can sometimes be obscured by shifts in demographic distributions—a topic revisited in this report.

The objectives of this research are to (a) compare the trend lines after some adjustments for level and scale only and determine if and how they differ; (b) describe the methodology of each assessment and identify similarities and differences; and (c) attempt to explain any observed differences based on comparable subsets or on special analysis. Subsets considered include grade within age, age within grade, and racial/ethnic groups.

This study arose from a NAEP Validity Studies Panel inquiry into potential issues affecting NAEP validity. As part of this inquiry, the panel proposed a two-component study to compare the separate NAEP statistical series. The first component was to address analysis of trends from the two series. The second component was to examine test-construct issues that may contribute to any divergence of the measured trends. Only the first component is addressed in this report. The second component was addressed, in part, by Dickinson and colleagues (2006).

The following chapters address these topics:

Chapter 2: Main NAEP and Long-Term Trend trends from 1990 to 2009

Chapter 3: Descriptions of the Main NAEP and Long-Term Trend surveys since 1990, with identification of similarities and differences

Chapter 4: Comparisons of trend measures limited to more nearly comparable populations based on modal age and modal grade

Chapter 5: An examination of population shifts and comparisons of subpopulation estimates from each of the NAEP surveys

Chapter 6: The effects of population shifts

Chapter 7: Conclusions

For discussion purposes, we use the term "elementary school" level to describe comparisons between statistics for Main NAEP grade 4 and Long-Term Trend age 9. Similarly, we use the term "middle school" to refer to comparisons between grade 8 and age 13 and the term "high school" to refer to comparisons between grade 12 and age 17.

2. NAEP Trends Since 1990

When measured by a simple linear regression model over the time period studied, several of the Main NAEP and Long Term Trend series show statistically significant positive trends. None exhibited a statistically detectable negative trend.

Differences between Main NAEP and Long-Term-Trend average annual change parameters are statistically significant only for mathematics at the elementary and middle school levels. Comparisons of the trend lines are complicated by (a) the fact that the two assessments use slightly different scales (discussed in Chapter 3), (b) bridge studies have shifted the level of each series somewhat, and (c) the two assessments have noncoinciding schedules. This chapter presents comparisons of the trend lines for the following:

- All grade 4 students from Main NAEP vs. all age 9 students from Long-Term Trend
- All grade 8 students from Main NAEP vs. all age 13 students from Long-Term Trend
- All grade 12 students from Main NAEP vs. all age 17 students from Long-Term Trend

The scale scores for Main NAEP were transformed to have the same starting values and the same population standard deviation as Long-Term Trend.[1] The problem of noncoinciding schedules was addressed by basing the comparison on the slope coefficient of a fitted linear trend line after putting both assessments on the same scale. A shift parameter[2] was added to the linear trend model to allow a constant slope shifted with introduction of the changes in exclusion and accommodations procedures and other methodologies studied in years with bridge samples. Chapter 3 gives background information on these issues.

Trend comparisons for mathematics are given in Tables 2.1 through 2.3 and Figures 2.1 through 2.3. Trend comparisons for reading are given in Tables 2.4 through 2.6 and Figures 2.4 through 2.6. Statistical model-based estimates and test results for all of the trend comparisons are summarized in Table 2.7.

Mathematics Trend Comparisons

Elementary and middle school data show positive mathematics score improvement on both assessments, with higher annual increases for the Main NAEP assessment. At the high school level, only Main NAEP shows a marginally significant

[1] Main NAEP mathematics scores were adjusted to equate the 1990 means on the two assessments, and the scale was adjusted on the basis of an empirical estimate of the population standard deviation so that subsequent adjusted estimates were on the Long-Term Trend scale for both series. The same procedure was applied to reading for 1992. Appendix A provides more details.

[2] See discussion of "Model-Based Results" later in this chapter.

($p = 0.0747$) positive trend, but the two trend lines cannot be declared to be statistically different.

Elementary School Level. Table 2.1 shows annual average scores for Main NAEP grade 4 and Long-Term Trend age 9. The Main NAEP trend data are shown both before and after applying the linear transformation to put its results on the same scale as the Long-Term Trend data. Figure 2.1 plots the transformed Main NAEP data and the Long-Term Trend data. The data points are plotted as circles for assessments that **do not use** the new accommodation and inclusion rules and as triangles for assessments that **do use** the new rules. The linear trend lines based on a linear regression model are plotted as dashed lines before the introduction of the new accommodation rules and as solid lines after.

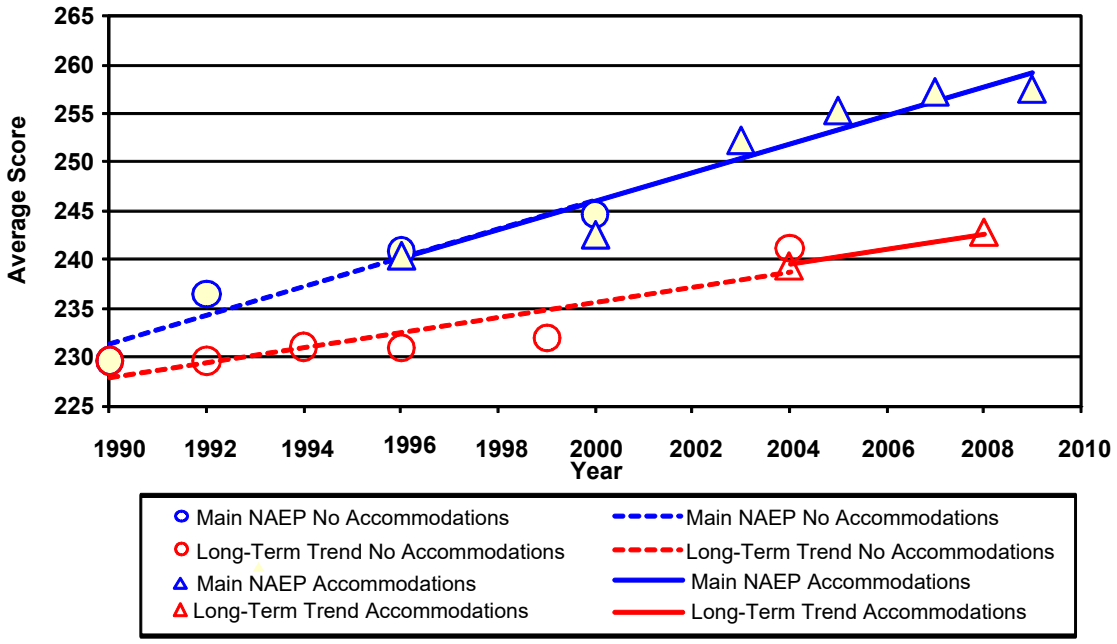
Both assessments show positive and statistically significant average annual change measures. The Main NAEP rate of increase, 1.48 scale points per year, is greater than that for Long-Term Trend, 0.78 scale points per year.

Table 2.1. Average mathematics scores by assessment year: Main NAEP grade 4 and Long Term Trend age 9

Assessment	Accom-modations Permitted	Assessment Year											
		1990	1992	1994	1996	1999	2000	2003	2004	2005	2007	2008	2009
Main NAEP Grade 4	Yes				224		226	235		238	240		240
	No	213	220		224		228						
Main NAEP Grade 4 Transformed	Yes				240		243	252		255	257		258
	No	230	237		241		245						
Long-Term Trend Age 9	Yes								239			243	
	No	230	230	231	231	232			241				

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1990 equivalent to the Long-Term Trend.

Figure 2.1. Average mathematics scores by assessment year: Main NAEP grade 4 (transformed) and Long-Term Trend age 9



Note: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

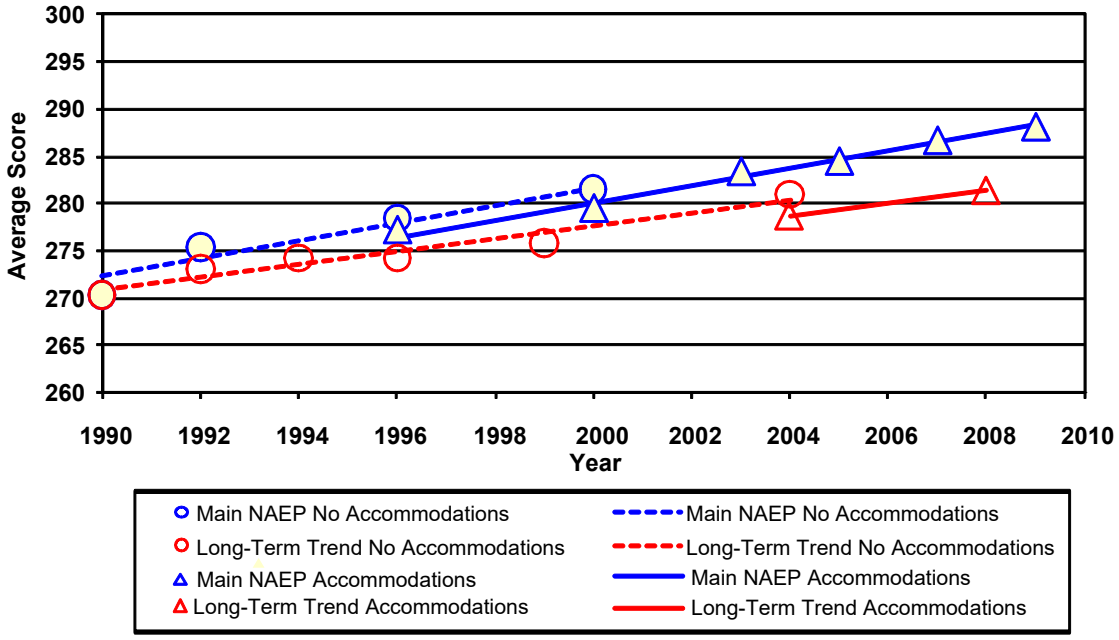
Middle School Level. Table 2.2 shows the annual average scores for Main NAEP grade 8 and Long-Term Trend age 13. The transformed Main NAEP and the Long-Term Trend data are shown in Figure 2.2. Both assessments show positive average annual increases. The Main NAEP rate of increase, 0.92 scale points per year, is greater than that for Long-Term Trend, 0.68 scale points per year.

Table 2.2. Average mathematics scores by assessment year: Main NAEP grade 8 and Long-Term Trend age 13

Assessment	Accom-modations Permitted	Assessment Year										
		1990	1992	1994	1996	1999	2000	2003	2004	2005	2007	2008	2009
Main NAEP Grade 8	Yes				270		273	278		279	281		283
	No	263	268		272		275						
Main NAEP Grade 8 Transformed	Yes				277		280	283		284	287		288
	No	270	275		279		282						
Long-Term Trend Age 13	Yes								279			281	
	No	270	273	274	274	276			281				

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1990 equivalent to the Long-Term Trend.

Figure 2.2. Average mathematics scores by assessment year: Main NAEP grade 8 (transformed) and Long-Term Trend age 13



Note: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

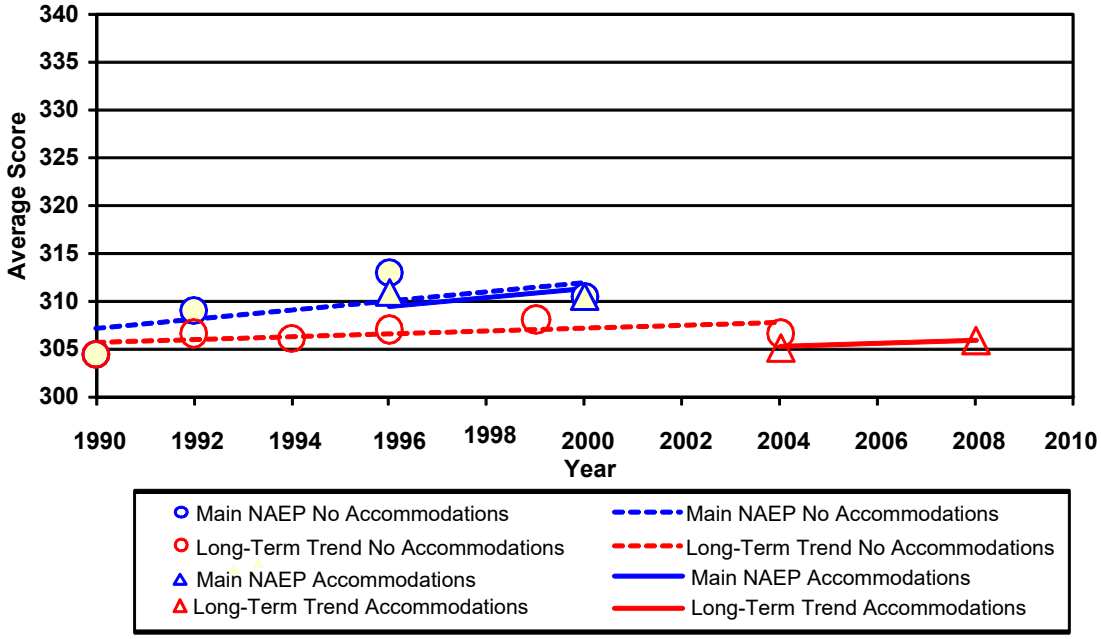
High School Level. Table 2.3 shows the annual average scores for Main NAEP grade 12 and Long-Term Trend age 17. The transformed Main NAEP and the Long-Term Trend data are shown in Figure 2.3. The estimated average annual increases are not statistically significant. The difference between the two trend lines is also not statistically significant. Note that the Main NAEP data exclude any data collected after 2000 due to lack of comparability issues. (See Chapter 3 for a discussion of assessment instrument frameworks.)

Table 2.3. Average mathematics scores by assessment year: Main NAEP grade 12 and Long-Term Trend age 17

Assessment	Accom-modations Permitted	Assessment Year							
		1990	1992	1994	1996	1999	2000	2004	2008
Main NAEP Grade 12	Yes				302		301		
	No	294	299		304		301		
Main NAEP Grade 12 Transformed	Yes				311		311		
	No	305	309		313		311		
Long-Term Trend Age 17	Yes							305	306
	No	305	307	306	307	308		307	

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1990 equivalent to the Long-Term Trend.

Figure 2.3. Average mathematics scores by assessment year: Main NAEP grade 12 (transformed) and Long-Term Trend age 17



Note: The 1990 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP (MT) transformed estimate.

Reading Score Comparisons

Long-term annual improvement estimates can be declared statistically significant only at the elementary level, where both assessments show positive change. The rates of annual change for the two assessments cannot be shown to be different at any of the three levels.

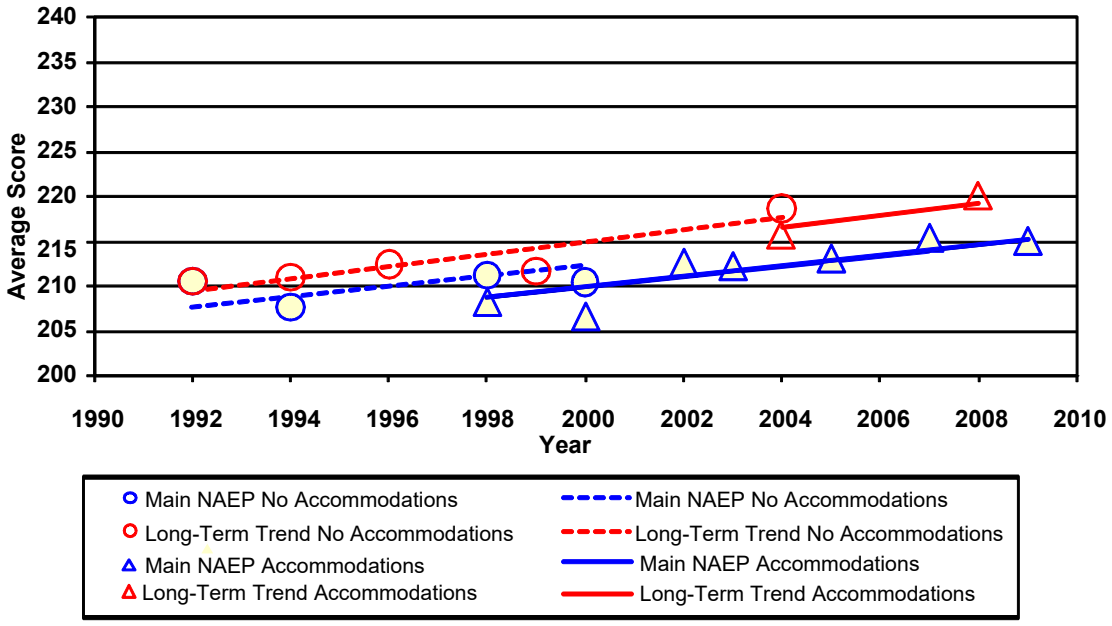
Elementary School Level. Table 2.4 shows the annual average reading scores for Main NAEP grade 4 and Long-Term Trend age 9. The transformed Main NAEP and the Long-Term Trend data are plotted in Figure 2.4. Both assessments show statistically significant positive average annual increases. The Main NAEP rate of increase, 0.48 scale points per year, is not statistically different from the Long-Term Trend rate of increase, 0.68 scale points per year.

Table 2.4. Average reading scores by assessment year: Main NAEP grade 4 and Long-Term Trend age 9

	Accom-modations Permitted	Assessment Year											
		1992	1994	1996	1998	1999	2000	2002	2003	2004	2005	2007	2008	2009
Main NAEP Grade 4	Yes				215		213	219	218		219	221		221
	No	217	214		217		217							
Main NAEP Grade 4 Trans-formed	Yes				208		207	213	212		213	215		215
	No	211	208		211		210							
Long-Term Trend Age 9	Yes									216			220	
	No	211	211	212		212				219				

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1992 equivalent to the Long-Term Trend.

Figure 2.4. Average reading scores by assessment year: Main NAEP grade 4 (transformed) and Long-Term Trend age 9



Note: The 1992 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

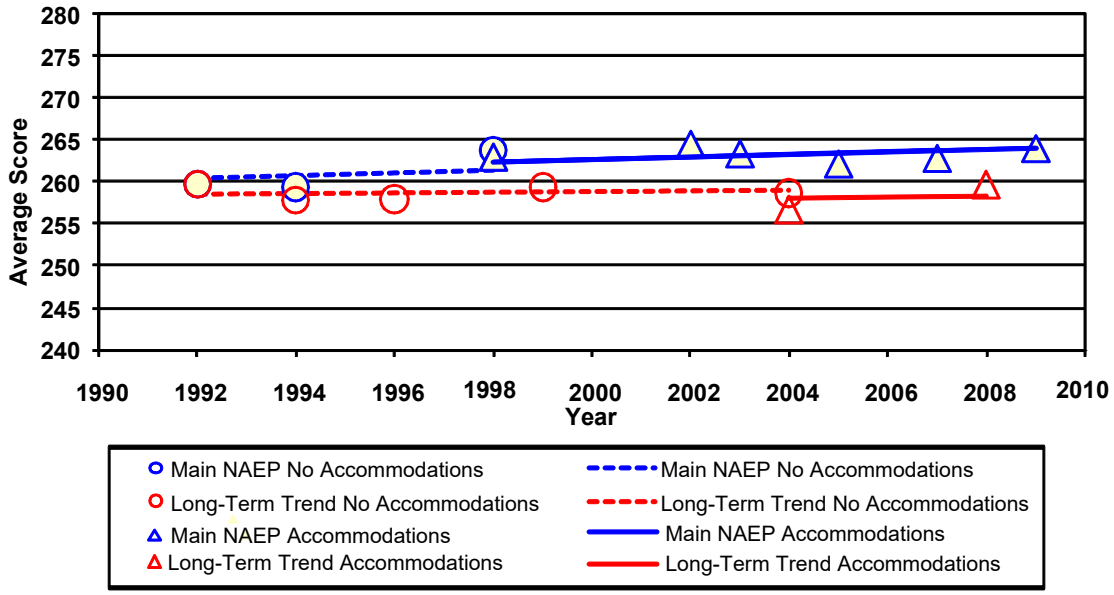
Middle School Level. Table 2.5 shows the annual average reading scores for Main NAEP grade 8 and Long-Term Trend age 13. The transformed Main NAEP and the Long-Term Trend data are plotted in Figure 2.5. Neither assessment shows statistically significant average annual change. The Main NAEP estimated rate of increase, 0.20 scale points per year, is not statistically different from the Long-Term Trend estimated rate of increase, 0.05 scale points per year.

Table 2.5. Average reading scores by assessment year: Main NAEP grade 8 and Long-Term Trend age 13

	Accom-modations Permitted	Assessment Year											
		1992	1994	1996	1998	1999	2002	2003	2004	2005	2007	2008	2009
Main NAEP Grade 8	Yes				263		264	263		262	263		264
	No	260	260		264								
Main NAEP Grade 8 Trans-formed	Yes				263		265	263		262	263		264
	No	260	259		264								
Long-Term Trend Age 13	Yes								257			260	
	No	260	258	258		259			259				

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1992 equivalent to the Long-Term Trend.

Figure 2.5. Average reading scores by assessment year: Main NAEP grade 8 (transformed) and Long-Term Trend age 13



Note: The 1992 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

High School Level. Table 2.6 shows the annual average reading scores for Main NAEP grade 12 and Long-Term Trend age 17. The transformed Main NAEP and the Long-Term Trend data are plotted in Figure 2.6. Neither assessment shows statistically significant average annual change. The Main NAEP estimated rate of

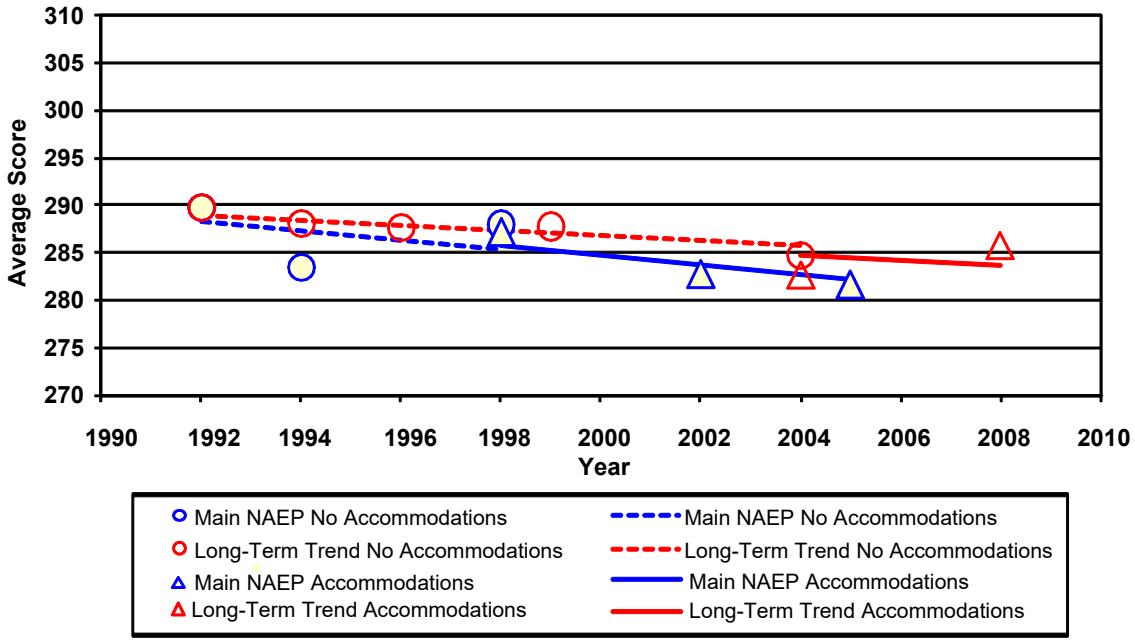
change, –0.49 scale points per year, is not statistically different from the Long-Term Trend estimated rate of change, –0.27 scale points per year.

Table 2.6. Average reading scores by assessment year: Main NAEP grade 12 and Long-Term Trend age 17

Assessment	Accom-modations Permitted	Assessment Year								
		1992	1994	1996	1998	1999	2002	2004	2005	2008
Main NAEP Grade 12	Yes				290		287		286	
	No	292	287		291					
Main NAEP Grade 12 Transformed	Yes				287		283		282	
	No	290	283		288					
Long-Term Trend Age 17	Yes							283		286
	No	290	288	288		288		285		

Note: The Main NAEP data are shown as reported and as transformed to have level and standard deviation in 1992 equivalent to the Long-Term Trend.

Figure 2.6. Average reading scores by assessment year: Main NAEP grade 12 (transformed) and Long-Term Trend age 17



Note: The 1992 data points are shown only in red but represent both the Long-Term Trend estimate and the Main NAEP transformed estimate.

Model-Based Results

An ordinary least squares model for linear trend was fit to the trend data from each statistical series. The model was a simple regression against year, with separate coefficients for the Main NAEP and Long-Term Trend assessments. It included a shift variable for the years with bridge studies associated with accommodations and

increased inclusion policies. The models for the two assessments were fit simultaneously to facilitate testing of the equivalence of the slope parameters associated with year. This slope parameter can be viewed as the average annual change in assessment scores. The combined model can be represented as

$$E(Y) = X_1(\beta_{11} + \beta_{12}X_{12} + \beta_{13}X_{13}) + (1 - X_1)(\beta_{21} + \beta_{22}X_{22} + \beta_{23}X_{23})$$

where Y = average score, $X_1 = 1$ for Main NAEP and $X_1 = 0$ for Long-Term Trend, X_{12} and X_{22} are indicator variables set to 1 only for assessments conducted with accommodations and increased inclusion, and X_{13} and X_{23} represent the year of the assessment. Table 2.7 summarizes the parameter estimates.

The discussion above covers the statistical significance of the slopes and of their differences. Note also that on the basis of this analysis, none of the accommodation effects was statistically significant.

For this evaluation of trend, the choice of models was limited to a simple linear trend over the study period. Other models might, in fact, fit the data better, but the linear model appears adequate to capture the directionality of change and the consistency between the two assessment types. Lack of a statistically detectable linear trend should not be construed to indicate no differences among years or to contradict findings of statistically significant differences in year-to-year comparisons in any NAEP report. These types of differences can exist even when the longer trend is relatively flat and not statistically significant.

Table 2.7. Estimated linear model parameters by subject and level

Level and Subject	Type	Intercept (Year = 2000)	Shift for Accommo-dation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary Mathematics	Main NAEP	246.06	–0.06	1.48	<0.0001			
	Long-Term Trend	235.68	0.74	0.78	0.0010	–0.70	0.0137	0.97
Middle School Mathematics	Main NAEP	281.55	–1.43	0.92	<0.0001			
	Long-Term Trend	277.64	–1.58	0.68	<0.0001	–0.25	0.0419	0.98
High School Mathematics	Main NAEP	312.01	–0.62	0.48	0.0747			
	Long-Term Trend	307.21	–2.39	0.15	0.3990	–0.33	0.2767	0.67
Elementary Reading	Main NAEP	211.70	–1.52	0.48	0.0018			
	Long-Term Trend	214.91	–1.11	0.68	0.0033	0.20	0.3846	0.81
Middle School Reading	Main NAEP	262.11	0.45	0.20	0.0657			
	Long-Term Trend	258.90	–1.01	0.05	0.7289	–0.15	0.4184	0.79
High School Reading	Main NAEP	284.42	0.28	–0.49	0.2087			
	Long-Term Trend	286.79	–0.91	-0.27	0.2964	0.22	0.6190	0.58

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the six models.

3. The Two Data Series

The purpose of this chapter is to review and document methodological similarities and differences between the two data series over the period 1990–2009. Each data series makes an effort to preserve trend measurement across time whenever possible by limiting changes in methodology and content. But change is inevitable if the data are to maintain relevance to the current national and world circumstances. Each series has at least one point at which significant changes were made and bridge assessments were carried out.

Methodological issues considered include population definitions, timing of the assessment during the school year, data collection methods, accommodation and inclusion, private school coverage, assessment instrument frameworks, reporting by standards or benchmarks, coordination with state assessments, scale scores, treatment of race/ethnicity, poststratification to external data, response rates, and use of bridge studies. Some topics that do not appear to pose any threat to the comparability of the two data series receive very little discussion, although in some cases it would be interesting to know more about them. Topics that appear to be a potential threat to comparability over the target period receive more attention.

The chapter closes with a summary of subject area coverage by year and assessment type (including bridge studies to adjust for changes in methodology within series) and a summary of potential methodological threats to trend comparability.

Population Definitions

During the early NAEP years, the NAEP focused on age-defined populations and attempted to address four age groups: age 9, age 13, age 17, and young adults ages 26–35. An out-of-school survey component addressed young adults through an area household sampling frame. Seventeen-year-olds were also sampled through the area household frame and, later, through school dropout lists. The out-of-school components were very expensive. The young adult household survey was discontinued after the 1973–1974 assessment, and the follow-up of dropouts was discontinued after the 1975–1976 assessment. The in-school components continue to the present time as the Long-Term Trend assessment (Chromy, Finkner, and Horvitz 2004).

The Long-Term Trend age definitions are based on date of birth and are coordinated with the timing of the assessment for each age group. Therefore, most students assessed during the scheduled period will be in the age group specified, but some may be younger or older depending on their exact birth date. Ages 9 and 13 are defined in terms of a calendar year, such that participants are age 9 or age 13 on January 1 of the assessment year. For example, for the 2008 Long-Term Trend assessment, students were eligible for the age 9 sample if they were born during calendar year 1998; students were eligible for the age 13 sample if they were born during calendar year 1994. Age 17 is defined by being age 17 on October 1. For the 2008 Long-Term Trend assessment, students were eligible for the age 17 sample if

they were born during the period October 1, 1990 through September 30, 1991 (NCES 2009a, p. 50).[3]

Age-by-grade sampling was introduced in 1983 (Hansen et al. 1987), but attempts to measure age and grade simultaneously were dropped in 1990.

Timing of Assessment during the School Year

Long-Term Trend is conducted at different times in the school year for the three age groups, with the period of field work being approximately 2 months for each age group. Age 13 students are assessed in the fall soon after school starts, age 9 students are assessed during January and February, and age 17 students are assessed from about mid-March through early May.

Given the age definitions, the age 13 sample tends to be a little younger than age 13 at the time of the assessment, the age 9 sample is closest to being age 9, and the age 17 sample is about one-half year younger on average than age 17.

Main NAEP is conducted during the period of January to March, with all three grades assessed concurrently.

Data Collection Methods

For the purpose of producing national statistics, both series are based on tests administered by specially trained contractor personnel. The early NAEP had group-administered and individually administered packages along with pacing with tape-recorded instructions and question reading. This methodology, which was continued by Long-Term Trend until 2004, was used in an attempt to separate reading ability from other cognitive traits being measured. For reading assessments, students were directed to read certain passages on their own and then the questions were read to them. For other subjects, the entire exercise was read to them. Tape-based pacing was also believed to give students an opportunity to address all the items in a booklet because students were instructed to proceed to the next item when the tape moved forward even if they had not answered the preceding item. The use of tape-recorded questions restricted the number of test versions or packages that could be administered in a single session because all the students in the room had to listen to the same recording.

In contrast, Main NAEP, since its inception in 1990, has required students to read their own booklets and allowed them to proceed at their own pace. This administration mode also allowed each student to work on a different booklet, which added statistical efficiency due to the reduced effect of sample clustering. In 2004, the Long-Term Trend reading and mathematics assessments were conducted in a bridge sample and a modified study sample. The bridge sample used the old

[3] The reference date for age 17 was redefined in the 1988 Long-Term Trend so that 17-year-olds would be 4 years older than 13-year-olds and the modal grade would be 12 rather than 11, as it was by the definition used in 1986 and earlier years (Rust 2004, p. 436).

methodology, and the modified study sample used a nonpaced methodology similar to that employed in Main NAEP (Vanneman et al. 2009, p. 50).

Accommodation and Inclusion

Before 1996, the NAEP did not allow accommodations for students with disabilities (SD) or English language learners (ELL). In 1996, a three-level split-sample experiment was conducted in conjunction with Main NAEP to test the effects of a change in exclusion criteria and the additional effects of providing accommodations in testing. Changing exclusion criteria alone did not have an appreciable effect on realized exclusion rates, so the experiment was repeated in 1998 and 2000 with two levels: (a) no change from prior years and (b) a combination of revised inclusion criteria and accommodations. Reading and mathematics were each assessed in 2 years, using the split-sample approaches (Rust 2004, p.447). In 2003 and later assessments, the NAEP has continued to allow accommodations under specified conditions. The details of the inclusion criteria were updated in 2006.

Historically, Long-Term Trend also excluded SD and ELL students. In 2004, a split-sample experiment was conducted as part of the mathematics and reading assessments to introduce accommodations (and other procedural changes) while providing a bridge with past assessments and establishing a new level for future assessments. Exclusion rates on the mathematics assessment dropped from 8% (using the old procedures) to 3% (using the new procedures) for age 9 and from 9% to 3% for age 13; similarly, exclusion rates on the reading assessment dropped from 9% to 6% for age 9 and from 9% to 5% for age 13 (Vanneman et al. 2009, p. 57). The drop in exclusions was more limited for reading than for mathematics because the NAEP does not allow the accommodation of reading the reading items to the student, which some states do allow on their own state assessments. Given the manner in which the reading construct is defined by the NAEP, this accommodation is not considered valid by the NAEP.

Private School Coverage

The Long-Term Trend assessment is designed to cover both public and private schools at the national level. The Main NAEP assessment covers both public and private schools, but the sample size for private schools is based on obtaining adequate precision for national estimates. State estimates are restricted to the population of public school students.

Assessment Instrument Frameworks

The impact of test differences and content is not directly addressed in this report. As noted in the introduction, this issue may be the topic of a subsequent study.

It is worth noting, however, that the National Assessment Governing Board (NAGB) updated the framework for Main NAEP in 2002 for reading and in 2005 for mathematics. The reading content objectives did not change, allowing a comparison of more recent estimates with those from earlier years (NCES 2009b). The updates to the mathematics assessment framework also did not change the design or content of the assessment sufficiently to interfere with comparisons across years at grades 4 and 8. However, the grade 12 changes were such that results for

grade 12 mathematics for 2005 forward cannot be compared with those of earlier (NCES 2009c).[4] Grade 12 mathematics data for 2005 and 2009 are excluded from consideration in this report.

The initial concept of Long-Term Trend was to have a large item pool and to release reports on only a small sample of items at each round of subject matter assessment. The balance of the items would be retained for comparison and release in future rounds. This approach provided an optimal way to measure trend by comparing student performance on exactly the same items at any two assessments. By the 1990s, Long-Term Trend had also adopted a scale-score approach to summarizing data across items.

The frameworks for Long-Term Trend continue to emphasize comparability for trend measurement and remain relatively unchanged (NCES 2009d). Changes in methodology were introduced in 2004 and required a bridge study. Perie and colleagues (2005, p.70) summarize some of the changes. The "I don't know" response option was dropped. Audio-paced tape presentation was discontinued in favor of self-paced assessment through each section. Accommodations for ELL/SD students were permitted.

Reporting by Performance Levels

The two assessments use different cut points and different terminology to describe performance levels. Main NAEP starts with verbal descriptors for achievement levels: basic, proficient, and advanced. Minimum scores associated with each level are then specified for each grade. For example, performance at the basic level in mathematics requires a score of at least 214 at grade 4 and at least 262 at grade 8 (NAEP 2010, pp. 18, 34).

Performance levels for Long-Term Trend start with score levels of 150, 200, 250, 300, and 350. Verbal descriptors are then developed for each scoring level. The same minimum scores are used to define performance level for all three age groups.

Because the conception and construction of performance levels are so different for the two assessment series, this report does not attempt to compare trends on the basis of performance levels.

Reporting by Percentile Scores

Both assessments also report trends in terms of percentile scores for the 10th, 25th, 50th, 75th, and 90th percentiles. This report is restricted to a study of *average* scores, but an examination of trends by percentiles would also be feasible. Visual examination of the trend plots for percentile scores and comparisons with plots of average scores seem to confirm the adequacy of limiting this comparative study to average scores.

[4] The framework for grade 12 mathematics was changed a second time in 2009, in consideration of the new emphasis on college and career readiness.

Coordination with State NAEP Assessments

Trial state assessments using the Main NAEP instruments began in 1990 in public schools at grades 4 and 8, but these assessments initially did not include all states. Main NAEP remained focused on national estimates by grade with separate samples used for state estimates. In 2002, the administration of the state NAEP assessments was shifted from local staff to contractor staff, making the administration of the state assessments more comparable to the administration of the national assessment. Accordingly, the state and national data for public schools could now be combined, providing a much larger sample for national estimates (Rust 2004, p. 448). Private schools continued to be sampled for national estimates only. In 2003 and later years, all states participated in state NAEP assessments, making the coordinated sampling for state and national estimates much more practical.

Because grade 12 was not included in the state assessments, the grade 12 sample size for Main NAEP is adequate only to support national estimates.

The Long-Term Trend assessment is designed for national estimates only and is administered by contractor personnel.

Scale Scores

Both Main NAEP and Long-Term Trend use 500-point scales for reading and mathematics. Because each student completes only a sample of the items, the student-level scores are plausible values based on a multiple imputation methodology and can be used only to describe the distribution of scores for groups of students, not for individual students.

To compare trends more easily from the two assessments in this report, a transformation of the scale scores for Main NAEP was performed so that both assessments had the same starting values in the initial years of the subject area series (i.e., the values in 1990 for mathematics and in 1992 for reading). Appendix A shows details.

Reporting Race/Ethnicity

Main NAEP collects both school-reported and self-reported measures of students' race/ethnicity. Prior to 2002, students' self-reported race/ethnicity was used in official reporting. In 2002, a decision was made to switch to school-reported race/ethnicity for reporting purposes. This decision was, in part, because of recent changes in Office of Management and Budget (OMB) requirements for the format of race/ethnicity survey questions and, in part, because of the lack of fit between subgroup totals based on student self-reports and population estimates based on the Current Population Survey (CPS). Subsequently, subgroup data for earlier years of Main NAEP were recalculated by using the school-reported race/ethnicity variable.

All the Main NAEP results for race/ethnic groups presented in this report derive from the school-based definition.[5]

In 2004, race/ethnicity data collection and reporting procedures for Long-Term Trend were revised to be consistent with those of Main NAEP. It was not, however, possible to adjust earlier years of the Long-Term Trend data using the new race/ethnic definitions, because the necessary data had not been collected. (Earlier years of Long-Term Trend depended on an observation protocol to record race/ethnicity.) The 2004 bridge study adjusts for this change in race/ethnicity reporting along with other procedural changes in Long-Term Trend.

Poststratification Issues

From the mid-1980s through 2002, the Main NAEP national estimates were poststratified to age, grade, region, and race/ethnicity estimates obtained from the CPS (Rust 2007, pp. 436–437). Poststratification for state estimates was never implemented because it was judged that independent data on population distributions at the state level were not better than the direct NAEP estimates.

Beginning in 2003, state and national samples were integrated. This process meant that (a) poststratifying the national and not the state would have led to the same student having two weights, and (b) the national samples were now so large that poststratification offered little benefit. Although the private school sample remained relatively small, no separate poststratification was undertaken for private schools separately because no suitable external demographic data are available that break out the population by the types of schools students attend.

Published estimates used in this report correspond to the nonpoststratified weights for the Main NAEP assessments in 2000 and subsequent years.

Long-Term Trend remains separate from state NAEP data, so those data potentially still could be poststratified, but poststratification was discontinued at the time of the 2004 bridge study to increase the comparability of Long-Term Trend and Main NAEP.

Thus, it appears that the two statistical series were treated in the same manner with regard to poststratification during the period covered by this report.

Response Rates

The passage of No Child Left Behind legislation provided a strong incentive for public school systems to participate in Main NAEP. Public school response rates increased for both the Main NAEP and the Long-Term Trend assessments. Increased response rates have to be viewed as an improvement in the quality of both data series, but this improvement may affect the trend if bias is reduced.

[5] The school-reported race/ethnicity variable is used to generate subgroup results in the NAEP Data Explorer; the Explorer analysis tool was used in this report.

Bridge Studies

Bridge studies for reading and mathematics were conducted in 1996, 1998, and 2000 for Main NAEP and in 2004 for Long-Term Trend. Table 3.1 shows the years and subjects included in this report along with the bridge studies used to adjust for changes in methods.

The Main NAEP bridge studies were focused on the impact of new accommodation and inclusion rules. The Long-Term Trend bridge study focused on a number of changes, including elimination of "I don't know" response options, dropping of audio-paced administration, adoption of the new accommodation and inclusion rules, and a change to no poststratification.

Table 3.1. Main NAEP assessment schedule for reading and mathematics

Year	Grade 4	Grade 8	Grade 12
1990	M	M	M
1992	R, M	R, M	M
1994	R	R	
1996	MB	MB	MB
1998	RB	RB	RB
2000	RB, MB	RB, MB	MB
2002	R	R	
2003	R, M	R, M	
2005	R, M	R, M	R, M*
2007	R, M	R, M	
2009	R, M	R, M	R†, M*

Note: R = reading assessment; M = mathematics assessment; B = bridge study.

*Mathematics assessments in 2005 and 2009 are based on new frameworks and are not comparable to prior years.

†Grade 12 reading results for 2009 were not released when this report was prepared.

Table 3.2. Long-Term Trend assessment schedule for reading and mathematics

Year	Age 9	Age 13	Age 17
1990	M	M	M
1992	R, M	R, M	R, M
1994	R, M	R, M	R, M
1996	R, M	R, M	R, M
1999	R, M	R, M	R, M
2004	RB, MB	RB, MB	RB, MB
2008	R, M	R, M	R, M

Note: R = reading assessment; M = mathematics assessment; B = bridge study.

Summary of Potential Methodological Threats to Trend Comparability

Although population definitions are quite different for the two data series, they were treated consistently within each series over the period involved. Age within grade and grade within age issues are discussed further in Chapter 4.

The timing of the data collection within the school year has different effects on the age at testing and on the portion of the year's curriculum to which the student has been exposed at time of testing. Timing of testing may not necessarily affect trend if data collection is scheduled consistently from year to year.

Data collection methods for Long-Term Trend were modified in 2004 to be the same as those used in Main NAEP. In addition to changes in frameworks, the "I don't know" response was dropped, audio-paced administration was discontinued, and accommodation and inclusion practices were made comparable to those of Main NAEP. A bridge study was conducted concurrent with this change.

Accommodation and inclusion policies changed for both series: in 1996 for Main NAEP and in 2004 for Long-Term Trend. Bridge studies were conducted for two successive assessments for Main NAEP and in a single year for Long-Term Trend.

For the purposes of producing national estimates, private school coverage was the same for both series.

The frameworks for the assessment instrument changed for both assessments during the period covered by this report. The Main NAEP reading framework changed in 2002, mathematics in 2005. Changes for grades 4 and 8 in both subjects and for grade 12 in reading were not considered substantial enough to require a bridge study. Major changes to the grade 12 mathematics framework in both 2005 and 2009, however, meant that results for those years were not considered comparable to results in earlier years. Changes to the Long-Term Trend framework were introduced in 2004 along with other procedural changes and were accompanied by a bridge study.

A change in coordination with state NAEP assessments occurred in 2003 for Main NAEP, but the change should have had no impact on national estimates.

Scale scores do not appear to be a trend comparability issue after adjustments for scale differences are made. As more data accumulate for the Main NAEP grade 12 mathematics measure, it may be possible to do a shift adjustment and scale correction to check comparability of the grade 12 mathematics trends, although the 2005 framework was used for only a single year.

Poststratification was dropped for Main NAEP when state and national samples were combined, based on newly consistent data collection methods for state and national purposes, starting in 2003. Poststratification was also dropped for Long-Term Trend, starting with the bridge study in 2004.

The increase in response rates accompanying the passage of No Child Left Behind legislation appears to have affected both series similarly.

In most instances, these potential threats to trend comparability are not considered serious enough to discourage comparisons of trends after adjusting for scale differences and accounting for bridge studies. The new assessment frameworks for the Main NAEP grade 12 mathematics introduced in 2005 and 2009 were handled by excluding 2005 and 2009 data from our analysis. Any conclusions about trend comparability under the older grade 12 mathematics framework may not extend to Main NAEP data generated under the grade 12 frameworks in use since 2005.

4. Age and Grade Effects

This chapter explores the effect of the distribution of grade samples by age in Main NAEP and the distribution of age sample by grade in Long-Term Trend. Table 4.1 shows these distributions over the period 1990–2009. The distributions do not always add to 100% because only the largest categories are shown in the tables; small subgroups (e.g., grade 4 students below age 9) are excluded because the sample is too small to provide reliable estimates. The Long-Term Trend grade distribution data for 1999 were available only for age 17.

Table 4.1. Age/grade distributions by assessment year: Main NAEP mathematics assessment for grades 4, 8, and 12 and Long-Term Trend mathematics assessment for ages 9, 13, and 17

	Percent Distribution by Assessment Year											
Age and Grade	**1990**	**1992**	**1994**	**1996**	**1999**	**2000**	**2003**	**2004**	**2005**	**2007**	**2008**	**2009**
Grade 4												
Age 9	60	57		62		63	61		61	61		61
Above age 9	40	42		37		37	38		39	39		39
Age 9												
Below grade 4	35	38	33	33	–			36			39	
Grade 4	65	62	66	66	–			64			61	
Grade 8												
Age 13	59	57		58		60	61		60	60		60
Above age 13	40	42		41		39	38		39	39		40
Age 13												
Below grade 8	36	37	38	36	–			38			40	
Grade 8	63	62	62	63	–			62			60	
Grade 12												
Age 17	66	67		64		64						
Above age 17	32	32		35		36						
Age 17												
Below grade 11	22	24	21	24	23			24			25	
Grade 11	70	70	73	71	74			72			71	
Above grade 11	8	6	6	6	3			4			4	

Note: In years with bridge studies, only the estimates based on the accommodated sample are shown. Modal grade data for the 1999 Long-Term Trend were available for age 17 only.

Table 4.1 shows that the modal ages for Main NAEP are 9, 13, and 17. Some year-to-year variation occurs in the proportion in the modal age, but no overall trend appears in the data. Table 4.1 also shows that the modal grades for Long-Term Trend are grades 4, 8, and 11. The variation in year-to-year proportions in modal grade also does not appear to follow any consistent trend pattern.

Tables 4.2 and 4.3 compare average scores by age and grade from both assessments in a year (1992) when both were conducted. Table 4.2 shows that students above the modal age in Main NAEP tend to get lower scores than those at the modal age. Note that all three grades are assessed during the same period (January through March).

No data have been examined to support this argument, but a higher proportion of students above the modal age may have been held back because of academic difficulty, language problems, or other reasons.

Table 4.2. Main NAEP scores in mathematics and reading by grade and age, 1992

Grade	When Assessed	Age	Birth Dates	Percent of Total	Average Score	
					Math	Reading
4	January to March	Age 9	1982	57	240	215
		Above age 9	1981 or earlier	42	232	204
8	January to March	Age 13	1978	57	282	269
		Above age 13	1977 or earlier	42	266	248
12	January to March	Age 17	October 1974–September 1975	67	314	302
		Above age 17	Before October 1974	32	299	284

Note: Scores are transformed to have levels and population standard deviations equivalent to the Long-Term Trend data in mathematics for 1990 and in reading for 1992.

The Long-Term Trend age definitions classify student ages by birth date (a cohort definition) rather than exact age at the time of assessment. As discussed in Chapter 3, ages 9 and 13 are based on birth year or age as of January 1 of the assessment year. Age 17 is defined for the Long-Term Trend assessment that is conducted in April and May, and "age 17" is based on age as of October 1 of the calendar year. As a result, 17-year-old students tend to be in an earlier academic year (their grade 11 year) at the time of assessment. Table 4.3 shows that students below the modal grade in Long-Term Trend have considerably lower scores than those in the modal grade. Only data for 1992 are shown; however, this relationship holds for the entire data series. The lower scores achieved by students below the modal grade may result either from lack of exposure to the topics of the assessment items, which are designed for the modal grade, or from being held back because of academic difficulty or language problems.

Table 4.3. Long-Term Trend scores in mathematics and reading by age and grade, 1992

Age	When Assessed	Birth Dates	Grade	Percent of Total	Average Score	
					Math	Reading
9	January and February, 1992	1982	Below grade 4	38	207	192
			Grade 4	62	242	224
13	Fall, after school starts,1991	1978	Below grade 8	37	258	243
			Grade 8	62	282	272
17	Mid-March through early May, 1992	October 1974–September 1975	Below grade 11	24	285	261
			Grade 11	70	313	301
			Above grade 11	6	318	300

Because this report focuses on comparing trends, no attempt was made to try to further understand the reasons for the score differences by grade within age or by age within grade. Further study would be of general interest in guiding future alternatives.

Table 4.4 shows the trend data for mathematics at the modal grade 4 and modal age 9. In Chapter 2, Table 2.1, the data were transformed to have the same starting point in 1990 and the same population standard deviation. When applying the same linear transformation to the modal age and grade, the starting points do not match, but these points are sufficiently close to give some perception of the difference in average annual change. The Main NAEP scores start out lower in 1990 and 1992, but the scores increase to values higher than the Long-Term Trend scores in the final years shown. The difference in trend lines at the elementary level was confirmed by testing the differences of the trend lines, using the same model applied in Chapter 2. This model fit individual linear trend lines and allowed a separate shift variable to account for introduction of accommodations in testing and broader inclusion policies and other methodological changes when those changes were accompanied by bridge studies.

Table 4.4. Average mathematics scores by assessment year for students who are both grade 4 and age 9: Main NAEP and Long-Term Trend

	Accom-modations	Assessment Year											
	Permitted	1990	1992	1994	1996	1999	2000	2003	2004	2005	2007	2008	2009
Main NAEP Transformed	Yes				241		244	254		257	259		259
	No	234	240		241		246						
Long-Term Trend	Yes								250			253	
	No	242	242	241	241				251				

Note: The transformation parameters that were applied in Chapter 2 to set the 1990 levels and standard deviations for the Main NAEP full sample data series equivalent to those for the Long-Term Trend data series were applied to the Main NAEP modal grade data series in this chapter.

Table 4.5 shows the result of the model-fitting process applied to all six combinations of level and subject. Only the elementary mathematics slope lines showed significant differences between the Main NAEP and Long-Term Trend average annual change (slope). Both elementary and middle school mathematics showed positive trends similar to the results in Chapter 2.

No slope differences were found significant for reading. Elementary reading showed significant positive trends in the all-student analysis for both Main NAEP and Long-Term Trend in Chapter 2. Here, only the Main NAEP slope remains significantly positive for the modal group. At both the middle and high school levels, Long-Term Trend shows negative annual change when restricted to the modal grade.

High school reading showed no significant trend in the analysis in Chapter 2, but reading shows a statistically significant negative trend for the Long-Term Trend data for 17-year-old students in grade 11. The Main NAEP grade 12 trend data are limited to the period 1990 through 2000, thus making it more difficult to detect a long-term trend in average scores. Note that at the high school level, the Long-Term Trend modal group was defined as age 17 and grade 11 because of small sample sizes for age 17 and grade 12. This group was not strictly comparable to the grade 12 and age 17 modal group defined for Main NAEP.

Table 4.5. Estimated linear model parameters by subject and level: In elementary school, age 9 and grade 4; middle school, age 13 and grade 8; high school, age 17 and grade 11 (Long-Term Trend) or grade 12 (Main NAEP)

School Level and Subject	Type	Intercept (Year = 2000)	Shift for Accommodation *	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R2
Elementary Math	Main	248.05	–0.67	1.42	<.0001			
	LTT	246.78	0.52	0.70	0.0109	–0.72	0.0342	0.92
Middle Math	Main	285.60	–1.72	0.74	<.0001			
	LTT	284.47	–0.47	0.42	0.0091	–0.32	0.0917	0.89
High Math	Main	315.88	–0.22	0.44	0.1182			
	LTT	312.42	–1.71	0.04	0.8348	–0.40	0.2264	0.48
Elementary Reading	Main	215.64	–3.48	0.53	0.0279			
	LTT	224.94	–1.39	0.42	0.1231	–0.11	0.7362	0.89
Middle Reading	Main	268.59	–1.41	–0.01	0.9450			
	LTT	266.27	0.73	–0.42	0.0472	–0.41	0.1461	0.58
High Reading	Main	266.74	0.00	0.00	1.0000			
	LTT	294.70	0.75	–0.67	0.0137	–0.67	0.1111	0.99

Note: Main = Main NAEP; LTT = Long-Term Trend.

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the six models.

Restricting the sample to modal age and modal grade had negligible effects on the comparability of trend measures when averaged over the period. Of the two statistically significant slope differences for the complete data models studied in Chapter 2, the difference remained statistically significant for elementary mathematics but not for middle school mathematics. In both cases, the Main NAEP annual change estimates were larger.

5. Performance by Racial and Ethnic Groups

As mentioned earlier, the performance of various racial and ethnic groups is of great importance to educational policymakers and the general public. This chapter presents the trends in average reading and mathematics performance for White, Black, and Hispanic students because these are the only racial/ethnic groups for which sufficient data are available.

An important question is whether or not the trends differ substantially for the Main NAEP and Long-Term Trend samples. To explore this question, six graphs, one for each of the three education levels in each subject area are presented. In turn, each graph contains six trend lines: the Main NAEP and Long-Term Trend trends for each of the three racial/ethnic groups. After these figures, significance tests are shown, highlighting the differences between the Main NAEP and Long-Term Trend trends. Because age and grade sampling may explain the differences, the significance tests are repeated for only those students who are in the modal age groups for Main NAEP and the modal grade groups for Long-Term Trend.

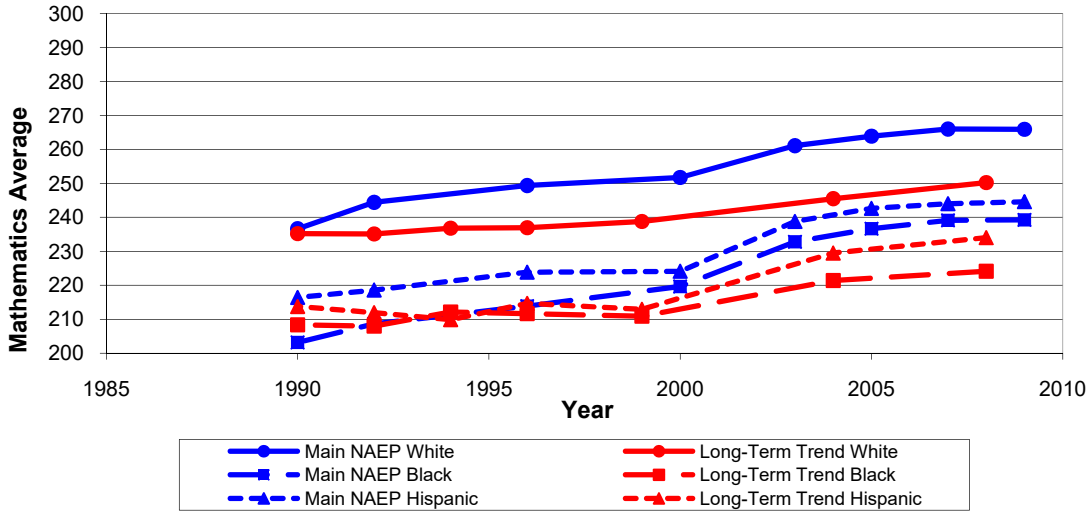
In all the figures below, blue symbols represent the Main NAEP averages and red symbols represent the Long-Term Trend averages. The White student sample is represented by a circle, the Black by a square, and the Hispanic by a triangle. The assessment averages are connected by solid, dashed, or dotted lines to help the reader follow the changes in average performance over the assessed years.

The average values for the Main NAEP samples have been transformed so that the national means and standard deviations are the same as those of the Long-Term Trend samples on the first year in which both samples were assessed. For reading, the first year was 1992; for mathematics, the first year was 1990. However, the racial/ethnic group averages are not individually constrained to be the same in the first year.

For simplicity, only the accommodated sample was used in "bridge" years in which separate samples were drawn to study the effect of allowing accommodations.

Figure 5.1 shows the average performance of elementary school students in mathematics.

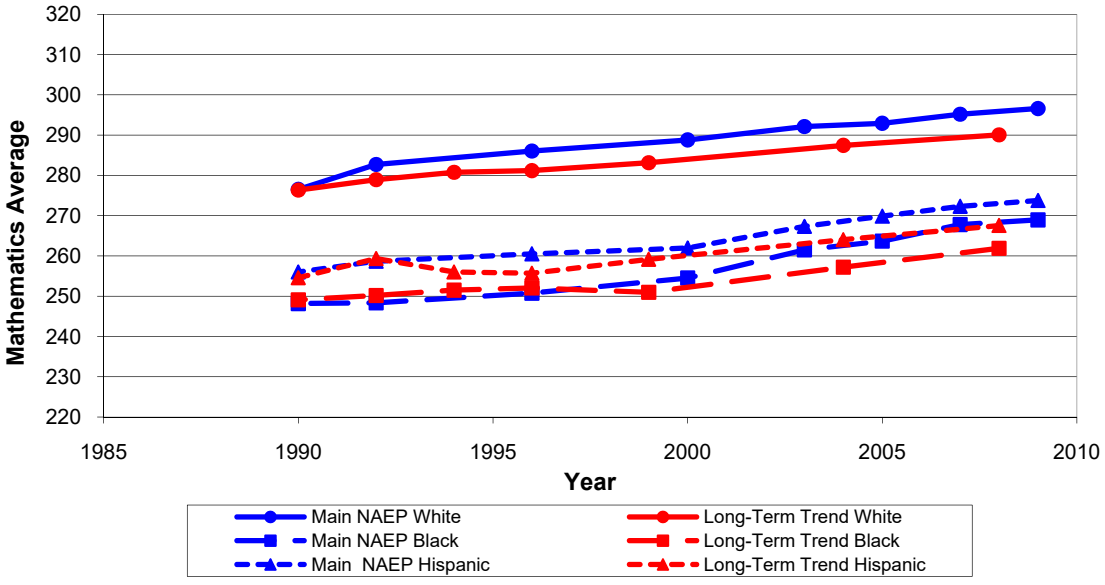
Figure 5.1. Average mathematics scores by assessment year and racial/ethnic group: Main NAEP grade 4 transformed scores and Long-Term Trend scores for age 9



In mathematics at the elementary level, the trends for all groups improved over time, with the Main NAEP trends improving more than those of the Long-Term Trend assessments. The Main NAEP samples also showed a larger increase than the Long-Term Trend samples for all racial/ethnic groupings.

Figure 5.2 shows that at the middle school level, each racial/ethnic group shows a gain and the Main NAEP assessment indicates larger gains than the Long-Term Trend assessment.

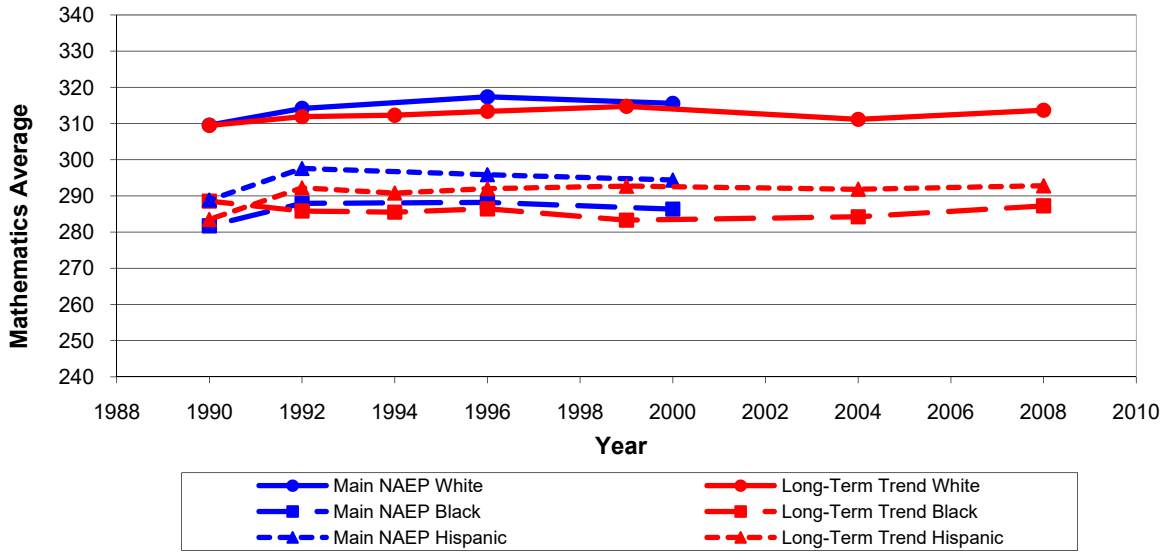
Figure 5.2. Average mathematics scores by assessment year and racial/ethnic group: Main NAEP grade 8 transformed scores and Long-Term Trend scores for age 13



For the high school samples, the Main NAEP trend line in mathematics ends in the year 2000 because later assessments used a different test framework that is deemed

not comparable. In the span of years where both assessments were administered, as shown in Figure 5.3, students in the Main NAEP grade 12 sample did somewhat better than those in the Long-Term Trend age 17 sample.

Figure 5.3. Average mathematics scores by assessment year and racial/ethnic group: Main NAEP grade 12 transformed scores and Long-Term Trend scores for age 17



As shown in Figure 5.4, the reading trends for elementary-level Main NAEP and Long-Term Trend are almost identical for Whites. A large gap occurs between the performance of White students and both Black students and Hispanic students. Both the Black students and the Hispanic students averaged a little higher in the Long-Term Trend assessments. Both the Main NAEP and Long-Term Trend lines show a slight increase from the 1992 values for White students and a somewhat larger gain for Black students and Hispanic students.

Figure 5.4. Average reading scores by assessment year and racial/ethnic group: Main NAEP grade 4 transformed scores and Long-Term Trend scores for age 9

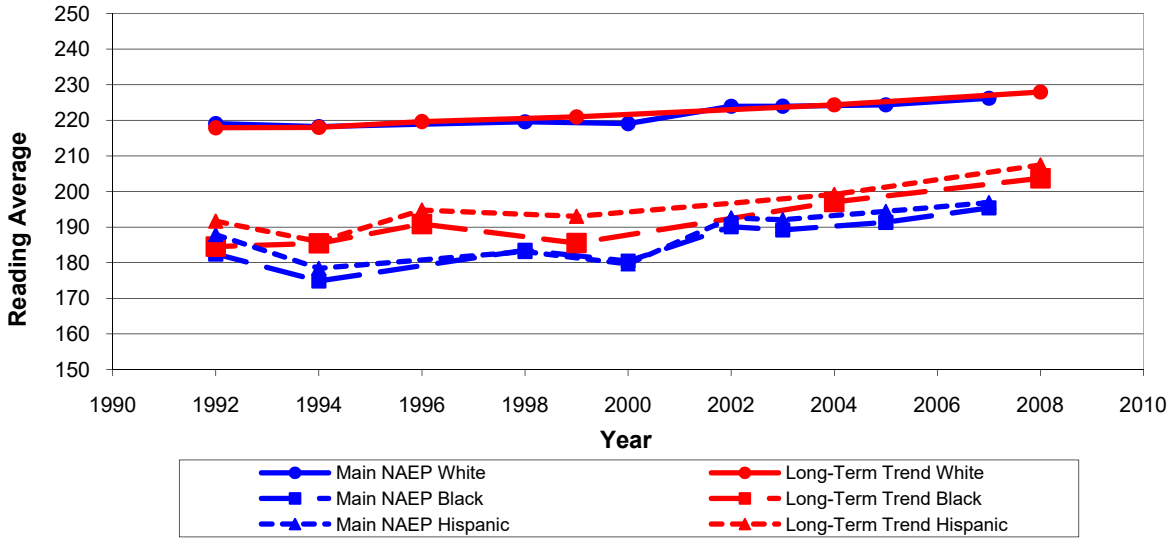
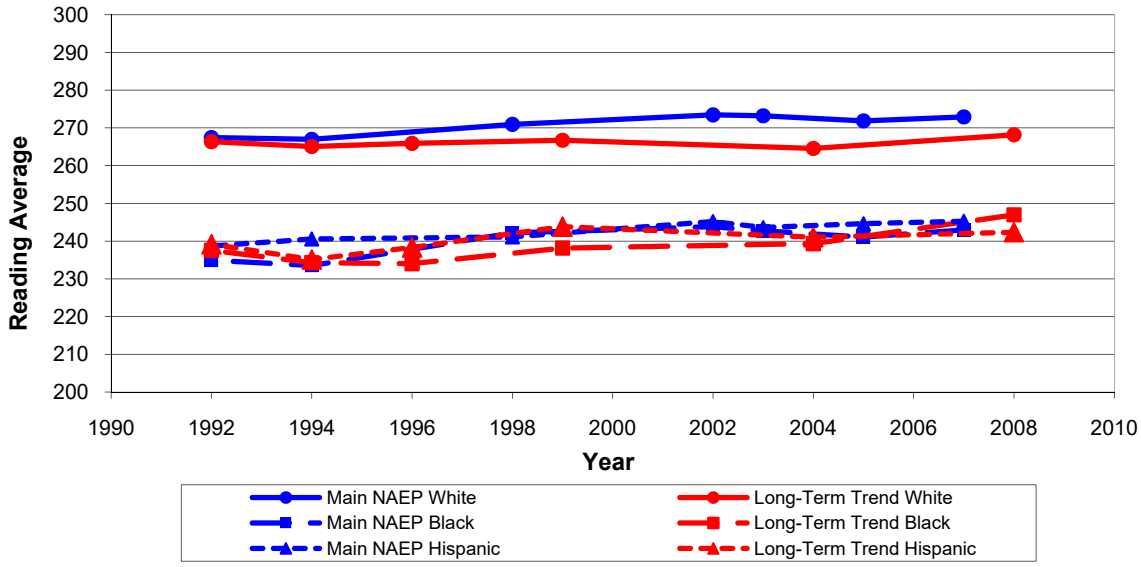


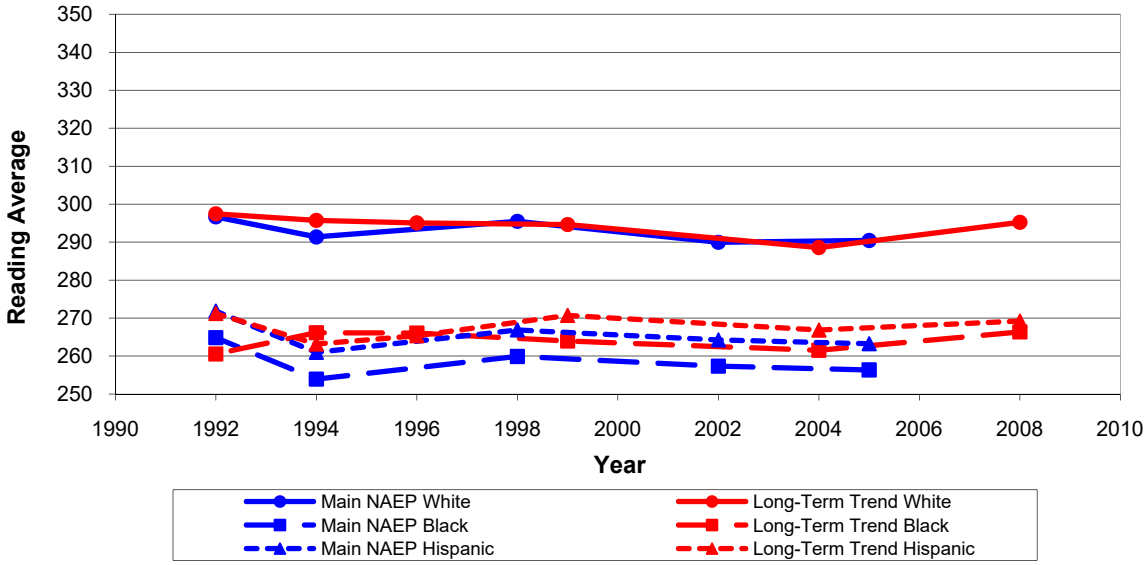
Figure 5.5 shows the reading trend lines for the Main NAEP at grade 8 and Long-Term Trend assessments at age 13. The results are similar to those observed for the earlier age and grade grouping, with students performing a little better on the Main NAEP assessments. Both the Main NAEP and Long-Term Trend trends show slight improvements over time.

Figure 5.5. Average reading scores by assessment year and racial/ethnic group: Main NAEP grade 8 transformed scores and Long-Term Trend scores for age 13



In Figure 5.6, for the high school samples, the performance gap between race/ethnic groups is still evident. At this grade level, little, if any, improvement over time seems to occur. Black and Hispanic students both do a little better in the Long-Term Trend assessment than in Main NAEP.

Figure 5.6. Average reading scores by assessment year and racial/ethnic group: Main NAEP grade 12 transformed scores and Long-Term Trend scores for age 17



Are the differences between trend lines statistically significant? To address this question, linear trend models were fit to the Main NAEP and Long-Term Trend averages to compare the two trends for the different racial/ethnic groupings separately. The models are the same as those in Chapter 2.

Tables 5.1 and 5.2 show the statistics for mathematics and reading, respectively. It is important to remember that at each educational level, the overall averages and standard deviations were set to be the same for the Main NAEP and Long-Term Trend samples at the beginning of the reading and mathematics trends. The differences between the Main NAEP and Long-Term Trend trends are encapsulated in their slopes. A significance test is given for the differences between these slopes.

The results in these tables show little difference between the trends of Main NAEP and Long-Term Trend. In mathematics, only two slopes are significantly different: those for White students and Black students at the elementary school level. None of the reading slopes is significantly different at the 0.05 level for any racial/ethnic group or at any educational level. In summary, only 2 of the 18 pairs of slopes were significantly different.

Table 5.1. Estimated linear model parameters for mathematics: By level, race/ethnic group, and type of assessment

Level and Group	Type	Intercept (Year = 2000)	Shift for Accom-modation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary White	Main	253.65	1.09	1.47	<0.0001			
	LTT	241.91	0.79	0.86	0.0004	–0.61	0.0229	0.97
Elementary Black	Main	223.02	0.45	2.02	<0.0001			
	LTT	216.82	–0.17	1.02	0.0013	–1.00	0.0099	0.95
Elementary Hispanic	Main	229.38	2.05	1.57	0.0003			
	LTT	220.07	5.05	1.10	0.0087	–0.47	0.3432	0.91
Middle School White	Main	289.09	–0.34	0.96	<0.0001			
	LTT	284.73	–0.69	0.79	<0.0001	–0.17	0.2384	0.97
Middle School Black	Main	285.33	–1.39	1.28	<0.0001			
	LTT	255.96	–1.26	0.81	0.0007	–0.47	0.0733	0.93
Middle School Hispanic	Main	264.43	0.11	0.92	<0.0001			
	LTT	260.96	1.02	0.63	0.0024	–0.29	0.2150	0.92
High School White	Main	317.76	–0.16	0.54	0.0777			
	LTT	313.76	–3.1	0.29	0.1547	–0.25	0.4711	0.62
High School Black	Main	289.59	–1.48	0.42	0.2509			
	LTT	285.15	1.55	–0.16	0.5166	–0.58	0.1982	0.27
High School Hispanic	Main	296.66	–0.89	0.32	0.4740			
	LTT	291.11	–0.27	0.24	0.4424	–0.08	0.8828	0.46

Note: Main = Main NAEP; LTT = Long-Term Trend.

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the nine models.

Table 5.2. Estimated linear model parameters for reading: By level, race/ethnic group, and type of assessment

Level and Group	Type	Intercept (Year = 2000)	Shift for Accom-modation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary White	Main	221.84	–0.44	0.55	0.0002			
	LTT	222.79	–0.68	0.67	<0.0001	0.13	0.4058	0.91
Elementary Black	Main	185.09	0.69	1.09	0.0036			
	LTT	193.02	0.27	1.18	0.0017	0.09	0.8359	0.85
Elementary Hispanic	Main	189.74	–2.95	1.20	0.0067			
	LTT	198.12	–1.65	1.14	0.0073	–0.05	0.9245	0.78
Middle School White	Main	270.37	1.11	0.31	0.0678			
	LTT	266.34	–1.34	0.23	0.0960	–0.08	0.7019	0.89
Middle School Black	Main	238.22	3.38	0.33	0.3433			
	LTT	240.01	0.71	0.41	0.1742	0.07	0.8700	0.57
Middle School Hispanic	Main	243.54	–1.09	0.54	0.0096			
	LTT	241.26	–2.019	0.44	0.0109	–0.10	0.6609	0.82
High School White	Main	292.42	0.23	–0.41	0.3424			
	LTT	294.58	–1.53	–0.19	0.4273	0.22	0.6513	0.44
High School Black	Main	258.06	0.44	–0.38	0.5079			
	LTT	264.58	–0.55	–0.02	0.9601	0.36	0.5788	0.53
High School Hispanic	Main	265.21	0.19	–0.37	0.5864			
	LTT	266.1	4.86	-0.49	0.2218	-0.12	0.8800	0.30

Note: Main = Main NAEP; LTT = Long-Term Trend.

* The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the nine models.

The graphic and statistical results support the notion that the trends are nearly the same. The gap between majority and minority students is clearly established in both mathematics and reading at all educational levels. The next question is: Why?

An obvious and quite possible explanation of the differences between the Main NAEP and Long-Term Trend means is the difference between grade and age populations. To examine this effect, those students who were *not* in both the modal grade and age of the samples were removed from the analysis. For example, students who were not 9-years-old were removed from the grade 4 Main NAEP sample and students who were not in grade 4 were removed from the Long-Term Trend 9-year-old sample.

The same statistical tests were performed for the modal age/grade samples as for the full student samples. Tables 5.3 and 5.4 show the results for mathematics and reading, respectively.

Table 5.3. Estimated linear model parameters for modal group: Mathematics, by level, race/ethnic group, and type of assessment

Level and Group	Type	Intercept (Year = 2000)	Shift for Accom-modation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary White	Main	255.16	0.74	1.42	<0.0001			
	LTT	253.37	–0.35	0.81	0.0015	–0.61	0.0350	0.95
Elementary Black	Main	226.56	–0.21	2.03	<0.0001			
	LTT	228.69	0.46	0.94	0.0121	–1.09	0.0223	0.93
Elementary Hispanic	Main	231.13	0.43	1.54	0.0016			
	LTT	231.93	3.03	1.07	0.0311	–0.46	0.4345	0.84
Middle School White	Main	292.54	–0.56	0.80	<0.0001			
	LTT	291.03	0.17	0.59	0.0002	–0.21	0.1722	0.95
Middle School Black	Main	263.95	–1.29	1.09	0.0004			
	LTT	265.95	–1.31	0.64	0.0288	–0.45	0.2117	0.84
Middle School Hispanic	Main	267.99	–0.43	0.81	0.0003			
	LTT	267.99	1.54	0.30	0.1806	–0.51	0.0968	0.82
High School White	Main	320.23	–0.10	0.16	0.5179			
	LTT	317.49	–2.43	0.17	0.2082	0.02	0.9536	0.71
High School Black	Main	293.66	–1.03	–0.13	0.6875			
	LTT	291.82	1.91	–0.21	0.2482	–0.08	0.8246	0.31
High School Hispanic	Main	298.15	0.06	–1.04	0.0594			
	LTT	2978.00	0.10	0.11	0.6733	1.15	0.0650	0.61

Note: Main = Main NAEP; LTT = Long-Term Trend.

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the nine models.

Table 5.4. Estimated linear model parameters for modal group: Reading by level, race/ethnic group, and type of assessment

Level and Group	Type	Intercept (Year = 2000)	Shift for Accom-modation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary White	Main	224.54	–1.22	0.54	<.0001			
	LTT	232.56	–0.24	0.31	0.1140	–0.23	0.3778	0.88
Elementary Black	Main	191.13	–0.29	0.94	0.0064			
	LTT	206.28	–0.43	1.00	0.0049	0.06	0.8836	0.91
Elementary Hispanic	Main	191.34	–2.99	1.14	<.0001			
	LTT	206.75	–2.17	0.94	0.0685	–0.19	0.7733	0.8
Middle School White	Main	276.01	–0.71	0.17	0.4238			
	LTT	272.93	–0.88	0.00	0.9875	–0.16	0.5503	0.52
Middle School Black	Main	247.09	1.09	0.20	0.5535			
	LTT	250.77	1.61	–0.06	0.8259	–0.27	0.5621	0.39
Middle School Hispanic	Main	250.06	–3.15	0.32	0.1924			
	LTT	247.07	2.20	–0.27	0.2673	–0.60	0.0811	0.28
High School White	Main	297.40	–0.23	–0.55	0.2096			
	LTT	300.45	–0.55	–0.42	0.0973	0.12	0.7953	0.69
High School Black	Main	265.56	0.55	–0.78	0.2445			
	LTT	275.62	–1.64	–0.28	0.4375	0.49	0.5073	0.73
High School Hispanic	Main	270.77	2.55	–1.19	0.0297			
	LTT	275.42	5.84	–0.86	0.0099	0.34	0.5335	0.8

Note: Main = Main NAEP; LTT = Long-Term Trend.

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the nine models.

The trends for the modal groups are quite similar to those in the full sample and do not seem to change the interpretation of results. In reading, the Main NAEP and Long-Term Trend slopes never differ significantly at the .05 level. In mathematics, the only significant results are again for the slopes of White students and Black students.

The similarities between the trend lines lead to a question of why both trends are necessary. That is, what policy questions can one trend address that the other cannot?

6. The Effect of Population Shifts

Chapter 2 presented the NAEP trends in mathematics and reading. Such trend lines represent changes in the populations being assessed as well as changes in their performances. In fact, it is possible for all subpopulations to raise or lower their average performances while the overall average shows a change in the opposite direction. This chapter presents "demographically standardized means" to show how the NAEP trend lines would look if subpopulations in the trend samples had remained the same size in different assessment years.

The figures in Chapter 5 show separate NAEP trend lines for several subpopulations based on race/ethnicity (White, Black, and Hispanic). A comparison of these subpopulations' trend lines with the overall average trend lines shown in Chapter 2 raises some concerns. Table 6.1 presents the reasons for these concerns and shows the average mean Main NAEP scores for different subgroups in different subject areas, grade levels, and assessment years. The first column shows the assessment years 1990 and 2009, the first and last years in the trends studied. The following columns give the means for the three largest racial/ethnic subgroups and for an "Other" group that combines Asian students, Native American students, and students from other small subpopulations. The final two columns show the overall published averages for these assessments and the demographically standardized means discussed below.

Table 6.1. Main NAEP trend: Gain/loss in mean performance by racial/ethnic subgroup and overall

Average Mathematics Score in Grade 4					Overall	
Year	White	Black	Hispanic	Other	Published Mean	Demographically Standardized
1990	219.8	187.5	200.3	220.2	213.1	213.1
2009	248.1	222.3	227.5	247.8	239.7	242.4
Difference	28.3	34.8	27.2	27.6	26.6	29.3
Average Mathematics Score in Grade 8					**Overall**	
Year	White	Black	Hispanic	Other	Published Mean	Demographically Standardized
1990	269.6	236.8	245.9	264.5	262.6	262.6
2009	292.9	260.9	266.4	293.2	282.9	285.9
Difference	23.3	24.1	20.6	28.7	20.4	23.4
Average Reading Score in Grade 4					**Overall**	
Year	White	Black	Hispanic	Other	Published Mean	Demographically Standardized
1992	224.3	192.0	196.8	214.4	216.7	216.7
2009	230.3	204.5	205.1	228.6	220.9	224.3
Difference	6.0	12.5	8.3	14.2	4.2	7.5
Average Reading Score in Grade 8					**Overall**	
Year	White	Black	Hispanic	Other	Published Mean	Demographically Standardized
1992-R2	267.0	237.4	240.8	263.4	260.0	260.0
2009	272.9	246.4	249.1	269.7	264.0	266.6
Difference	5.9	9.0	8.3	6.3	4.0	6.6

The bottom rows in the four grade and subject groupings of Table 6.1 display the differences (gain or loss) in mean performance over the trend years and show reason for concern about population shifts. The good news is that in all trend lines, all subgroups gained, as did the overall published mean. However, it is curious that the gains for every racial/ethnic group are larger than the overall "published" gains. The overall published means seem to diminish the success achieved by the subgroups. Two questions arise: What is happening? What can be done about it?

Underlying this phenomenon is a shift in the relative size of the NAEP racial/ethnic subpopulations over the trend years. Table 6.2 shows the percentage of students from each racial/ethnic group in the grade 4 mathematics population. Note that the percentage of White students diminished from 75% in 1990 to 56% in 2009, while the Hispanic students increased from 6% to 21%. The Black population stayed about the same (18% and 16%), and the "Other" group increased from 2% to 8%. This change in the mix of students in different years is sufficient to explain the fact that all racial/ethnic groups gained more than the overall gain. The shifts in subpopulation sizes are similar for other NAEP trend data.

Table 6.2. Racial/ethnic distributions by year: Main NAEP mathematics assessment for grade 4

Year	White (%)	Black (%)	Hispanic (%)	Other (%)
1990-R2	75	18	6	2
1992-R2	73	17	6	3
1996-R2	72	16	8	4
1996	66	16	11	7
2000	64	16	15	6
2003	60	17	18	6
2005	58	16	19	7
2007	57	16	20	7
2009	56	16	21	8

Note: R2 designates the sample that was allowed accommodations in bridge years.

For descriptive purposes, the mathematics and reading trends for grade 4 and grade 8 students in the Main NAEP samples were analyzed. The grade 12 Main NAEP samples were not analyzed because the mathematics test frameworks were substantially changed in 2005 and 2009, thus truncating the grade 12 trend in that subject. In this analysis, the Main NAEP data were *not* transformed to increase comparability to the Long-Term Trend sample. The focus here is comparing the demographically standardized means with the overall means that are published in NAEP reports, such as the 2009 Mathematics Report Card (NCES 2010).

Before the demographically standardized means are described, it is useful to first look at the overall means that are usually published. Trend data are collected on student populations in different years. The population within each year consists of several subpopulations that may vary in size and performance. One may assume that the subpopulations are mutually exclusive and exhaustive and that their definitions do not change over time. Under these conditions, the overall mean at a particular time can be computed

$$(1)\quad x_t = \sum_k p_{tk} x_{tk}$$

where p_{tk} ($t = 1,2,\ldots,T$; $k = 1,2,\ldots,K$) is the proportion of students in subpopulation k at time t and x_{tk} is the average value of x for that group at that time.

Beaton and Chromy (2007) have shown how the difference between two means can be partitioned into components as a result of changes in performance, changes in populations, and their interactions. A side component of that work is demographically standardized means, defined here as

$$(2)\quad x_{st} = \sum_k p_{1k} x_{tk}$$

That is, the proportion in each category is kept the same, while the values of x_{1k} change over time. In this way, the demographically standardized means reflect the changes in group performance, not changes in subpopulation membership.

Any proportions may be used for demographic standardization as long as they sum to 1.00. With trends over time, it is sensible to use values of the p_{1k}, the proportions at the earliest point in the series.

Tables 6.3 through 6.6 show the results for four Main NAEP trends. Each table row contains the assessment year, the published mean score, the demographically standardized mean, and the difference between the demographically standardized and published means. In the earliest year, the results are identical; they must be, because the proportions are the same. Over the following years, as the population changes became more pronounced, the demographically standardized means became more divergent—ending roughly 3 scale points higher than the published means. Further, in Table 6.1, the demographically standardized means are always found amid the subpopulation means, not below them as the published means are.

Table 6.3. Main NAEP grade 4 mathematics scores by year: Published and demographically standardized mean scores

Math	Grade 4	Main	
Year	Published	Demographically Standard	Difference
1990	213.1	213.1	0.0
1992	219.7	219.8	0.1
1996	223.9	224.1	0.2
1996	223.5	224.6	1.1
2000	225.6	227.5	1.9
2003	234.9	237.3	2.4
2005	237.9	240.3	2.4
2007	239.7	242.4	2.7
2009	239.7	242.4	2.7

Table 6.4. Main NAEP grade 8 mathematics scores by year: Published and demographically standardized mean scores

Year	Published	Demographically Standard	Difference
1992	262.6	262.6	0.0
1994	268.4	268.5	0.1
1996	272.0	272.4	0.4
1996	270.5	271.7	1.3
2000	273.1	275.2	2.1
2003	277.6	279.9	2.4
2005	278.8	281.3	2.5
2007	281.3	284.3	2.9
2009	282.9	285.9	3.0

Table 6.5. Main NAEP grade 4 reading scores by year: Published and demographically standardized mean scores

Year	Published	Demographically Standard	Diff.
1992	216.7	216.7	0.0
1994	214.3	214.6	0.3
1998	217.3	218.1	0.8
1998	214.8	216.9	2.1
2000	213.4	216.2	2.8
2002	218.6	221.5	2.9
2003	218.2	221.3	3.1
2005	219.0	222.1	3.2
2007	221.0	224.2	3.2
2009	220.9	224.3	3.3

Table 6.6. Main NAEP grade 8 reading scores by year: Published and demographically standardized mean scores

Year	Published	Demographically Standard	Diff.
1992	260.0	260.0	0.0
1994	259.6	259.5	-0.1
1998	263.6	264.2	0.5
1998	262.9	263.6	0.6
2002	264.3	265.8	1.4
2003	263.3	265.4	2.1
2005	262.2	264.4	2.2
2007	262.8	265.4	2.6
2009	264.0	266.6	2.6

According to the NAEP data shown here, the demographically standardized means make a slightly different statement by separating the performance of racial/ethnic subgroups from population shifts.[6] In these instances, all racial/ethnic subgroups show gains, and the demographically standardized means present slightly larger trend values. Both the published and demographically standardized trend lines should be considered in evaluating NAEP trends.

It is noteworthy that the racial/ethnic groups analyzed here differ in performance as well as in their population proportions. Demographically standardizing by gender would make little difference, because the proportion of boys and girls does not

[6] For those who wish to duplicate the computations in this chapter, all basic data in this report were produced by NAEP's Data Tool. However, the confidentially concerns of the U.S. Department of Education require that percentages be presented as whole numbers, although averages may be presented to two decimal places. Rounding the percentages to integers introduced small but unacceptable errors into the demographically standardized means. Therefore, special analyses were done using more precise values. Consequently, computing the demographically standardized means from NAEP Data Tool output will not reproduce exactly the values presented in this chapter.

change substantially over the assessment years. Other groupings of students could also be used; however, the demographic standardization of trends is useful only when the subgroups are of interest to policymakers or the general public.

In addition to the comparisons of reported and demographically standardized estimates discussed above, an analysis comparing the Main NAEP and Long-Term Trend lines was developed for the demographically standardized estimates. This analysis is comparable to the analyses in Chapters 2 and 4 except that the demographically standardized scores are used as dependent variables rather than the reported scores. Table 6.7 shows results comparable to those in Tables 2.7 and 4.5.

Table 6.7. Estimated linear model parameters by subject and level: Demographically standardized data

Level and Subject	Type	Intercept (Year = 2000)	Shift for Accom-modation*	Annual Change (Slope)	P-Value for Slope	Slope Difference	P-Value for Slope Difference	Model R^2
Elementary Mathematics	Main	247.91	–0.34	1.66	<0.0001			
	LTT	236.83	0.99	0.92	0.0005	–0.74	0.0197	0.97
Middle School Mathematics	Main	282.41	–0.68	1.02	<0.0001			
	LTT	278.62	–0.62	0.79	<0.0001	–0.23	0.0882	0.98
High School Mathematics	Main	316.19	–3.35	0.93	0.0441			
	LTT	307.83	–2.23	0.22	0.2303	–0.71	0.1308	0.70
Elementary Reading	Main	214.43	–1.90	0.79	0.0003			
	LTT	216.39	–0.57	0.82	<0.0001	0.03	0.8832	0.91
Middle School Reading	Main	262.97	1.15	0.34	0.0364			
	LTT	260.09	–0.87	0.28	0.0363	–0.06	0.7642	0.88
High School Reading	Main	284.90	0.26	–0.39	0.3725			
	LTT	287.52	–0.66	–0.18	0.4569	0.21	0.6717	0.42

Note: Main = Main NAEP; LTT = Long-Term Trend.

*The models included a shift variable to indicate accommodated assessment. The shift parameter was not statistically significant for any of the six models.

Demographic standardization to an early year race/ethnic group distribution removes some of the effect of demographic shifts that have occurred over the period. When compared with Table 2.7, additional statistically significant measures of annual change are detected for the Main NAEP high school mathematics scores and for the Long-Term Trend middle school reading scores. Across all models, the estimates of annual change are shifted in a positive direction when using demographically standardized data.

Comparisons of the annual rate of change for the two models show only one statistically significant difference between Main NAEP and Long-Term Trend. Middle school mathematics trends can no longer be declared different; detectable differences between the trend lines continue for elementary mathematics.

In summary, this chapter shows that changes in the assessment population may cloud the achievements of individual subpopulations. In the period studied here, the difference between the overall average and the demographically standardized average differ by approximately 3 points on the NAEP scales. Three points of change in state averages would result in small differences in rank orders. Whether this change is worth exploring further may be considered in other venues.

7. Conclusions

Summary of Investigative Findings

Average scores in mathematics and reading on Main NAEP and Long-Term Trend were examined at three levels: elementary school, middle school, and high school. The focus was on the average rate of change over the entire period, because the two assessments were conducted simultaneously only in the early years. Shorter-term differences based on nonmatching years may still show apparently different measures of progress for the two surveys. Direct comparisons of two specific years within either survey may also show significant trends that are not necessarily reflected in the long-term measures addressed in this report.

In trends over the period 1990–2009, as measured by average annual change in similarly scaled average scores, Main NAEP and Long-Term Trend produce highly similar results. Differences between the trend lines can be claimed only for mathematics at the elementary and middle school levels. The Main NAEP trend exhibits the higher average annual change in these cases, but both trends show positive progress in elementary and middle school mathematics. Both assessments also indicate positive progress in elementary school reading scores.

Table 7.1 shows the change in score that would be expected on each assessment over a period of 10 years. The right two data columns show the 10-year change in terms of Long-Term Trend scale points; the left two data columns show the 10-year change in terms of population standard deviations.[7] Based on the model, Main NAEP shows a 10-year gain in elementary school mathematics of 14.77 scale points or 0.45 population standard deviations. Long-Term Trend shows smaller mathematics gains of 7.82 scale points or 0.24 population standard deviations. The differences between the two series are statistically significant. Smaller gains in mathematics scores are shown at the middle school level, and the series differences remain statistically significant.

In reading, only elementary-level students showed gains over 10 years, and the two series could not be shown to be different.

[7] The change in terms of population standard deviations is based on the Long-Term Trend standard deviations used to equate the scale of the two NAEP series. See Appendix A for the values used for each subject and grade.

Table 7.1. Summary of model-based trend comparisons: All students in age group or grade

	10 Years' Progress in Population Standard Deviations				10 Years' Progress in Long-Term Trend Scale Points			
	Math		Reading		Math		Reading	
Elementary School Level								
Main NAEP	0.45	**	0.12	**	14.77	**	4.77	**
Long-Term Trend	0.24	**	0.17	**	7.82	**	6.80	**
Difference	–0.21	*	0.05		–6.96	*	2.03	
Middle School Level								
Main NAEP	0.30	**	0.05		9.22	**	2.00	
Long-Term Trend	0.22	**	0.01		6.75	**	0.52	
Difference	–0.08	*	–0.04		–2.47	*	–1.48	
High School Level								
Main NAEP	0.15		–0.11		4.78		–4.91	
Long-Term Trend	0.05		–0.06		1.46		–2.69	
Difference	–0.11		0.05		–3.33		2.22	

*Statistical significance ($p < 0.05$). **High level of statistical significance ($p < 0.01$).

An examination of methodologies used in the two assessments reveals many differences between the two surveys. Population definitions, timing of the data collection, data collection methods, frameworks, and scaling procedures were some aspects of methodology that were different for the two surveys. Most of these distinct methodologies continued over the entire period and are reflected in the comparison of trend measures.

Some methodological features also changed over the period under study. Most notably, accommodation and inclusion policies were amended to include more SDs and more ELLs by offering selective accommodations. These changes were made over the period 1996–2000 for Main NAEP and in 2004 for Long-Term Trend. In both cases, bridge studies were implemented to correct for the impact of methodology change on trend measures. In 2004, Long-Term Trend also modified data collection procedures to be more similar to those used in Main NAEP.

Some changes in the assessment frameworks in Main NAEP were not judged severe enough to affect trend measures. The exception was the grade12 mathematics frameworks; 2005 and 2009 grade12 data are based on new frameworks, will start a new trend line, and were not included in the analyses for this report. The Long-Term Trend frameworks remained relatively unchanged.

Scale adjustment and inclusion of shift parameters for bridge studies were considered sufficient to facilitate a comparison of the trend lines while admitting other unique methodological features as part of each assessment.

Because population definitions are so different for the two surveys, special comparisons of modal groups (modal age within grade and modal grade within age)

were examined using the model-fitting approach to estimate annual rates of change and to compare them across the two surveys (Table 7.2). Restricting the definition to the modal groups results in smaller sample sizes and less ability to detect differences. Similar results were obtained for both surveys on the basis of these smaller student samples. Differences between the surveys in annual rates of changes were detected only for elementary mathematics, although both surveys showed positive progress in elementary and middle school mathematics. Modal reading scores showed improvement at the elementary level (Main NAEP only). Modal reading scores declined for the middle and high school levels as measured by the Long-Term Trend survey.

Table 7.2. Summary of model-based trend comparisons: Students in modal age or grade

	10 Years' Progress in Population Standard Deviations				10 Years' Progress in LTT Scale Points			
	Math		Reading		Math		Reading	
Elementary School Level								
Main NAEP	0.43	**	0.13	*	14.18	**	5.35	*
Long-Term Trend	0.21	*	0.10		6.96	*	4.21	
Difference	–0.22	*	–0.03		–7.22	*	–1.14	
Middle School Level								
Main NAEP	0.24	**	0.00		7.35	**	–0.13	
Long-Term Trend	0.13	**	–0.11	*	4.16	**	–4.21	*
Difference	–0.10		–0.10		–3.19		–4.09	
High School Level								
Main NAEP	0.14		0.00		4.39		0.00	
Long-Term Trend	0.01		–0.16	*	0.38		–6.72	*
Difference	–0.13		–0.16		–4.01		–6.72	

* indicates statistical significance (p <0.05). ** indicates a high level of statistical significance (p < 0.01).

The average scores by racial/ethnic groups (White, Black, and Hispanic) were compared graphically. Although large differences are evident across the groups, the trend lines are relatively similar for Main NAEP and Long-Term Trend within each racial/ethnic group and exhibit the same types of differences between surveys observed for the combined population.

Finally, the impact of changes in the population distribution by racial/ethnic group was examined for Main NAEP. Changes in the population distribution were shown to mask some of the progress that occurs within race/ethnic subgroups when computing population averages on the basis of current year distributions. This situation occurs when the population shifts toward those groups exhibiting lower average performance. Because these population shifts occur for both Main NAEP and Long-Term Trend, average trend may be adjusted for both surveys with some standardization, but the difference or lack of difference between the two trend lines is likely to remain.

Demographically standardized trend comparisons are shown in Table 7.3. When compared with Table 7.1, Table 7.3 shows that significant differences in the trend

lines remain only for elementary mathematics. Generally, the scores have more positive (or less negative) trends when adjusted demographically. Changes in high school–level mathematics scores are now shown to be positive for Main NAEP.

Table 7.3. Summary of model-based trend comparisons: Demographically standardized estimates for all students

	10 Years' Progress in Population Standard Deviations				10 Years' Progress in LTT Scale Points			
	Math		Reading		Math		Reading	
Elementary School Level								
Main NAEP	0.50	*, **	0.20	**	16.56	**	7.87	**
Long-Term Trend	0.28	**	0.20	**	9.21	**	8.18	**
Difference	–0.22	*	0.01		–7.35	*	0.31	
Middle School Level								
Main NAEP	0.33	**	0.09	*	10.20	**	3.38	*
Long-Term Trend	0.25	**	0.07	*	7.92	**	2.81	*
Difference	–0.07		–0.01		–2.29		–0.57	
High School Level								
Main NAEP	0.30	*	–0.09		9.35	*	–3.86	
Long-Term Trend	0.07		–0.04		2.21		–1.80	
Difference	–0.23		0.05		–7.14		2.06	

* indicates statistical significance ($p < 0.05$). ** indicates a high level of statistical significance ($p < 0.01$).

Standardization by gender was also considered, but it showed no impact on trend.

Recommendations

Even though we are satisfied that both assessments give similar results, we wonder which assessment gives better results. Higher trend lines do not answer this question; they tell us only that the results are slightly different. We need to understand the possible uses for these assessment results to determine their validity. Let us first look at the two designs and their possible uses.

Although there are many other differences between the Main NAEP and the Long-Term Trend assessments, the choice between age sampling and grade sampling is of singular importance. The two sampling methods address different although overlapping views of what the assessment is about. NAEP now selects separate samples for the age (Long-Term Trend) trends and the grade (Main NAEP) trends. For a short period, 1983–1990, NAEP collected data for ages and grades simultaneously. This history leads us to the following recommendation:

Recommendation 1: Consider the possible advantages of adding a supplementary age sample to the Main NAEP sample.

Because both age and grade sampling have advantages, we recommend considering doing both age and grade sampling in the Main NAEP sample. This has already been done, and the process is fairly easy. The question is whether the expansion of the Main NAEP sample is worth the effort.

First, let us consider the essential differences between age and grade samples:

- **Age sampling** addresses the question of what people at various ages—both in and out of school—know and can do. Age is attractive because it is easy to define and understand. It also has the same meaning in different jurisdictions, such as states. Ideally, all people of a certain age could be selected for this sample. With age sampling, students in ungraded schools who are now omitted can be selected. In principle, age sampling could include home-schooled children, those in institutions, employed and unemployed drop-outs, and early entrants to college. Age sampling takes a large view of the important variables used to judge the effectiveness of an educational system.

- **Grade sampling** addresses the issue of what students in various grades know and can do. Grade sampling is defined only for in-school students; drop-outs, home-schoolers, and so on, are not assessed. An advantage of this approach is bureaucratic in the sense that decisions are usually made by grade; for example, curricula are defined and standards are set by grade. However, "grade" may have different meanings depending on entry age, retention policies, and so on. In practice, NAEP includes private schools in its national estimates but excludes them from the state estimates because state authorities have less control over private schools.

At the beginning, the NAEP was clearly interested in looking at the nation's performance as a whole and chose age sampling. State and district results were not permitted and therefore were not a decisive issue. The sampling included in-school 9-, 13-, and 17-year-old students. Out-of-school 17-year-olds and young adults ages 26 to 35 were assessed with a household survey. This design produced performance estimates for the selected age levels and contrasts of the performance of in-school and out-of-school 17-year-olds but could not estimate the performance of all students in any particular grade. Out-of-school sampling of adults was discontinued after the 1973–1974 assessment. Out-of-school sampling of 17-year-olds was discontinued after the 1975–1976 assessment.

The 1983–1984 NAEP assessment introduced simultaneous age and grade sampling. The age/grade (grage) sample consisted of students who were either age 9 or grade 4, age 13 or grade 8, and age 17 or grade 11. No out-of-school assessment was done. The resulting sample could be used to estimate either in-school age or grade proficiency population parameters. The first result of this assessment was the *Reading Report Card* (1985), which reported the results by age exclusively. Some grade results were included in the NAEP 1983–1984 Technical Report (Beaton 1987), which was not widely circulated. By 1990, when the Trial State Assessment was introduced, interest in age sampling was minimal, so the sample, which became the Main NAEP sample, was by grade only. The Long-Term Trend lines were continued by using separate in-school, age-only samples.

Which estimates of trends are superior? Both the Main NAEP and Long-Term Trend approaches have pros and cons.

- The Main NAEP approach has the advantage of more up-to-date assessment frameworks and the ability to estimate proficiency on several subscales in both mathematics and reading. The database is sufficient to estimate the proficiency of public school students in states and other jurisdictions. The Main NAEP grade samples are superior for decisions about curricula and standards.

- The Long-Term Trend approach has the advantage of starting in the early 1970s and so goes back about 20 more years. One principle of the Long-Term Trend NAEP design involved retaining a pool of unreleased items to be reused in succeeding assessments to obtain exact comparability at the item level. The available data show some fairly large increases in student performance that would be lost if the grade-only sample were used. The age samples may make it possible to estimate the number of out-of-school students at its selected age levels using data available from other sources.

Thus, we suggest that the NAEP consider returning to age and grade sampling in the Main NAEP assessments. The first step would involve organizing and preparing a hypothetical report showing what could be possible by adding an age dimension to the Main NAEP sample. This step requires input from all NAEP audiences. Reporting of the grade dimension is likely to remain the same as at present. The age sample size and use would depend at first on what data would be easily available from present questionnaires or from public sources. The cost and feasibility of gathering information about out-of-school youth should also be considered. Of particular interest would be estimates of the percentage of students at particular ages who are in different school grades and also the percentage who are out of school. The advantages of placing the out-of-school youth into different categories (e.g., employed, unemployed) should be considered and the cost estimated.

The existing Long-Term Trend might be continued by transforming it onto the Main NAEP metric. This transformation will not be precise but should be adequate for most practical purposes. Separate Long-Term Trend samples could then be dropped.

This first step would show what could be gained by adding age samples to the Main NAEP samples and using only the existing surveys or adding a few survey items. The effect on assessment procedures in the schools would be minimal, involving only a larger list of eligible students. Because Westat has done such sampling in the past, getting reasonably accurate cost estimates is feasible.

What this first step will *not* do is investigate what out-of-school students know and can do. It will not separate different categories of such students. Getting such information would require other techniques, such as household surveys similar to those used in the past. However, going through this process would involve a wholesale review of what NAEP itself can gather and present to educational policymakers and the general public.

The first step in this recommendation should lead to thinking about what the NAEP can do and what it can afford. Further work would depend on the results of the first step.

The in-school age sampling supplement could be based on a national subsample allocated to obtain adequate precision at the national level only. Large samples required for state estimates would primarily use the current Main NAEP methodologies. Private schools should be included in the subsample just as they are currently included in both Main NAEP and Long-Term Trend for the national estimates.

Recommendation 2: Before implementing an age supplement to Main NAEP, complete an investigation of the history of individual items in each of the two surveys.

Particular emphasis should be placed on determining whether any and how many of the items originally included in the Long-Term Trend assessment are still in an unreleased category for measuring trend at the item level. A similar item matching analysis for Main NAEP would also help highlight each assessment's ability to obtain comparable measures over an extended number of periodic surveys.

Recently completed analyses of items have dealt primarily with their match to the specified framework, particularly when the framework has been adjusted or revised.

Recommendation 3: Consider adding subpopulation standardized trend lines to the present reports.

As discussed in Chapter 6, the present national trend lines represent population shifts as well as changes in student performance. It is quite likely that the present NAEP national trend lines are nearly flat but that all racial/ethnic subpopulations have improved their performances. Although the present trends lines should be continued, their interpretation would be enhanced by presenting what the trend would be if the sizes of the demographic subgroups had not shifted. This is precisely what demographic standardization does.

Demographic standardization is simply an estimation methodology for sorting out change that results from improvement within groups versus change in group mix. Groups other than race/ethnicity should be considered. This methodology is closely allied to the methodology for poststratification to marginal controls. Extending this methodology could allow simultaneous standardizing in several dimensions. Demographic standardization should not replace current estimation methods that reflect how things are, given our population mix. Demographic standardization gives a measure of progress that takes account of progress within subgroups and ignores the impact of change in the population mix; it should be considered a supplementary progress reporting mechanism.

Recommendation 4: Initiate an investigation of the in-school student population size by age and grade at the time of assessment and develop estimates of the equivalently aged population not enrolled in school. This change will allow a better understanding of the total success of the educational process, including measures of both retention and the scores obtained by those retained.

This recommendation is particularly important for the grade12 and age 17 populations because current assessment schedules and procedures miss a large,

unknown portion of the total population that might be tested were they still enrolled and attending regularly. Although an out-of-school performance assessment *per se* is not proposed here, it seems important to know the size of the population that cannot be assessed because they are not enrolled. Although some portion of this population may be ahead of the normal progression and be enrolled in college, a much higher proportion likely represents a key failure of the system.

References

Beaton, A. E. (Ed.). (1987). *Implementing the new design: NAEP 1983–84 technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A. E., & Chromy, J. R. (2007). *Partitioning NAEP trend data: A report of the NAEP Validity Studies Panel.* Palo Alto, CA: American Institutes for Research.

Chromy, J. R., Finkner, A. L., & Horvitz, D. G. (2004). Survey design issues. In L. V. Jones, & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 383–425). Bloomington, IN: Phi Kappa Delta Educational Foundation.

Dickinson, E. R., Taylor, L. R., Koger, M. E., Deatz, R. C., & Koger, L. E. (2006). *Alignment of long term trend and main NAEP.* Alexandria, VA: HumRRO.

Dillon, S. (2009a, April 28). Achievement gap for US students hasn't narrowed. In *New York Times,* p. A1. Retrieved September 21, 2010, from http://www.nytimes.com/2009/04/29/education/29scores.html

Dillon, S. (2009b, July 15). Racial gap in testing sees shift by region. *New York Times.* p. A10. Retrieved September 21, 2010, from http://www.nytimes.com/2009/07/15/education/15educ.html?_r=1

Hansen, M. H., Tepping, B. J., Lago, J. A., & Burke, J. (1987). Sample selection and instrument collection. In A. E. Beaton (Ed.), *The NAEP 1983–84 technical report* (pp. 79–96). Princeton, NJ: Educational Testing Service.

Hauser, R. M., Brown, B. V., & Prosser, W. R. (Eds.). (1997). *Indicators of children's well-being.* New York: Russell Sage Foundation. Retrieved September 21, 2010, from http://books.google.com/books?id=ho-LO51XufYC&printsec=frontcover&source=gbs_navlinks_s#v=onepage&q=&f=false

National Center for Education Statistics (NCES). (2009a). *NAEP 2008 trends in academic progress* (NCES 2009-479). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/pdf/main2008/2009479.pdf

NCES. (2009b). *What does the NAEP reading assessment measure?* Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/reading/whatmeasure.asp

NCES. (2009c). *What does the NAEP mathematics assessment measure?* Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/mathematics/whatmeasure.asp.

NCES. (2009d). *What are the differences between long-term trend NAEP and main NAEP?* Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/about/ltt_main_diff.asp

NCES. (2010). *The nation's report card: Mathematics 2009 at grades 4 a*nd 8 (NCES 2010-451). Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/pdf/main2009/2010451.pdf

Perie, M., Moran, R., Lutkus, A. D., & Tirre, W. (2005). *The nation's report card. NAEP 2004: Trends in academic progress. Three decades of student performance in reading and mathematics* (NCES 2005-464). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/pdf/main2005/2005464.pdf

Rust, K. (2004). Sampling and field operations at Westat, 1983 to 2001. In L. V. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 427–448). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Stancavage, F. B., Beaton, A. E., Behuniak, P., et al. (2003). *NAEP validity studies: An agenda for NAEP validity research. A report of the NAEP Validity Studies Panel* (NCES 2003-07). Palo Alto, CA: American Institutes for Research. Retrieved September 21, 2010, from http://nces/ed/gov/pubs2003/200307/pdf

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement gaps: How black and white students in public schools perform in mathematics and reading on the National Assessment of Educational Progress. Statistical analysis report* (NCES 2009-455).Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Education Statistics. Retrieved September 21, 2010, from http://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf

Appendix A. Adjusting the Metric of the Main National Assessment of Educational Progress (NAEP) Samples

Comparing Main NAEP and Long-Term Trend trend lines is made more complex by the fact that the trend lines are constructed in different metrics. The metrics for both sets of trend lines are based on data from calibration samples and judgmental decisions about the reporting of results. The judgmental decisions were made for a number of reasons that include avoiding both negative scores and confusion with the scales of other commonly used tests. The result is that the two data series use similar but not identical metrics, which produce slight differences in trend lines.

To reduce the effect of the reporting metric, we chose to transform the Main NAEP assessment data into the metric of the Long-Term Trend assessment. To do this, we chose to make the mean and standard deviation of student proficiency for the Main NAEP and Long-Term Trend samples the same in the first assessment year in which both samples were assessed. The first year was 1992 for reading and 1990 for mathematics. The transformation was done separately for each age and grade combination.

A simple linear transformation was used. When this transformation is used for individual students, its form is

$$Y_i = b_0 + b_1 X_i$$

where X_i is a student's plausible value on the Main NAEP assessment scale, Y_i is the transformed value, which is in the Long-Term Trend metric, and

$$b_0 = \text{Avg}(Y) - b_1(\text{Avg}(X))$$

$$b_1 = S_y / S_x$$

where S_y and S_x are the standard deviations of the Y and X distributions, respectively. Because student data were not available on the Internet, this model was applied to the summary statistics that were available. The coefficients of the transformation were computed from the first year of the data series and then applied to the following years.

The transformation is quite simple. The result is a transformed Main NAEP distribution that has precisely the same reading or mathematics mean and standard deviation as the corresponding Long-Term Trend sample. The transformation affects the size of a gap (e.g., Males vs. Females) but not the *t*-test used to test the significance of this difference.

The form of the Main NAEP distribution is *not* changed by the transformation: a U-shaped distribution is still U-shaped, although the location and spread may differ. The rank-order of the X_i will not change.

After the transformations were applied, each pair of trend lines starts at the same point, and the changes in performance over the following years are largely due to other factors, not metric differences. We must be cautious, however, because the transformations are based on fallible data, which entail some sampling error.

The computation of the transformation proceeded as follows:

1. The basic data were downloaded from the Internet. These data included the means, standard deviations, and standard errors for each of the Main NAEP and Long-Term Trend samples. The statistics are shown in Table A-1.

2. The transformation constants b_0 and b_1 were computed from Table A-1 and are shown in Table A-2.

3. To check the computations, the transformed Main NAEP values are also shown in Table A-2. The transformed means and standard deviations are the same as those for the Long-Term Trend sample to at least two decimal places.

4. Note that the standard errors of the transformed Main NAEP samples are not the same as those of the Long-Term Trend sample. Both the Main NAEP and Long-Term Trend standard errors were computed by the jackknife method, and the transformed standard errors are the values that would be computed if the transformed data were jackknifed in the same way. These transformed standard errors were computed by multiplying the untransformed Main NAEP standard errors by b_1.

Table A-1. Basic data for transformations

1992 Reading		Main Average	Main SD	Main SE	LTT Average	LTT SD	LTT SE
Grade 4	Age 9	216.74	35.57	0.935	210.52	40.35	0.606
Grade 8	Age 13	260.04	35.89	0.919	259.79	39.40	1.205
Grade 11	Age 17	292.15	32.81	0.550	289.74	43.03	1.118

1990 Mathematics		Main Average	Main SD	Main SE	LTT Average	LTT SD	LTT SE
Grade 4	Age 9	213.07	31.79	0.927	229.64	32.94	0.849
Grade 8	Age 13	262.55	36.02	1.280	270.40	31.07	0.885
Grade 11	Age 17	294.15	35.73	1.110	304.56	31.10	0.905

Note: Main = Main NAEP; LTT = Long-Term Trend.

Table A-2. Translation results

1992 Reading		b_0	b_1	Transformed Main Average*	Transformed Main SD*	Transformed Main SE*
Grade 4	Age 9	–35.30048	1.13420	210.52	40.35	1.061
Grade 8	Age 13	–25.64199	1.09768	259.79	39.40	1.008
Grade 11	Age 17	–93.35925	1.31130	289.74	43.03	0.721

1990 Mathematics		b_0	b_1	Transformed Main Average*	Transformed Main SD*	Transformed Main SE*
Grade 4	Age 9	8.84688	1.03625	229.64	32.94	0.961
Grade 8	Age 13	43.90589	0.86267	270.40	31.07	1.104
Grade 11	Age 17	48.49124	0.87051	304.56	31.10	0.966

Note: b_0 and b_1 = transformation constants, as defined in text; Main = Main NAEP.