Partitioning NAEP Trend Data

Albert E. Beaton
Professor Emeritus, Boston College

James R. Chromy
Senior Fellow, RTI International

September 2007
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. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

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:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
RAND

Larry Hedges
University of Chicago

Gerunda Hughes
Howard University

Robert Linn
University of Colorado

Donald M. McLaughlin
American Institutes for Research

Ina V.S. Mullis
Boston College

Jeffrey Nellhaus
Massachusetts State Department of Education

P. David Pearson
University of California, Berkeley

Lorrie Shepard
University of Colorado

David Thissen
University of North Carolina-Chapel Hill

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
1070 Arastradero Road, Suite 200
Palo Alto, CA 94304-1334
Phone: 650/ 843-8192
Fax: 650/ 858-0958

Acknowledgments

The authors are indebted to Donald McLaughlin, David Grissmer, Jack Buckley, and the NAEP Validity Studies Panel for their helpful comments.

CONTENTS

Introduction....1
The NAEP Data....2
Definition of Partitioning....4
Reading Results....6
Mathematics Results....10
Discussion....13
References....15

LIST OF EXHIBITS

Figure 1. Published and Standardized Mean Scale Scores: Reading, Age 13....1
Figure 2. Published and Standardized Mean Scale Scores: Mathematics, Age 13....2
Table 1. NAEP Reading Data: Long-term Trend, Age 13....3
Table 2: NAEP Mathematics Data: Long Term Trend, Age 13....4
Table 3. Reading Partitioning Summary: 1975–2004....7
Table 4. Reading Partitioning Details: 1975–2004....8
Table 5. Published and Standardized Means: Reading, Age 13....9
Table 6. Reading Partitioning Summary: 1994–2004....9
Table 7. Reading Partitioning Details: 1994–2004....10
Table 8. Mathematics Partitioning Summary: 1978–2004....10
Table 9. Mathematics Partitioning Details: 1978–2004....11
Table 10. Published and Standardized Means: Mathematics, Age 13....12
Table 11. Mathematics Partitioning Summary: 1994–2004....12
Table 12. Mathematics Partitioning Details: 1994–2004....13

Introduction

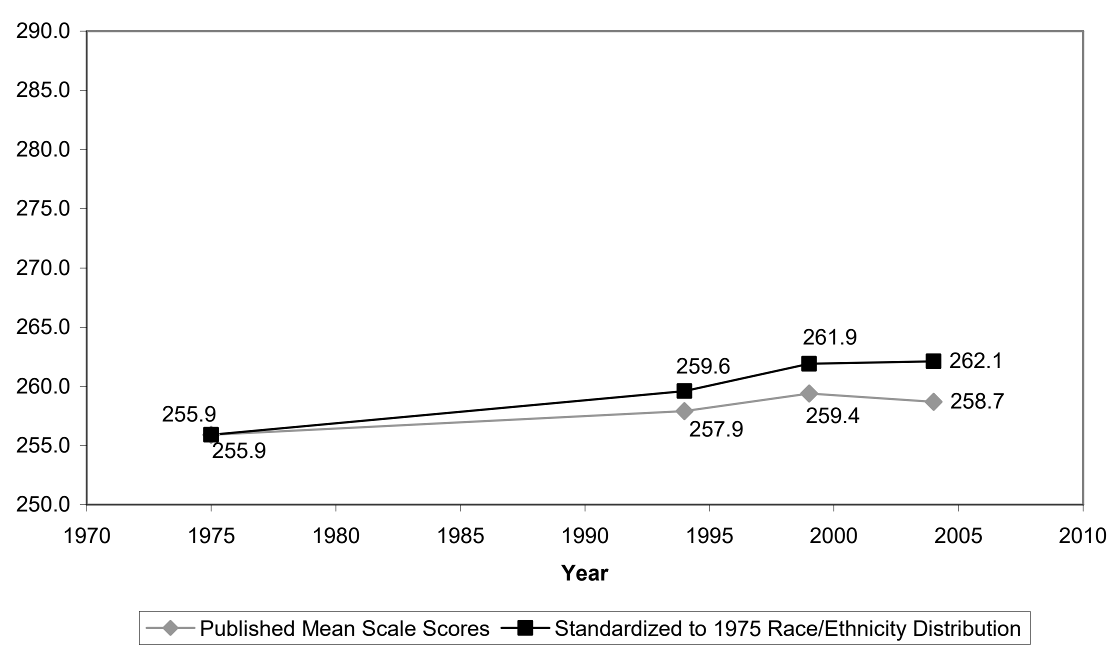
Fundamental to statistical analyses is the comparison of means of one variable from two or more populations. Population samples may be constructed (i.e., experimental and control groups), or they may be natural groupings (i.e., students at a particular grade in different years). If the populations are similar, the mean comparisons are straightforward; if not, the question arises as to whether the mean differences are due to differences in the variable or differences in the populations. Partitioning analysis is a way of distinguishing between these differences.

This paper is a demonstration of how partitioning analysis can be used to help separate changes in reading and mathematical proficiency from changes in school populations over assessment years. NAEP reading and mathematics trend data were readily available from published NAEP reports. Subgroup means were published separately for White, Black, Hispanic, and "Other" students. We selected 13-year-old students from four assessment years as sufficient for this demonstration.

Partitioning analysis separates the difference between two means into three parts: proficiency effect, population effect, and joint effect. The proficiency effect is the change in means attributable to changes in student ability, the population effect is the part attributable to population changes, and the joint effect is the part attributable to the way that the population and proficiency work together. Partitioning analysis makes it simple to compute a well-known statistic, the standardized mean, which estimates what the mean would have been if the percentages of the various subgroups had remained the same.

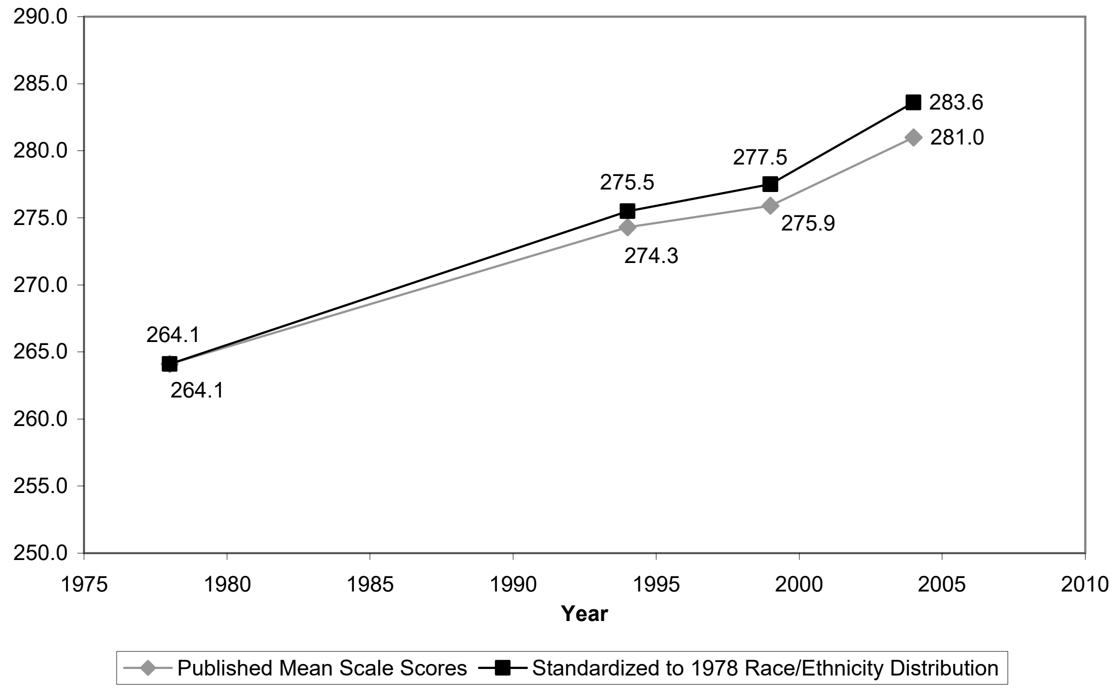
The results for 13-year-olds in reading are shown in figure 1 where the gray line shows the published means and the black line shows the standardized means for comparison. The numerical values are shown in table 5 on page 9.

Figure 1. Published and Standardized Mean Scale Scores: Reading, Age 13



The results for 13-year-olds in mathematics are shown in figure 2 where, again, the gray line shows the published means and the black line shows the standardized means for comparison. The numerical values are shown in table 10 on page 12.

Figure 2. Published and Standardized Mean Scale Scores: Mathematics, Age 13



The two figures summarize the results: In both reading and mathematics the performances of students show improvement, but the shift in populations diminishes the published means. The results are explored further below.

The scope of this demonstration is severely limited by the time and resources that were available. We did not have time or resources to analyze samples from the NAEP database and examine standard errors of the statistics. Nor did we have multi-way tables to explore their possibilities. We hope that future research will expand the scope and utility of partitioning analysis.

The NAEP Data

NAEP data have been collected for nearly 40 years resulting in a huge and complex database. There are long-term trend data, cross-sectional data, state data, etc. To protect privacy, there is a rigorous procedure required to gain access to the data. Our analyses are limited to published data and data available to the public on the NAEP Web site.

For this demonstration, we went to published data that would be of wide interest to policymakers and for which long-term trend data were readily available: reading and mathematics for 13-year-old students. We chose to use the data from an early assessment, the last available assessment, and two in-between assessments for these analyses. The in-between points were chosen to demonstrate short-term partitioning of trend data.

The data were gathered from the Long-term Trend Database,[1] which is available on the NAEP Web site. The reading and mathematics average scores, as well as the percentages of each racial/ethnic group, are available back to the early days of NAEP. In this database, the average score is presented to two decimal places and the percentages are presented as integers. We used more precise percentages for White, Black, and Hispanic students since they were available up until 1999 in *NAEP 1999 Long-term Trend Technical Analysis Report* (Allen, McClellan, and Stoeckel, 2005). The 2004 data used the percentages from the Long-term Trend Database. The "Other" category was calculated as 100 minus the sum of the percentages in the other racial/ethnic categories.

We note that long-term trends inevitably face changes in perspective and procedure. NAEP has gone to extraordinary lengths to keep the trend lines as accurate as possible. We feel comfortable interpreting these trend data and are confident the data are sufficient for demonstrating partitioning analyses.

We chose statistics from the second reading assessment (1975) and from the years 1994, 1999, and 2004. The first year for reading was 1971, but we did not use these data since Hispanic statistics were not published. The year 2004 was the last assessment for which data were available. For mathematics, the first assessment for which full data were available was 1978. The data from this year was selected along with the assessment years 1994, 1999, and 2004.

The reading and mathematics data are shown in tables 1 and 2 respectively. The racial/ethnic groupings use the standard NAEP definitions. The "Other" group contains Asians, American Indians, Pacific Islanders, etc., as well as students whose groupings are unknown.

Table 1. NAEP Reading Data: Long-term Trend, Age 13

Race/ Ethnicity	1975		1994		1999		2004	
	Percent of Total	Scale Score	Percent of Total	Scale Score	Percent of Total	Scale Score	Percent of Total	Scale Score
Total	100.0	255.94	100.0	257.88	100.0	259.42	100	258.69
White	80.9	262.08	73.8	265.08	69.8	266.72	64	265.97
Black	12.7	225.75	14.7	234.31	16.4	238.17	15	244.38
Hispanic	4.9	232.50	8.0	235.14	10.3	243.83	16	242.45
Other	*1.5*	*255.56*	*3.5*	*257.38*	*3.5*	*257.89*	*5*	*264.73*

[1] To access the NAEP Data Explorer for long term trend data, go to http://nces.ed.gov/nationsreportcard , click on Analyze Data, and then click Continue to the long-term trend version of NDE.

Table 2: NAEP Mathematics Data: Long Term Trend, Age 13

	1978		1994		1999		2004	
Race/ Ethnicity	**Percent of Total**	**Scale Score**	**Percent of Total**	**Scale Score**	**Percent of Total**	**Scale Score**	**Percent of Total**	**Scale Score**
Total	100.0	264.13	100.0	274.33	100.0	275.85	100	281.00
White	80.2	271.57	72.9	280.77	71.5	283.14	66	288.35
Black	13.1	229.59	15.3	251.50	15.3	250.98	15	261.75
Hispanic	5.8	237.95	8.1	256.00	9.6	259.16	15	265.11
Other	*0.9*	*272.50*	*3.7*	*283.61*	*3.6*	*282.62*	*4*	*292.42*

The data are shown to the number of decimal places that were available and that we used in the following calculations. In the following tables we will use fewer decimal places.

Definition of Partitioning

The differences among means are explored in many scientific studies. The means come from different populations and, if the populations are similar, then the mean comparisons are straightforward. However, if the populations differ, the mean comparisons are problematic. For example, the problem has occurred and partitioning was applied to explore the SAT decline from 1960 to 1972 (Beaton, Hilton, and Schrader, 1977) and to the study of the NAEP reading anomaly (Beaton and Zwick, 1990).

On a specific schedule, NAEP measures student performance in school subjects such as reading and mathematics, and the performances are compared over time. The mean performances may change because of changes in the students' ability, changes in the populations of students, or for other reasons. In this report, we focus on the effects of changes in student performances and changes in student population distributions.

In this demonstration, the student performance variable will be either reading or mathematics scale scores and the population distribution will be based on student race/ethnicity at different assessment years designated by the subscript *t*. Let us further assume that for each survey year, the overall population can be partitioned into *K* similarly-defined, mutually exclusive, and exhaustive groups. In this demonstration, *K=4* groups will represent White, Black, Hispanic, and "Other" students.

For each survey year designated by *t*, two *K-th* order column vectors can be defined: the vector P_t contains the estimated proportion of the population in each of the *K* groups; and X_t contains the estimated means of the students in the *K* subgroups. The overall population mean for time *t* can be defined as $\overline{X}_t = P_t'X_t$.

The difference in overall means at two times (*t=1* and *t=2*) can then be written as

$$\begin{aligned}\bar{X}_2 - \bar{X}_1 &= P_2'X_2 - P_1'X_1 \\ &= (P_1 + P_2 - P_1)'(X_1 + X_2 - X_1) - P_1'X_1 \\ &= P_1'(X_2 - X_1) + (P_2 - P_1)'X_1 + (P_2 - P_1)'(X_2 - X_1)\,.\end{aligned}$$

Thus, the difference between the two means can be partitioned into three parts:

- the performance effect, $P_1'(X_2 - X_1)$, that displays the mean gain or loss if the population remains the same as in population 1;
- the population effect, $(P_2 - P_1)'X_1$, displays the gain or loss due to changes in population if the subgroup means remain the same as in population1; and
- the joint effect, $(P_2 - P_1)'(X_2 - X_1)$, displays the joint effect of performance and population. This term will be positive if the subgroups that are increasing (decreasing) most in relative size are also increasing (decreasing) most in relative performance, and will be negative if the predominant subgroup performance changes and population changes are in opposing directions.[2]

The three effect components are each vector products, which are weighted sums of the basic data. The components of these sums can be used for diagnostic purposes by analyzing them separately to see how much each subgroup contributed to the overall effects. This usage will be shown in the results sections.

Note that partitioning can be done to any pair of means that have common subgroup definitions. If there are more than two populations, the mean comparisons may be done in pairs as in the examples below. We begin by comparing $\bar{X}_1$ and $\bar{X}_4$, the first and last year for which we have data. We then investigate where the changes happened by comparing $\bar{X}_1$ and $\bar{X}_2$, then $\bar{X}_2$ and $\bar{X}_3$, and finally $\bar{X}_3$ and $\bar{X}_4$. Thus, the $\bar{X}_1$ and $\bar{X}_4$ comparisons are broken up into component parts.

Standardization

Partitioning analysis is closely related to standardization, a well-known statistical technique (see Mosteller and Tukey, 1977). Let us assume that, as above, that the population for any survey year can be partitioned into *K* commonly defined, mutually exclusive, and exhaustive subgroups. The general idea is to decide on a standard population that specifies the proportion of the overall population represented by each of the *K* groups. The standardized mean uses the standard population and the actual means for each group. In the notation here, the standardized mean for any survey year t is $\bar{X}_{t,std} = P_{std}'X_t$ where P_{std}' is a *K-th* order vector that represents the standard population and X_t is the vector of means for survey year *t*. Although not necessary, we have used the earliest year in the series for the standard population and thus $\bar{X}_{1,std} = P_1'X_1 = \bar{X}_1$.

[2] Donald McLaughlin has suggested breaking the joint effect into two lesser components. We have not investigated his suggestion at this time.

Assuming the base year population is used as the standardizing population, the standardized mean for survey year *t (t>1)* can be shown to be $\overline{X}_{t,std=1} = P_1' X_1 + P_1'(X_t - X_1)$ or the base year overall mean plus the proficiency effect. The standardized mean is useful for portraying a trend line under the assumption that the population membership is constant.

Variance Estimation

Variance estimates of the components of the partitioning process or of standardized means are simple if the standardizing population or the individual survey year population vector components can be assumed to be measured without error or with negligible error relative to the mean vectors, and if the estimates of the mean vectors in different survey years are based on independent samples. Under these assumptions, the various partitioning components can be expressed as some linear combination of mean vectors; *e.g.*, $L = A'X_t + B'X_{t'}$. The variance-covariance matrix of the mean vectors can be represented by $Var(X_t) = \Sigma_t$ and can be estimated using appropriate survey analysis software. Then the variance of estimates of these linear combinations can be expressed as $Var(\hat{L}) = A'\hat{\Sigma}_t A + B'\hat{\Sigma}_{t'} B$. If these simplifying assumptions fail, either a complex linearization or replication methods may be used to obtain an approximate variance estimate for the resulting nonlinear statistics.

Reading Results

The reading data in table 1 on page 3 show the means for 4 years: 1975, 1994, 1999, and 2004. First we examine the trend from 1975 to 2004, since NAEP procedures had substantially stabilized in the eighties;[3] and then we examine the shorter, more homogeneous trend from 1994 to 2004.

The reading data in table 1[4] show that the mean was 255.94 in 1975 and 258.69 in 2004, for a change of about 3 points on the NAEP reading scale. The results of the partitioning analysis are shown below in the first row of table 3. The first column notes the two partitioning years, and the next two columns report the reading means for those two years (Year 1 and Year 2). The next column is the difference between those means. The final three columns list the proficiency effect (6.1), population effect:
(-4.3), and joint effect (1.2).

The proficiency effect indicates that the difference between means would have been 6.1 points if the population had remained the same as in 1975, while the change in subgroup means remain as measured. The population effect shows how the changes in the relative size of subgroups affect the overall mean difference. The joint effect shows a tendency for large changes (in both population and proficiency) in the same direction to predominate over changes in the opposite directions. We will look more closely at this below.

Before proceeding, the results are worth a cautionary note. Tests may change over time in subtle ways that affect the mean scores since, for example, test items become more or less relevant to the students or their curricula. The populations may

[3] The procedures for recording the race/ethnicity of students are more comparable among the later years. Comparisons with 1975 based on recorded race/ethnicity may be problematic.
[4] Average scores will be rounded to one decimal place for the rest of this report.

change as the result of immigration or emigration, but population changes may also be due to differences in racial/ethnic definitions or student self-perceptions. It is important not to overly simplify the interpretation of results.

In order to establish when these changes took place, the remaining rows in Table 3 show the partitioning results for adjacent pairs of the years for which data are available, that is 1975–1994, 1994–1999, and 1999–2004. These rows show that the largest change was between 1975 and 1994 where the proficiency effect was 3.7, the population effect was -1.8, and the joint effect was minimal.

Table 3. Reading Partitioning Summary: 1975–2004*

Years	Mean Scale Score		Mean Difference	Partitioned Effects		
	Year 1	Year 2		Proficiency	Population	Joint
75–04	255.9	258.9	3.0	6.1	-4.3	1.2
75–94	255.9	257.9	1.9	3.7	-1.8	0.1
94–99	257.9	259.4	1.5	2.5	-1.2	0.2
99–04	259.4	258.9	-0.5	0.6	-1.0	0.0

*The precision of the partitioning process is influenced by the number of decimal places published in the resource documents used to generate this table. In most cases, the partitioned effects sum to the mean (unadjusted) difference between 2 years to within 1/10th of a scale point. The 2004 population distribution data were published in whole percentage points, only causing larger errors in the partitioning process. To provide consistent calculation for demonstration purposes, the overall scale score for 2004 was recalculated as a weighted sum of the population group estimates using the whole percentages as population weights. As a result the total score for 2004 was changed from 258.7 to 258.9 in tables 3, 4, 6, and 7 for the purposes of this demonstration.

The subgroup details are shown below in table 4. This table is similar in format to table 3 except that columns are inserted for the subgroup percentages for each year.

Table 4. Reading Partitioning Details: 1975–2004

Years	Race/ Ethnicity	First Year % of Total	First Year Scale Score	Second Year % of Total	Second Year Scale Score	Mean Difference	Partitioned Effects Prof.	Partitioned Effects Pop.	Partitioned Effects Joint
75–04	*Total*	*100.0*	*255.9*	*100*	*258.9*	*3.0*	*6.1*	*-4.3*	*1.2*
	White	80.9	262.1	64	266.0	3.9	3.1	-44.3	-0.7
	Black	12.7	225.8	15	244.4	18.6	2.4	5.2	0.4
	Hispanic	4.9	232.5	16	242.5	9.9	0.5	25.8	1.1
	Other	1.5	255.6	5	264.7	9.2	0.1	8.9	0.3
75–94	*Total*	*100.0*	*255.9*	*100.0*	*257.9*	*1.9*	*3.7*	*-1.8*	*0.1*
	White	80.9	262.1	73.8	265.1	3.0	2.4	-18.6	-0.2
	Black	12.7	225.8	14.7	234.3	8.6	1.1	4.5	0.2
	Hispanic	4.9	232.5	8.0	235.1	2.6	0.1	7.2	0.1
	Other	1.5	255.6	3.5	257.4	1.8	0.0	5.1	0.0
94–99	*Total*	*100.0*	*257.9*	*100.0*	*259.4*	*1.5*	*2.5*	*-1.2*	*0.2*
	White	73.8	265.1	69.8	266.7	1.6	1.2	-10.6	-0.1
	Black	14.7	234.3	16.4	238.2	3.9	0.6	4.0	0.1
	Hispanic	8.0	235.1	10.3	243.8	8.7	0.7	5.4	0.2
	Other	3.5	257.4	3.5	257.9	0.5	0.0	0.0	0.0
99–04	*Total*	*100.0*	*259.4*	*100*	*258.9*	*-0.5*	*0.6*	*-1.0*	*0.0*
	White	69.8	266.7	64	266.0	-0.8	-0.5	-15.5	0.0
	Black	16.4	238.2	15	244.4	6.2	1.0	-3.3	-0.1
	Hispanic	10.3	243.8	16	242.5	-1.4	-0.1	13.9	-0.1
	Other	3.5	257.9	5	264.7	6.8	0.2	3.9	0.1

The first panel in table 4 has the subgroup details for the entire span from 1975–2004. We note that the means of *all* subgroups improved, with the Black subgroup increasing the most (18.6 points).

The entries under partitioned effects for race/ethnic group show the contribution of each group to the overall effects. They should not be interpreted as proficiency effects, population effects, or joint effects at the race/ethnic group level. Since changes in proficiency are weighted by the first year population distribution, the largest impacts on the proficiency effect tend to accrue to the largest race/ethnic group when all show similar improvements in scale scores. Recall that population effects hold the year scale scores constant and just show the impact of population distribution changes. The direction and magnitude of the contribution to the total effect is governed primarily by the change in population percent for each race/ethnic group.

The remaining panels in table 4 show the group details for each link in the overall partition. It is interesting to note that the White, Black, and Hispanic groups generally contribute positively to the proficiency effect until the 1999–2004 link where there is a slight dip (-0.5) in the overall mean difference and where the population shift reduces the small positive proficiency effect.

The standardized means are shown below in table 5 and on page 1 in figure 1. Note that the standard population is the year 1975 assessment students, and that the later populations are compared to it. In this study, the standardized means are always higher than the published means, except for 1975, which is algebraically identical. The higher standardized means indicates that population shifts have a negative effect on the trend values.

Table 5. Published and Standardized Means: Reading, Age 13

Reading	1975	1994	1999	2004
Published Mean Scale Scores	255.9	257.9	259.4	258.7
Standardized to 1975 Race/Ethnicity Distribution	255.9	259.6	261.9	262.1

The shorter trend analysis for 1994–2004 is shown in table 6, where the overall mean difference is just 0.8. The proficiency effect is 3.0, the population effect is -2.6, and the joint effect is .6. Therefore, the increase in performance is diminished by the population shift.

Table 6. Reading Partitioning Summary: 1994–2004

Years	Mean Scale Score		Mean Difference	Partitioned Effects		
	Year 1	Year 2		Proficiency	Population	Joint
94–04	257.9	258.9	1.0	3.0	-2.6	0.6
94–99	257.9	259.4	1.5	2.5	-1.2	0.2
99–04	259.4	258.9	-0.5	0.6	-1.0	0.0

The subgroup details are given in table 7. Each subgroup improved performance between 1994 and 2004 and thus contributed to the positive proficiency effect of 3.0. The population effect counteracted the proficiency effect with large decreases in the White population and large increases in the growing Hispanic group.

Table 7. Reading Partitioning Details: 1994–2004

Years	Race/ Ethnicity	Year 1		Year 2		Mean Difference	Partitioned Effects		
		% of Total	Scale Score	% of Total	Scale Score		Prof.	Pop.	Joint
94–04	*Total*	*100*	*257.9*	*100*	*258.9*	*1.0*	*3.0*	*-2.6*	*0.6*
	White	73.8	265.1	64	266.0	0.9	0.7	-26.0	-0.1
	Black	14.7	234.3	15	244.4	10.1	1.5	0.7	0.0
	Hispanic	8	235.1	16	242.5	7.3	0.6	18.8	0.6
	Other	3.5	257.4	5	264.7	7.4	0.3	3.9	0.1
94–99	*Total*	*100*	*257.9*	*100*	*259.4*	*1.5*	*2.5*	*-1.2*	*0.2*
	White	73.8	265.1	69.8	266.7	1.6	1.2	-10.6	-0.1
	Black	14.7	234.3	16.4	238.2	3.9	0.6	4.0	0.1
	Hispanic	8	235.1	10.3	243.8	8.7	0.7	5.4	0.2
	Other	3.5	257.4	3.5	257.9	0.5	0.0	0.0	0.0
99–04	*Total*	*100*	*259.4*	*100*	*258.9*	*-0.5*	*0.6*	*-1.0*	*0.0*
	White	69.8	266.7	64	266.0	-0.8	-0.5	-15.5	0.0
	Black	16.4	238.2	15	244.4	6.2	1.0	-3.3	-0.1
	Hispanic	10.3	243.8	16	242.5	-1.4	-0.1	13.9	-0.1
	Other	3.5	257.9	5	264.7	6.8	0.2	3.9	0.1

Our general conclusion is that these important populations are all improving in reading, but the published trend means are lessened by population shifts.

Mathematics Results

The mathematics data in table 2 on page 4 show the means for 4 years: 1978, 1994, 1999, and 2004. First, we examine the trend from 1978 to 2004, and then we examine the shorter trend from 1994 to 2004.

The mathematics data in table 2 show that the mean was 264.13 in 1978 and 281.00 in 2004, for a change of 16.87 points on the NAEP mathematics scale. The summary results of the partitioning analysis are shown in the first row of Table 8. The columns are the same as for the reading summary, but with mathematics results inserted. The final three columns are the proficiency effect (19.4), population effect (-3.9), and joint effect (1.3).

Table 8. Mathematics Partitioning Summary: 1978–2004

Years	Mean Scale Score		Mean Difference	Partitioned Effects		
	Year 1	Year 2		Proficiency	Population	Joint
78–04	264.1	281.0	16.9	19.4	-3.9	1.3
78–94	264.1	274.3	10.2	11.4	-1.7	0.5
94–99	274.3	275.9	1.5	1.9	-0.4	0.0
99–04	275.9	281.0	5.1	6.3	-1.2	0.0

Analogous to reading, the proficiency effect indicates that the difference between means would have been 19.4 points if the population had remained the same as in 1978, while the subgroup means remained as reported. The population effect shows how the changes in the relative size of subgroups affect the overall mean difference. The positive joint effect shows a tendency for the subgroups that have an increasing population proportion to have larger mean increases, and vice versa. We will look more closely at this below.

In order to establish when these changes took place, the remaining rows in table 8 show the partitioning results for adjacent pairs of the years for which data are available, that is 1978–1994, 1994–1999, and 1999–2004. These rows show that the largest change was between 1978 and 1994 where the proficiency effect was 11.4, the population effect was -1.7, and the joint effect was 0.5.

The subgroup details are shown in table 9. This table is similar in form to table 8 except for the addition of subgroup percentages for each pair of years.

Table 9. Mathematics Partitioning Details: 1978–2004

		Year 1		Year 2			Partitioned Effects		
Years	Race/ Ethnicity	% of Total	Scale Score	% of Total	Scale Score	Mean Difference	Prof.	Pop.	Joint
78–04	*Total*	*100.0*	*264.1*	*100*	*281.0*	*16.9*	*19.4*	*-3.9*	*1.3*
	White	80.2	271.6	66	288.4	16.8	13.5	-38.6	-2.4
	Black	13.1	229.6	15	261.8	32.2	4.2	4.4	0.6
	Hispanic	5.8	238.0	15	265.1	27.2	1.6	21.9	2.5
	Other	0.9	272.5	4	292.4	19.9	0.2	8.4	0.6
78–94	*Total*	*100.0*	*264.1*	*100*	*274.3*	*10.2*	*11.4*	*-1.7*	*0.5*
	White	80.2	271.6	72.9	280.8	9.2	7.4	-19.8	-0.7
	Black	13.1	229.6	15.3	251.5	21.9	2.9	5.1	0.5
	Hispanic	5.8	238.0	8.1	256.0	18.1	1.0	5.5	0.4
	Other	0.9	272.5	3.7	283.6	11.1	0.1	7.6	0.3
94–99	*Total*	*100.0*	*274.3*	*100.0*	*275.9*	*1.5*	*1.9*	*-0.4*	*0.0*
	White	72.9	280.8	71.5	283.1	2.4	1.7	-3.9	0.0
	Black	15.3	251.5	15.3	251.0	-0.5	-0.1	0.0	0.0
	Hispanic	8.1	256.0	9.6	259.2	3.2	0.3	3.8	0.0
	Other	3.7	283.6	3.6	282.6	-1.0	0.0	-0.3	0.0
99–04	*Total*	*100.0*	*275.9*	*100*	*281.0*	*5.1*	*6.3*	*-1.2*	*0.0*
	White	71.5	283.1	66	288.4	5.2	3.7	-15.6	-0.3
	Black	15.3	251.0	15	261.8	10.8	1.6	-0.8	0.0
	Hispanic	9.6	259.2	15	265.1	5.9	0.6	14.0	0.3
	Other	3.6	282.6	4	292.4	9.8	0.4	1.1	0.0

The first panel in table 9 has the subgroup details for the entire span from 1978–2004. We note that the means of *all* subgroups improved.

The remaining panels in table 9 show the subgroup details for each link in the overall partition. It is interesting to note that the White, Black, and Hispanic subgroups generally contribute positively to the proficiency effect except for the miniscule contribution of -0.1 for Blacks in the 1994–1999 partition. In other words, each racial/ethnic proficiency improved, but the shift in populations brought about an attenuated mean difference.

The standardized means are shown in table 10 below and in figure 2 on page 2. Note that the standard population is defined as 1978 and that the later populations are compared to it. In this demonstration, the standardized mean is always higher than the published mean, indicating that population shifts have a negative effect on the trend values.

Table 10. Published and Standardized Means: Mathematics, Age 13

Mathematics	1978	1994	1999	2004
Published Mean Scale Scores	264.1	274.3	275.9	281.0
Standardized to 1978 Race/Ethnicity Distribution	264.1	275.5	277.5	283.6

The trend analysis for 1994–2004 is shown in table 11 where the overall mean difference is 6.7. The proficiency effect is 8.2, the population effect is -1.6, and the joint effect is 0.1. Therefore, the increase in performance is diminished by the population shift. In the lower rows, each link shows a positive proficiency effect, negative population effect, and small joint effect, which also indicates that improvement in performance is attenuated by the population shifts.

Table 11. Mathematics Partitioning Summary: 1994–2004

Years	Mean Scale Score		Mean Difference	Partitioned Effects		
	Year 1	Year 2		Proficiency	Population	Joint
94–04	274.3	281.0	6.7	8.2	-1.6	0.1
94–99	274.3	275.9	1.5	1.9	-0.4	0.0
99–04	275.9	281.0	5.1	6.3	-1.2	0.0

The subgroup details are given in table 12. Looking at the first panel that covers the 1994–2004 span, each subgroup improved performance and thus contributed to the positive proficiency effect of 8.2. The population effect counteracted the proficiency effect. The other two panels show that the proficiency effect was smaller in the 1994–1999 years (1.9 points) than in the 1999–2004 years (6.3 points).

Table 12. Mathematics Partitioning Details: 1994–2004

Years	Race/ Ethnicity	Year 1		Year 2		Mean Difference	Partitioned Effects		
		% of Total	Scale Score	% of Total	Scale Score		Prof.	Pop.	Joint
94–04	*Total*	*100.0*	*274.3*	*100*	*281.0*	*6.7*	*8.2*	*-1.6*	*0.1*
	White	72.9	280.8	66	288.4	7.6	5.5	-19.4	-0.5
	Black	15.3	251.5	15	261.8	10.3	1.6	-0.8	0.0
	Hispanic	8.1	256	15	265.1	9.1	0.7	17.7	0.6
	Other	3.7	283.6	4	292.4	8.8	0.3	0.9	0.0
94–99	*Total*	*100.0*	*274.3*	*100.0*	*275.9*	*1.5*	*1.9*	*-0.4*	*0.0*
	White	72.9	280.8	71.5	283.1	2.4	1.7	-3.9	0.0
	Black	15.3	251.5	15.3	251.0	-0.5	-0.1	0.0	0.0
	Hispanic	8.1	256.0	9.6	259.2	3.2	0.3	3.8	0.0
	Other	3.7	283.6	3.6	282.6	-1.0	0.0	-0.3	0.0
99–04	*Total*	*100.0*	*275.9*	*100*	*281.0*	*5.1*	*6.3*	*-1.2*	*0.0*
	White	71.5	283.1	66	288.4	5.2	3.7	-15.6	-0.3
	Black	15.3	251.0	15	261.8	10.8	1.6	-0.8	0.0
	Hispanic	9.6	259.2	15	265.1	5.9	0.6	14.0	0.3
	Other	3.6	282.6	4	292.4	9.8	0.4	1.1	0.0

Our general conclusion is that these important populations are all improving in mathematics, but the published trend means are lessened by population shifts.

Discussion

This demonstration has shown the usefulness of partitioning analysis in answering questions about how much the difference between two averages is attributable to population shifts as opposed to changes in the ability of various subgroups. Partitioning is a straight-forward and simple approach to the proficiency versus population issue. It makes clear what the assumptions and limitations are. The necessary data are clear and often available in published reports.

The examples analyzed here are the reading and mathematics performances of 13-year-old students in NAEP national samples in different school years. The data were classified by racial/ethnic groupings. The results showed that each racial/ethnic group improved during the selected time spans, while the population shifts diminished the measure of increased performance.

The results suggest speculation and future research. What were the population changes? Were the changes due to immigration/emigration or due to changes in NAEP's sampling and inclusion policies? Partitioning does not answer these questions, but suggests further research for empirical answers.

However, partitioning lacks an important property that requires further research: standard errors. The results in this paper do not have any indicator of statistical accuracy, which is highly desirable. We believe that we can approach this issue using available software, but application would require using the micro-data in the NAEP public use data tape. Our time and budget constraints did not allow us to pursue this avenue.

There may be other approaches that address the performance/population issues.[5] The issue is important in a number of different scientific studies, notably international comparisons. We think that partitioning is worth further study and expansion.

[5] Jack Buckley suggested a regression approach that is worth pursuing in the future.

References

Beaton, Albert E., Hilton, Thomas L., & Schrader, W. B. (1977). *Changes in the verbal abilities of high school seniors, college entrants and SAT candidates between 1960 and 1972*. Princeton, NJ: Educational Testing Service.

Beaton, Albert E. & Zwick, Rebecca (foreword by J.W. Tukey). (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985–86 reading anomaly*. Princeton, NJ: Educational Testing Service.

Mosteller, Frederick & Tukey, John W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Allen, Nancy L., McClellan, Catherine A., & Stoeckel, Joan J. (2005). *NAEP 1999 long-term trend technical analysis report: Three decades of student performance.* (NCES 2005-484). Washington, DC: U.S. Department of Education, National Center for Education Statistics.