THE NAEP 1985-86 READING ANOMALY:
A TECHNICAL REPORT

Albert E. Beaton

in collaboration with

John J. Ferris
Eugene G. Johnson
Janet R. Johnson
Robert J. Mislevy
Rebecca Zwick

Educational Testing Service, Princeton, NJ
February 1988

The National Assessment of Educational Progress is funded by the U.S. Department of Education under a grant to Educational Testing Service. National Assessment is an education research project mandated by Congress to collect data over time on the performance of young Americans in various learning areas. It makes available information on assessment procedures to state and local education agencies.

The work upon which this report is based was performed pursuant to Grant No. NIE-G-83-0011 of the Office for Educational Research and Improvement, Center for Education Statistics. It does not, however, necessarily reflect the views of that agency.

The NAEP 1985-86 Reading Anomaly:
A Technical Report

Table of Contents

Executive Summary

Albert E. Beaton

The 1985-86 National Assessment of Educational Progress (NAEP) was designed to assess the proficiency of 9-, 13-, and 17-year-olds as well as third, seventh, and eleventh grade students in a number of subject areas, including reading, mathematics, science, computer competence, and, for 17-year-olds and eleventh graders, history and literature. This national assessment was also designed to measure changes in the performance in reading, mathematics, and science from comparable groups measure in past assessments.

Responsible reporting of national results implies an application of judgment as well as an insistence on high technical standards and rigorous quality control. The results of the analysis of the NAEP data and its comparison with the results of the 1983-84, and previous, reading assessments suggest a substantial decrease in performance in reading for 9- and 17-year-olds but a slight improvement for 13-year-olds. There were no substantial changes in performance from past assessments in mathematics or science at any age levels. The decreases in reading performance were anomalous enough for the NAEP staff to question the accuracy of the assessment results. Eminent, outside educational researchers also examined the NAEP reading results and concurred with the NAEP staff in their judgment that the results were unusual and advised further investigation.

The NAEP staff have examined a number of hypotheses about what might have caused such anomalous results. They have looked for changes in the student population, failures in the sampling process, modifications of the NAEP design and administration, lapses in quality control, computer bugs, and so forth. They have also explored whether or not some external event, such as the Challenger tragedy, might have affected student performance.

The results of the studies of the reading anomaly are inconclusive. Some hypotheses, such as inaccuracies in sampling, scaling, and quality control, can be ruled out beyond any reasonable doubt. However, some changes in the assessment process are inevitable, and these changes are documented in this report. The possibility that one or a combination of such changes may have resulted in the declines in reading proficiency cannot be ruled out. The effect of such changes cannot be estimated from existing data.

The 1987-88 National Assessment has been modified to gather data to measure the effect of the changes made in the 1985-86 assessment. Equivalent random samples of students will be measured using the 1983-84, 1985-86, and 1987-88 reading assessment methodologies, and the results will be compared. Comparison of the samples using 1983-84 and 1985-86 technology should be able to isolate how much, if any, the changes in methodology affected the reading results. If the estimated effect of methodological changes is trivial, then the 1985-86 results suggest a substantial decline in reading proficiency. If the estimated effect of methodological changes is substantial, then the results of the 1985-86 assessment can be adjusted for their disadvantage. In either case, the 1985-86 reading results will be released as soon as these analyses can be completed.

Acknowledgments

This report is the culmination of many months of research into the unusual results of the NAEP 1986 reading assessment. The challenge of the anomalous reading data was met with creativity and dedication by the psychometricians and statisticians who collaborated in writing the report: Jim Ferris, Gene Johnson, Janet Johnson, Bob Mislevy, and Rebecca Zwick.

The authors are particularly indebted for the counsel of NAEP's Design and Analysis Committee, whose members are Robert Linn (Chair), John B. Carroll, Robert Glaser, Bert Green, Jr., Sylvia Johnson, Ingram Olkin, Tej Pandey, Richard Snow, and John W. Tukey. The authors are also indebted to W. B. Schrader and John W. Tukey, who served as special consultants.

The authors were assisted in the formulation and investigation of their hypotheses by a talented and dedicated team of data analysts under the direction of John Barone. Significant contributions of time, energy, and ideas in the data analysis effort were made by Laurie Barnett, David Freund, Doug Gentile, Bruce Kaplan, Ed Kulick, Laura McCamley, Norma Norris, Al Rogers, and Ira Sample. Jo-Ling Liang ably assisted the research effort.

Ina Mullis provided many important suggestions and comments. We are also beholden to Henry Braun, Archie Lapointe, and Sam Messick for their valuable reviews.

Morris Hansen, Renee Slobasky, Nancy Caldwell, and Keith Rust of Westat, Inc., also deserve special mention for their assistance.

We would also like to thank the Office for Educational Research and Improvement, Center for Education Statistics, of the Department of Education for its help. In particular, we would like to thank Gary Phillips, Gene Owen, and the anonymous reviewers of an earlier draft of this document.

The report was prepared under the outstanding editorial supervision of Debbie Kline. Mary Varone, Marcy Yates, and Sharon Stewart provided word processing assistance and careful attention in the preparation of the manuscript.

Albert E. Beaton
Director of Data Analysis
February 1988

Chapter 1

INTRODUCTION

Albert E. Beaton

The NAEP reading results for the 1986[1] assessment are anomalous, and the NAEP staff as well as its Design and Analysis Committee question whether these results are accurate measures of educational progress. At ages 9 and 17, the estimated reading proficiency of students appears to be sharply lower than the previous assessment in 1984, whereas there was a slight rise in reading proficiency at age 13. The apparent declines in reading proficiency at age 9 and especially age 17 are so large over the two-year period that we doubt that actual changes of this size would have been unnoticed. We believe that these results should not be used for estimating trends in American education until supported by corroborating evidence.

Although the process of selecting and administering a national assessment and then estimating the proficiencies of the nation's students seems straightforward, the process is sensitive to many different extraneous influences. In one sense, the decline at age 17 in the students' performance is really quite small, since the average assessment question was answered correctly by only 3 percent fewer students, but this decline is much larger than would be expected in two years. Average student performance changes slowly, and an assessment must be precise enough to separate the trends in student performance from other factors that might influence the assessment results. Factors that might influence the estimated trends in student performance in an extraneous way include changes in population composition, selection of an unrepresentative sample, and changes in the motivation of the students at the time of assessment. Changes in assessment technology involving assessment instruments, administrative procedures, and analytic methods may also have noticeable effects on the estimated performance of the nation's students. The NAEP staff have not been able to conclude that the 1986 results are in fact the result of an actual decline in student performance and not the result of some factors unrelated to student reading proficiency.

Since there is still a reasonable doubt about the 1986 reading trend results, the NAEP staff have had to decide whether or not to publish them at this time. There are two ways to err in making such a decision: the results could be published when they are in fact wrong or they could not be published when the decline is in fact real. Since either error is possible, the choice between them depends on one's judgment as to which type of error is more likely and which could have a more deleterious effect on the students in

[1] Throughout this report, the assessments conducted during the 1983-84, 1985-86, and 1987-88 school years are referred to respectively as the 1984, 1986, and 1988 assessments.

American schools. Our decision has been to forgo publishing trend results until more evidence about their accuracy has been obtained.

* * *

Substantial time and energy has been spent in trying to understand why the results seem anomalous. At first, it was assumed that there must have been a data processing error, but this possibility was soon ruled out beyond any reasonable doubt. Afterwards, many different hypotheses about technical issues were investigated; the results were ultimately inconclusive. Although many possible causes have been ruled out, some cannot be investigated without additional data. Two eminent consultants[2] outside of the NAEP staff were consulted for their insights. The anomaly was further discussed at several meetings of the NAEP Design and Analysis Committee.[3]

The Design and Analysis Committee has advised the NAEP staff to:

- revise the design of the 1988 assessment to gather data for exploring the technical factors that might have produced the apparent decline;

- withhold the publication of reading trend results until the 1988 data are available to support or discredit the apparent decline in reading proficiency; and

- withhold the distribution of the reading data until a paper summarizing reservations about the data is prepared.

We have followed the recommendations of the Design and Analysis Committee.

* * *

The purpose of this report is threefold: to describe the results of the 1986 assessment; to describe the research that has been done to detect technical problems that might have lowered the estimated reading performance; and to describe the steps that have been taken to collect additional data that may illuminate why the results are so different from previous assessments.

[2] Dr. John Tukey, Statistics Professor Emeritus, Princeton University and Dr. W. B. Schrader, former Director of Statistical Analysis at Educational Testing Service.

[3] Members of the Design and Analysis Committee: Robert Linn (Chair), John B. Carroll, Robert Glaser, Bert Green, Jr., Sylvia Johnson, Ingram Olkin, Tej Pandey, Richard Snow, and John W. Tukey. Barbara Shapiro served as observer for the NAEP Assessment Policy Committee.

The results of the 1986 assessment are described in Chapter 2. We note that the sharp declines occur at only two of the three ages at which reading was assessed, and there is no corresponding decline in mathematics or science, which were also assessed. The reading decline at age 17 was the most substantial and therefore studied most closely. The decline was found to be pervasive, occurring for all of the traditional NAEP reporting groups, such as all regions of the country, sexes, racial/ethnic groupings, etc.

Chapter 3 contains a summary of the design of the 1986 NAEP and a brief summary of the explorations into the possible causes of the anomalous results, including the following hypotheses:

- the population of students attending American schools has changed, or that the NAEP sample did not accurately represent that population;

- the assessment instruments were not comparable with past instruments;

- changes in administrative procedures made the results not comparable with past results;

- the decline was due to lapses in quality control during the data processing;

- the decline is an artifact of the scaling procedures that were used;

- the students responded in substantially different ways than in the past to the reading exercises that were presented;

- some assessment booklets or blocks of exercises were flawed; and

- some external event occurred that might have made the students less involved in the assessment process.

Detailed reports of these and other hypotheses are provided in the following chapters.

Chapter 13 contains a brief summary of conclusions and a description of new data that will be collected during the 1988 assessment for the purpose of determining whether or not a substantial decline in average reading proficiency actually took place.

Chapter 2

1986 READING RESULTS

Albert E. Beaton

The data from the 1986 reading assessment suggest that the decline in reading proficiency between 1984 and 1986 is 11.4 points at age 17 and 5.6 points at age 9. Assuming that growth between ages 13 and 17 was linear in 1984, the estimated average reading proficiency of 17-year-olds in 1986 is at about the same level as 16-year-olds in 1984. Assuming linear growth at the early ages, the estimated average reading proficiency of 9-year-olds in 1986 is at about the same level as 8.5-year-olds in 1984. We do not believe that declines of this magnitude over a two-year period could have gone unnoticed by educators.

The national trends in average reading proficiency for years 1971-1984 for 9-, 13-, and 17-year-olds are shown in Table 2-1. The data up to 1984 were originally published after the 1984 data collection in The Reading Report Card: Progress Toward Excellence in Our Schools (ETS, 1985).

Figures 2-1, 2-2, and 2-3 compare the estimated distributions of reading proficiency in the 1984 and 1986 assessments for the three age groups respectively. The top panel in each figure depicts the 1984 estimated distribution and the middle panel depicts the estimated distribution for 1986. The distributions all cover the same range (50 to 450) and the location of the mean and the size of the standard deviations are indicated on the abscissa. The frequencies are reported in tenths of percents of the estimated population. The bottom panel superimposes the two distributions; a plus sign (+) indicates that a larger proportion of the population was at that point in 1984 and a minus sign (-) indicates a larger proportion in 1986.

We note that for all ages there are estimated to be both more very high scorers and very low scorers, indicating an increase in variance. The distribution at age 9 shows a somewhat larger proportion of low scorers and a smaller proportion in the middle. The distribution at age 13 shows a smaller proportion of students scoring near the mean with more students scoring at both extremes. At age 17 there is a small increase in the proportion of students scoring very highly but a very large increase in the proportion with very low scores.

Table 2-2 contains some basic statistics relevant to the change in reading proficiency between 1984 and 1986. The top panel of this table displays the sizes of the NAEP subsamples used in estimating reading proficiency trends. The remaining panels contain the estimated average performance on the NAEP reading scale and the estimated standard deviations for each age group in each of the two years.

The decline in average reading performance is also evident in the average percents-correct metric that was used for reporting trends in the past. The changes in average percents correct are shown in Table 2-3. The declines from 1984 to 1986 are large for the 9- and 17-year-olds and there is a slight increase in average percent correct at the 13-year-old level. The declines are much larger than the estimated changes between 1980 and 1984.

The 1986 results in mathematics and science do not show corresponding declines. The estimated trends in mathematics and science proficiency will be published shortly. We note, however, that there were substantial differences between the reading trend assessment and the assessments in mathematics and science: for example, mathematics and science were last assessed in 1982 using the paced, aural presentation method of administration. The 1986 students participating in the trend estimation for mathematics and science were also assessed using a tape recorder, eliminating the need for the student to read the assessment instructions or questions. Both the 1984 and 1986 students participating in the reading trend analyses were assessed using the read-and-respond method. It is not possible at this time to establish whether a decline would have taken place if similar methods had been used for mathematics and science.

Table 2-1

Weighted Reading Proficiency Means and Jackknifed Standard Errors*

National Trend Results Across Five Assessments

	Age 9	Age 13	Age 17
1971	207.3 (1.0)	255.2 (0.9)	285.4 (1.2)
1975	210.2 (0.7)	256.0 (0.8)	286.1 (0.8)
1980	214.8 (1.1)	258.5 (0.9)	285.8 (1.4)
1984	212.9 (1.0)	258.0 (0.7)	288.8 (0.9)
1986	207.3 (1.4)	260.4 (1.1)	277.4 (1.0)

* Standard errors in parentheses

Figure 2-1
Comparison of 1984 and 1986 Reading Scores Using First Plausible Scale Value
Age 9, Weighted

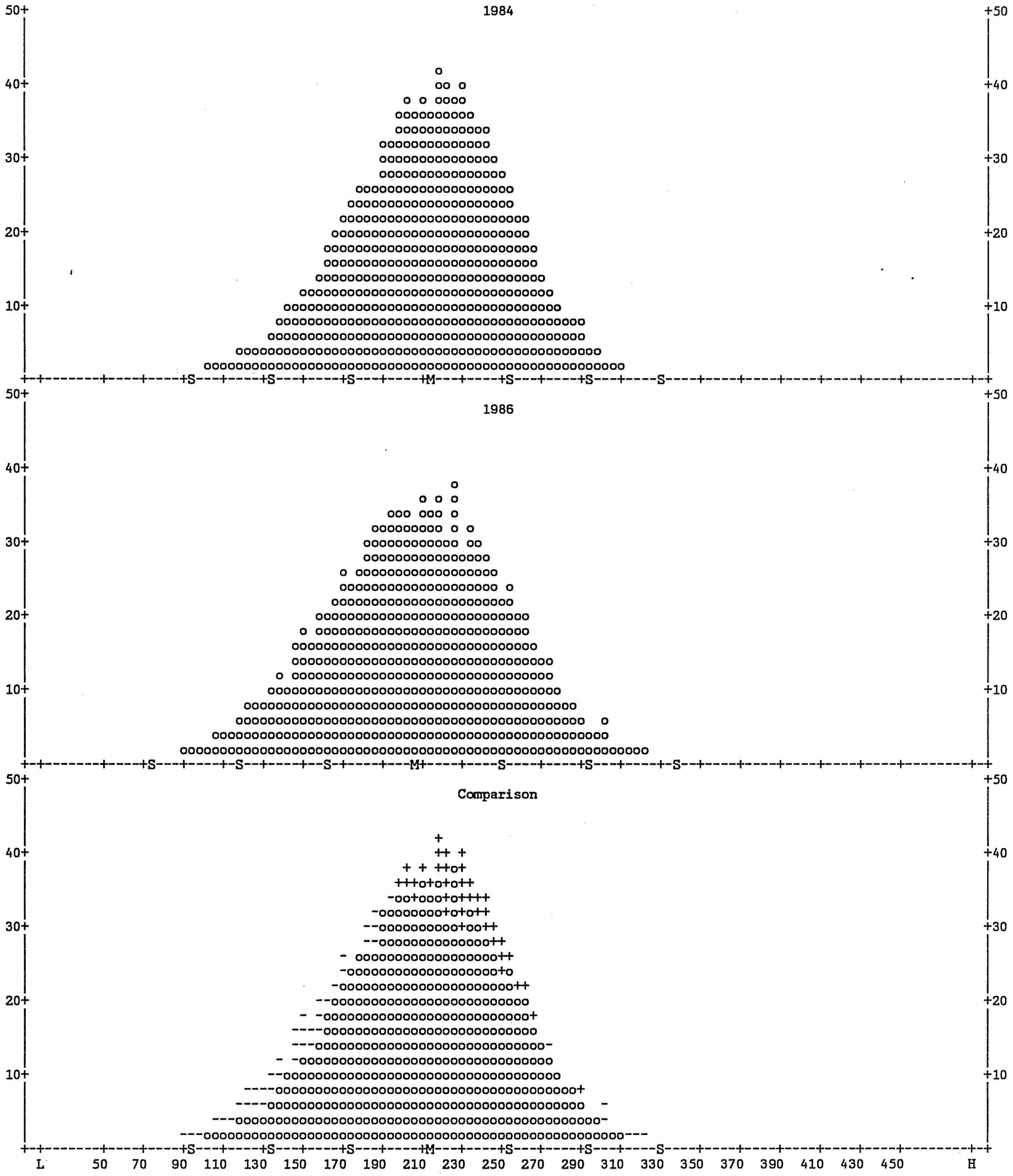


Figure 2-2
Comparison of 1984 and 1986 Reading Scores Using First Plausible Scale Value
Age 13, Weighted

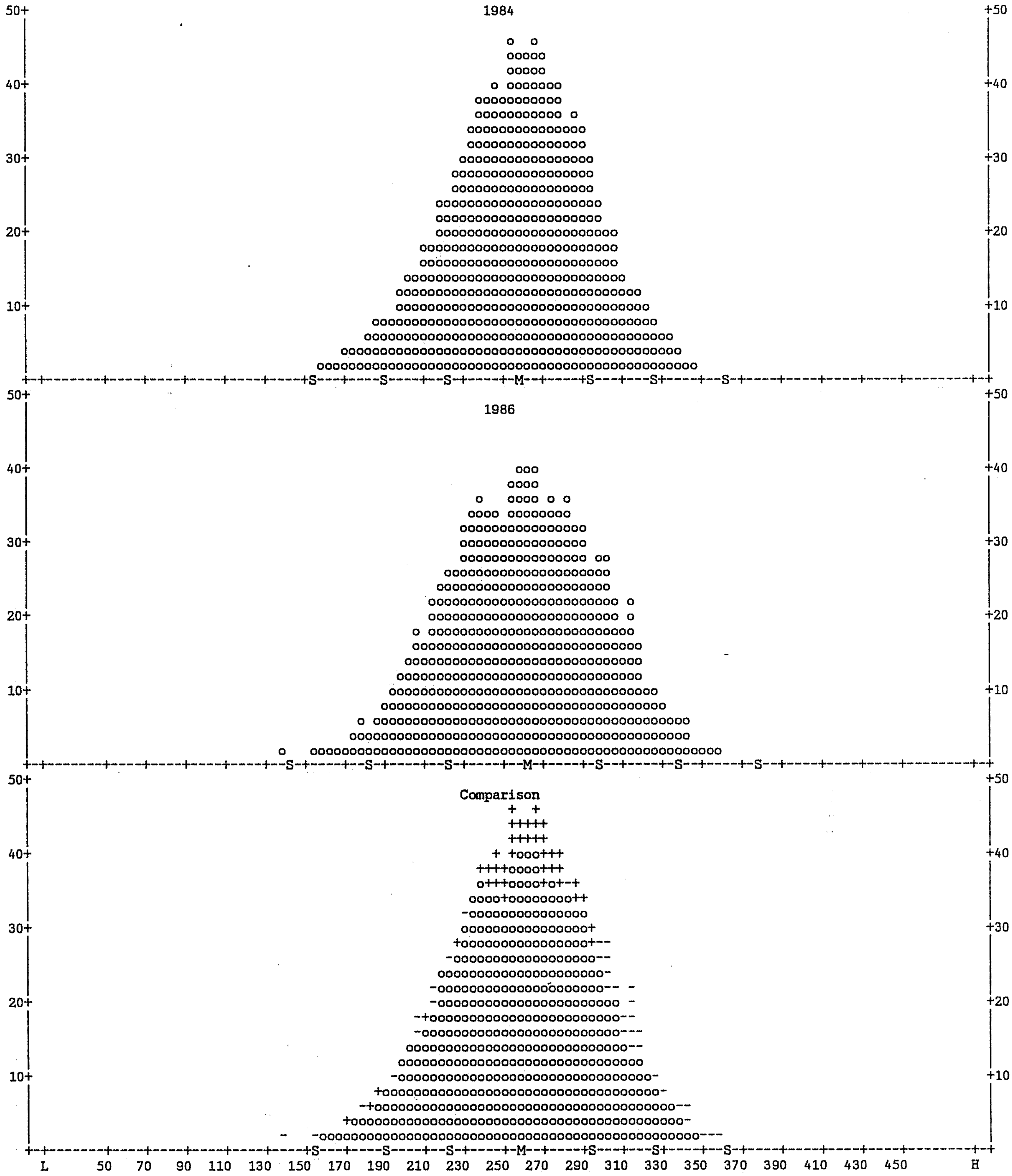


Figure 2-3
Comparison of 1984 and 1986 Reading Scores Using First Plausible Scale Value
Age 17, Weighted

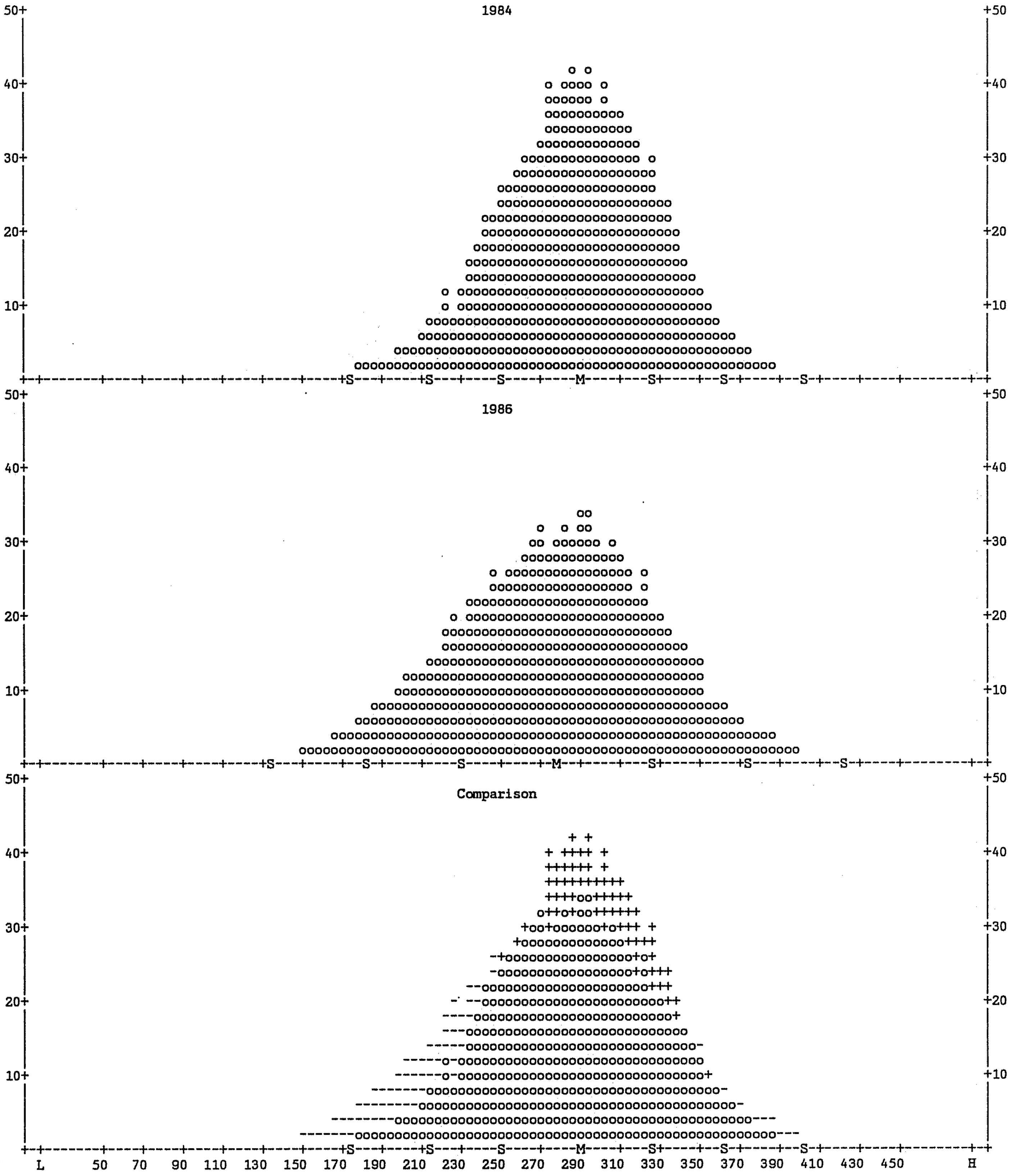


Table 2-2

Reading Proficiency Results
(Scale Values)

Sample Sizes

	1984	1986
Age 9	16799*	6932**
13	17535*	6200**
17	18984*	15976*

Average Reading Proficiencies

	1984	1986	Difference
Age 9	212.9	207.3	-5.6
13	258.0	260.4	+2.4
17	288.8	277.4	-11.4

Standard Deviations

	1984	1986
Age 9	40.1	44.1
13	36.1	40.0
17	39.6	49.4

*data from BIB sample (see Chapter 3)
**data from Bridge (pseudo-BIB) sample (see Chapter 3)

Standard errors are typically less than 1.1

Table 2-3

Mean Percent Correct for 1984/1986 Reading Trend Items

	Number of Items	1984 Mean P*	1986 Mean P*	Difference
Age 9	30	64.2	60.8	-3.4
13	24	66.7	66.8	+ .1
17	39	76.4	73.1	-3.3

Average Sample Sizes per Item

	1984	1986
Age 9	1958	2105
13	2183	1925
17	2343	4003

Differences Between Mean P's
(2 point trends)**

	1984-1980	1986-1984
Age 9	-1.0 [35 items]	-3.4 [30 items]
13	0.0 [43 items]	0.1 [24 items]
17	0.3 [35 items]	-3.3 [39 items]

* Mean P is the average of item-level percent correct where item percent correct = R/(R+W+IDK) with R, W, and IDK being the weighted frequencies of right, wrong, and I Don't Know responses to the item.

** based on total set of reading comprehension trend items

Chapter 3

OVERVIEW OF THE INVESTIGATIONS INTO THE READING ANOMALY

Albert E. Beaton

The purpose of this chapter is to present an overview of the investigations into the reading anomaly. The first part of the chapter contains a brief description of parts of the NAEP assessment design that are relevant to the decline in estimated reading performance. Next, the types of hypotheses that were investigated are presented and the major results summarized. Finally, the special samples included in the 1988 assessment to isolate the effects of assessment alterations are discussed.

The anomalous results in reading have been based on the three samples collected in 1986 for analyzing trends and on their comparisons with corresponding samples collected in 1984 and earlier. Most of the studies of the reading anomaly have focussed on the 1984 and 1986 samples, although a few have also used data from earlier assessments. It is useful, therefore, to consider the properties of the 1986 trend samples and to contrast those samples with the corresponding data collected in 1984.

It should be noted that not all of the data collected by NAEP in 1984 or 1986 are relevant to the estimated decline in reading performance. The overall designs of both the 1984 and 1986 samples were complicated because both were expected to introduce new assessment technology as well as allow the measurement of changes in performance from past assessments. For example, NAEP has historically assessed populations of students at certain ages (9, 13, and 17) but NAEP now collects data for selected grades as well. Since the students in the NAEP grades who were not at the appropriate ages are not relevant to the anomalous trends in reading performance, the NAEP data collected on grade-only eligible students will not be described. Only the samples appropriate for studying the anomaly will be described here.

The data used in these studies were collected from national probability samples of students aged 9, 13, and 17 in American schools, both public and private. The methods and curriculum areas of assessment for the 1984 and 1986 samples are shown in Table 3-1.

In both years, some form of BIB spiralling was used. BIB spiralling involves assigning different assessment booklets to different students within a particular assessment session. Every assessment booklet contains three blocks of items, which may be from one curriculum area (e.g., reading) or from several. The student is expected to read the instructions and respond to the assessment questions.

Table 3-1

Administrative Method and Curriculum Areas

Age	1984	1986
9	BIB* Spiralling Reading, Writing	Pseudo-BIB Spiralling Reading, Math, Science
13	BIB Spiralling Reading, Writing	Pseudo-BIB Spiralling Reading, Math, Science
17	BIB Spiralling Reading, Writing	BIB Spiralling Reading, Math, Science, Computer Competence

*Balanced Incomplete Block

The 1984 samples were assessed in reading and writing at all ages. Each student, therefore, received a booklet containing either all reading exercises, all writing exercises, or some combination of both.

In 1986, pseudo-BIB spiralling was used at ages 9 and 13. Under pseudo-BIB spiralling, students were given a booklet that contained one reading, one mathematics, and one science block. The mathematics and science blocks were administered using a tape recorder, with both the assessment instructions and exercises administered to the student aurally. The tape recorder was turned off when the reading block was administered.

The 17-year-old students in 1986 were assessed using full BIB spiralling, but the mix of curriculum areas was different from 1984. In 1986, reading was administered with mathematics, science, and computer competence (as well as a special probe of history and literature), not with writing, as it had been in 1984.

The sample sizes used in estimating performance are shown in Table 2-2 of Chapter 2. Other differences between assessment years are described in Chapter 7.

* * *

When the steep declines in reading proficiency at ages 9 and 17 were noticed, a number of hypotheses as to the reason for the decline were developed and studied. These hypotheses fall into eight general classifications, which are discussed in the following sections of this chapter.

Population/Sample Hypotheses

The NAEP population is very precisely defined and the sample carefully drawn. However, a sharp change in the population--such as an increase in the number of traditionally low-scoring students--would probably result in a decline in average proficiency. Also, we must be assured that the sample that was actually drawn is not unusual and is truly representative of the intended student population.

Detailed study shows no reason to believe that the NAEP sample is not representative of the nation's students. First, the NAEP sample produces the best available estimate of the numbers of in-school 9-, 13-, and 17-year-old students since its sampling weights are post-stratified using information from the Census Bureau, the Current Population Survey, and the NAEP samples themselves. In any case, there were no sharp changes in the NAEP estimates of population sizes since 1984.

Chapter 4 describes the NAEP sampling process. There is no substantial difference in the percentage of students excluded from NAEP because of limited English proficiency, behavioral disorders, or physical or mental handicap. We have not found any reason to believe that a substantial change in the dropout rate for 17-year-olds occurred.

Chapter 5 explains the attributes of low scorers to see if any discernible subgroup of students had an unusual increase in low scores. The proportion of low scorers increased in all major subgroups of students. Chapter 5 also examines whether the decline was concentrated in a few schools and concludes that it was not.

Chapter 6 investigates the effect of the slight changes found in population sizes. The evidence shows that the decline at age 17 was pervasive, occurring in all of the groups for which NAEP has traditionally reported results. In fact,

- both boys and girls declined, with the decline for boys substantially larger than that for girls;

- all racial and ethnic groups declined;

- all regions of the country declined, with the decline being least in the Northeast; and

- students whose parents did not graduate from high school, students whose parents did, and students whose parents had some education beyond high school all declined.

We have therefore concluded that neither population shifts nor the composition of the NAEP sample contributed to the reading anomaly.

Measuring Instrument Hypotheses

There were a number of seemingly minor differences in the measuring instruments used in NAEP between 1984 and 1986. These changes are documented in Chapter 7.

We have no reason to doubt the validity of the NAEP 1986 reading assessment since the reading results do behave as expected. The median correlation among the reading blocks and between the reading, mathematics, science, and computer competence blocks is shown in Table 3-2. The reading blocks contain fewer items than the blocks in other curriculum areas, and the estimated average reliability of individual reading items is estimated to be greater than in other curriculum areas.

Before the changes in the measuring instruments were made, they were judged by professionals to be so minor as not to affect the students' responses. For example, there was a change in the number of items in an assessment booklet, but there was also a corresponding change in the amount of time allocated to respond. We cannot now be sure that these minor changes did not have a major effect. We have, however, introduced some special studies into the 1988 assessment to measure the effect of changing the measuring instruments; this redesign is discussed in Chapter 13.

Administrative Changes Hypotheses

A number of changes were made in the design specifications and administrative procedures as discussed in Chapter 7. For example, the average number of students in an assessment session increased at age 17 from approximately 20 students in 1984 to approximately 35 students in 1986. Also, all three age groups were assessed in the spring of 1986, and a Language Minority Component was added.

In this report, we have examined some of these factors. In addition, based on field observations, Westat has reviewed administrative procedures used in 1984 and 1986 and has not found changes which are likely to have affected only reading for the 9- and 17-year-olds. However, we cannot be sure that seemingly minor changes in the design specifications and resulting procedures have not had an effect; thus, the redesigned 1988 assessment repeats both 1984 and 1986 assessment procedures.

The Westat report is shown in Appendix C.

Quality Control Hypotheses

A natural place to look for error is in the data processing to assure that the apparent decline was not caused by a processing blunder. NAEP data are already subject to strict quality control procedures (see Implementing the New Design: The 1983-84 NAEP Technical Report (Beaton, 1986)). To assure independently the accuracy of the data, we selected a copy of each

Table 3-2

Correlations Among NAEP Blocks
1986 Assessment, Age 17

			Reading	Math	Science	Computer Competence
			R	M	S	C
Reading	R	N	15	12	14	28
		median	.65	.60	.54	.38
		range	(.48 - .75)	(.46 - .65)	(.39 - .66)	(.19 - .57)
Math	M	N		55	28	14
		median		.74	.62	.52
		range		(.58 - .92)	(.48 - .80)	(.24 - .60)
Science	S	N			55	12
		median			.62	.51
		range			(.46 - .72)	(.22 - .63)
Computer Competence	C	N				15
		median				.57
		range				(.40 - .66)

booklet at random and assured that student responses in the assessment booklets were accurately recorded in the database. A study of the database, which is described in Chapter 8, shows the database to be very accurate. An external consultant, Dr. W. B. Schrader, also reviewed this process and found no basis for questioning the database or the scoring keys. His report is shown in Appendix B.

Computations of proportions passing various items have been done by several programs, and the results are in agreement.

We believe that errors in the database and computational errors can be ruled out as an explanation of the decline.

Scaling Hypotheses

NAEP uses a complex process to estimate the distribution of reading proficiency. We investigated whether or not the decline could be an artifact of the scaling process.

An approximate method has been developed for estimating average proficiency on the reading scale from the average percentage of items that the students answered correctly, without any scaling of the data. This method is described in Chapter 9. It shows that the decline in the average proportion answering items correctly is consistent with the decline in reading proficiency estimated from the scaling procedure.

We therefore rule out the scaling process as the cause of the decline in reading proficiency.

Item Level Hypotheses

At the individual item level, several particular questions were pursued. Were one or a few items so dramatically different that the decline is attributable to only a few items? Was there a change in the way that students responded to items? These hypotheses are studied in Chapter 10.

In summary, there was neither one nor a few items that changed enough to affect the entire results. In general, the 17-year-old students were less likely to respond correctly to an item, more likely to respond incorrectly or I Don't Know, and slightly less likely to omit or not reach items. These changes in the I Don't Know, omit, and not reached rates were found to contribute little to the decline. The decline does not, therefore, seem to be associated with any individual items or changes in response patterns.

Booklet and Block Hypotheses

A student might respond differently to a reading exercise when the exercise is placed in a booklet with mathematics or science exercises, so the effect of changing the context of items was studied. This study is reported in Chapter 11.

The study of the context of reading blocks shows little effect on the reading anomaly. One reading block seems a little out of line with the others when placed in the context of non-reading blocks, but even when the booklets containing this mix of blocks are removed, the sharp decline in estimated reading proficiency remains.

Other Hypotheses

Other hypotheses have also been explored. For example:

- the external event hypothesis. We looked for some event in 1985-86 that might have affected the way the students responded to NAEP. We found one--the Challenger disaster--which occurred during the last week of the assessment of the 9-year-olds. The study of this hypothesis is presented in Chapter 12.

  The data for 9-year-olds was separated by day of assessment and reviewed for any large increase in the number of low scorers immediately after the Challenger disaster occurred.

  No substantial change in the proportion of low scorers was discerned.

- the hypothesis that the 1984 assessment results were unusually high. This hypothesis was investigated by performing comparisons of 1986 with earlier years. Although the 1984 average reading performance was higher at age 17 than in previous years, the decline in 1986 would still be substantial even if compared to the results of the earlier assessments.

* * *

Although a number of possible explanations for the estimated decline in reading proficiency have been eliminated, several remain. It is still possible that the seemingly minor changes in the assessment booklets or in administrative procedures may have had a sufficient effect on the responses of students to produce such anomalous results. We have, therefore, modified the 1988 sample to gather data that will lead to a clarification of these issues.

Included in the 1988 assessment at each age level will be three randomly equivalent samples of students:

- a sample using the same assessment booklets and administrative procedures as in the 1984 NAEP.

- a sample using the same assessment booklets and administrative procedures as in the 1986 NAEP.

- a sample using the new booklets and procedures intended for the 1988 assessment.

In other words, we intend to reproduce the assessment conditions of the past two assessments as closely as possible.

Reproducing the conditions of the 1984 and 1986 assessments on randomly equivalent samples from the same population makes it possible to separate the effects of changing methodology from changes in the distribution of reading proficiency. The results that will be compared are discussed in Chapter 13. The comparisons may show that the apparent decline is real, that it is artifactual and due to changes in assessment design, or by some combination of real decline and methodological change. In any case, this study should allow professionally responsible release of the 1986 reading data.

Chapter 4

THE NAEP POPULATIONS AND SAMPLES

Eugene G. Johnson

In order to increase the power of NAEP data as well as to increase statistical and administrative efficiency, the NAEP sampling design for the 1986 assessment was improved in a number of ways over the design used in the 1984 assessment. It is thus appropriate to investigate whether or not some characteristic of the new sample design might have had some effect on the observed decline in the NAEP reading results. In particular, it is important to know whether or not the 1986 sampling and weighting procedures resulted in a realized sample in which the lower performing students are overrepresented.

As in 1984, the target population in 1986 consisted of 9-year-olds, 13-year-olds, and 17-year-olds enrolled in public and private elementary and secondary schools, along with other students in the modal grade for each of these three ages. As in 1984, the 1986 sample was a multistage probability sample consisting of four stages of selection. The first stage of selection, the primary sampling units (PSUs), consisted of counties or groups of counties. The second stage of selection consisted of elementary and secondary schools. The assignment of sessions to sampled schools comprised the third stage of sampling, and the fourth stage involved the selection of students within schools and their assignment to sessions.

There were a number of changes in the sampling design between 1984 and 1986. A major change was a shift to assessment in the spring for each age class, coupled with a change in the age definitions for the 9-and 13-year-old students. In previous assessments, 13-year-olds were defined on a calendar year basis and were assessed in the fall; 9-year-olds were also defined on a calendar year basis and were assessed in the winter; and 17-year-olds were defined on an October-through-September basis and assessed in the spring. For the main assessment in 1986, the students of each of the three age classes were defined as the students born between October 1 and September 30 of the appropriate years preceding the assessment. All ages were assessed in the spring for the main assessment of 1986. This change in the time of assessment and age definition necessitated the use of bridge samples for the 9- and 13-year-old students. These bridge samples were assessed at the same time of year as in 1984 and used the old age definitions. Because they provide data comparable to earlier assessments, the 1986 bridge samples were used for all analyses of trends in reading proficiencies for the 9- and 13-year-old students. The data from the 1986 main assessment of 17-year-old students were used for the measurement of trends in reading achievement, since those data are based on a sample with the same age definition and time of assessment used in previous assessments.

Consequently, the measurement of the 1986 trend point in reading was based on samples from populations comparable to those from previous assessments. The issue in determining if the sampling and weighting procedures are, at least in part, responsible for the reading decline lies in determining if these samples were properly representative of the target populations. To judge how representative the 1986 samples are, it is helpful to consider how the 1986 sampling design changed from that of 1984.

The main changes in the design for 1986 are as follows:

1) The definition of PSUs and their stratification differed from previous assessments. Generally, whole metropolitan statistical areas (MSAs) were defined as PSUs, rather than the individual counties that were used in the previous definitions. The PSUs in 1986 were stratified by region, size (MSA or non-MSA), and percent minority. In 1984, the PSUs were stratified by region and size of community only. The number of PSUs selected for the 1986 sample was increased to 94 from the 64 used in 1984.

2) A subsample of 64 PSUs, selected from the full sample of 94, was used for the bridge assessments.

3) Throughout the selection process in 1986, Black students and Hispanic students were oversampled at about twice the rate of other subgroups. This differed from 1984, when extreme-low socioeconomic status big city schools and extreme-rural schools were oversampled.

4) To accommodate the Language Minority Probe, supplemental samples were drawn of students in the schools sampled for the regular NAEP. (Since these supplemental samples were drawn only in the main assessment, and not in the bridge assessments, only the age 17 trend data could have been affected by this probe.)

5) Except for the larger schools, all eligible students were assessed, resulting in a moderately larger number of students assessed per school than in previous assessments. The assessment session sizes for the main assessment of students of age 17 or in grade 11 in 1986 were somewhat larger than those in 1984, averaging 35.2 students in 1986 versus 20.3 students in 1984. The assessment session sizes for the 1986 bridge samples for ages 9 and 13 were smaller (less than 25 students) and close to the session sizes obtained in 1984.

These changes in the sampling design were made to increase efficiency, to improve the precision of estimates of achievement for Black and Hispanic students, to accommodate the Language Minority Probe, and, in the case of the increased session sizes, to accommodate the desire of the schools to minimize disruption.

Turning first to the effect of change in session size on achievement, we considered the percent-correct statistics from the 1986 age 17 assessment for the items in one of the reading blocks (block R4), comparing these p-values for students in large sessions (51 or more students), medium sessions (26-50 students), and small sessions (1-25 students). The tendency was for the p-values to be highest for the large sessions, lowest for the medium sessions, and intermediate for the small sessions (but close to those for the large sessions). The average of the p-values were 70.2 for large sessions, 61.4 for medium sessions, and 68.0 for small sessions. These results appear contrary to the hypothesis that the large sessions would have the lowest results because of potentially chaotic conditions. Since the highest results were obtained by students in large sessions, and since these comprise 86 percent of the sample, we conclude that the larger session sizes in 1986 could not have contributed to the decline in reading scores.

Because of the design changes mentioned previously, the resulting samples in 1986 are somewhat different from those in 1984, but, if the sampling and subsequent weighting procedures were carried out properly, the 1986 samples would be, after weighting, representative of the 1986 target populations. A series of checks were made at the time of sample selection to verify that the sample was representative of the target population. One check was made after the selection of the school sample by inflating the estimated number of eligible students in each school by the inverse of the school selection probability and then checking the results against total school enrollment in the frame from which the sample was drawn. Other checks were made when students were sampled within schools, by verifying at the school and PSU level that the yield of students obtained was as anticipated. Careful checks were made of the sampling procedures in any case where a discrepancy was noted. The result of these checks was that any discrepancies that remained were the result of inaccuracies in the measure of size of the school, the effect of which was appropriately accounted for in the course of the weighting procedures. Since similar steps were taken to verify the integrity of the sample in 1984, there appears little possibility that errors in the sample selection procedures could have caused the changes in reading performance.

The weighting procedures for 1986 included computing the student's base weight, the reciprocal of the probability that the student was invited to a particular session. These base weights were adjusted for nonresponse, then subjected to a trimming algorithm to reduce a few excessively large weights. The weights were further adjusted by a post-stratification procedure to reduce the sampling error of estimates relating to student populations that correspond to several subgroups of the total population. Post-stratification was performed by adjusting the weights of the sampled students so that the resulting estimates of the total number of students in a number of specified subgroups of the population corresponded to independently determined population totals based on information from the Current Population Survey, the 1980 Census, and from NAEP. The subpopulations were defined in terms of race, ethnicity, census region, and sampling descriptor of community (based on the size and degree of urbanization of a county). The weighting procedures used in 1986 are essentially the same as those used in 1984. A check of the weighting procedures from beginning to end did not reveal any errors. The process and programs used to perform each stage of weighting were checked, and found to have performed as specified. Checks on sums of weights showed that they were consistent with the independent estimates of total enrollment.

For the decline in reading performance to be caused by an error in the sampling or weighting procedures, it is necessary for a group of students with low performance to be overrepresented in the final weighted sample, relative to the more able students. This overrepresentation would have to be substantial in order to contribute in a meaningful way to the drop in estimated reading performance. If incorrect weighting had occurred of the magnitude necessary to explain the drop, then the checks on the weighting procedures and particularly on the sums of the weights should have revealed some discrepancy--they did not.

Tables 4-1, 4-2, and 4-3 examine whether there has been a shift in the sample in 1986 in the direction of lower-performing students. Table 4-1 addresses the assessed 17-year-olds in 1984 and 1986 and shows mean percent correct statistics (across all 39 1984/1986 change items), by year, within each category of the main NAEP reporting groups. The table also shows the sampled frequency and the weighted frequency for each of the reporting groups, as well as the changes between 1984 and 1986 in the mean percent correct, the sampled frequency, and the weighted frequency. Table 4-2 provides the same information for the 9-year-olds, and is based on the 30 change items. Table 4-3, based on 24 change items, is for the 13-year-old assessed students.

We can see from Table 4-1 that reading performance declined within every demographic subgroup considered. The general pattern of decline is that the traditionally lower performing subgroups of the population (e.g., students below modal grade, students in disadvantaged urban environments) had a greater decline than the traditionally higher performing subgroups. Table 4-1 also shows that, if there has been any shift in the 1986 sample relative to the 1984 sample, it has been toward students in the modal grade and towards students with more highly educated parents. (The increase in the number of sampled Black and Hispanic students is to be expected since they were oversampled--this increase is more than corrected by the weighting.) It appears from Table 4-1 that, if there has been a shift in the population between 1984 and 1986, then it has tended to be away from the lower performing subgroups identified here.

Although the decline in scores for Black students and (especially) Hispanic students is greater than for the population as a whole, Table 4-1 shows that the weighted proportions of those subgroups has declined slightly between 1984 and 1986. This means that the decline in reading performance of the population as a whole cannot be due to an overrepresentation of Black and Hispanic students in 1986. (The fact that White students also declined also supports this contention.) Furthermore, since the weights for the ethnic groups are adjusted in both years to independently determined population totals, it is not likely that incorrect weighting could have led to incorrect representation of those groups in the population in either year.

The results for the age 17 male and female students deserve comment. The decline in reading performance for male students is noticeably more

Table 4-1

Changes in Performance, Sampled Frequency, and Weighted Frequency

Age 17 Reading

	Mean P*			Sampled Frequency			Weighted Frequency		
					%			%	
	1984	1986	Change	1984	1986	Change	1984	1986	Change
Total	76.4	73.1	-3.3	---	---	---	---	---	---
Male	74.2	70.1	-4.1	49.8	49.8	0.0	50.9	50.6	-.3
Female	78.6	76.1	-2.5	50.2	50.2	0.0	49.1	49.4	.3
White	78.8	75.8	-3.0	73.2	70.4	-2.8	75.9	76.7	.8
Black	67.4	63.8	-3.6	14.4	16.6	2.2	14.0	13.1	-.9
Hispanic	69.4	62.5	-6.9	9.2	10.3	1.1	7.6	7.3	-.3
Other	75.4	67.6	-7.8	3.2	2.7	-.5	2.5	2.9	.4
<modal grade	63.5	58.0	-5.5	17.5	17.3	-.2	21.5	21.5	0.0
Modal grade	79.5	76.6	-2.9	74.2	75.8	1.6	68.3	70.4	2.1
>modal grade	82.9	80.7	-2.2	8.3	7.0	-1.3	10.2	8.0	-2.2
NE	77.4	75.8	-1.6	22.5	19.9	-2.6	24.4	23.6	-.8
SE	75.4	70.0	-5.4	25.5	25.5	0.0	22.6	21.3	-1.3
Central	76.1	74.2	-1.9	28.2	26.0	-2.2	27.2	28.9	1.7
West	76.5	71.9	-4.6	23.8	28.6	4.8	25.8	26.2	.4
Rural	74.2	70.8	-3.4	6.4	4.3	-2.1	5.4	4.4	-1.0
Low metro	67.3	59.8	-7.5	10.0	7.2	-2.8	9.3	5.9	-3.4
High metro	81.3	80.7	-.6	13.1	11.5	-1.6	16.1	12.3	-3.8
Parental Education									
<HS	69.3	63.1	-6.2	12.1	9.2	-2.9	11.7	8.7	-3.0
Grad HS	74.0	68.7	-5.3	35.3	28.3	-7.0	34.5	27.7	-6.8
HS+	79.6	75.2	-4.4	14.7	21.9	7.2	14.6	21.5	6.9
Grad college	81.1	79.1	-2.0	33.5	36.7	3.2	35.0	38.2	3.2

Number of items = 39

Average number of students per item in total sample: 1984 = 2343; 1986 = 4003

*Mean P is the average of item-level percent correct where item percent correct = R/(R+W+IDK) with R, W, and IDK being the weighted frequencies of right, wrong, and I Don't Know responses to the item.

Table 4-2

Changes in Performance, Sampled Frequency,
and Weighted Frequency

Age 9 Reading

		Mean P*			Sampled Frequency %			Weighted Frequency %	
	1984	1986	Change	1984	1986	Change	1984	1986	Change
Total	64.2	60.8	-3.4	---	---	---	---	---	---
Male	61.9	58.5	-3.4	50.5	50.0	-.5	49.6	49.5	-.1
Female	66.4	63.1	-3.3	49.5	50.5	.5	50.4	50.5	.1
White	68.2	65.0	-3.2	67.8	66.1	-1.7	71.9	72.9	1.0
Black	51.5	48.0	-3.5	14.1	11.9	-2.2	13.8	13.5	-.3
Hispanic	53.7	48.1	-5.6	13.5	14.5	1.0	11.2	10.0	-1.2
Other	64.4	59.9	-4.5	4.6	7.6	3.0	3.0	3.6	.6
<modal grade	47.6	47.2	-.4	29.7	32.9	3.2	23.6	32.5	8.9
Modal grade	69.2	67.2	-2.0	69.9	66.7	-3.2	76.1	67.2	-8.9
>modal grade	83.5	79.7	-3.8	.4	.3	-.1	.3	.3	0.0
NE	65.6	63.0	-2.6	22.5	24.8	2.3	22.2	21.3	-.9
SE	61.9	57.2	-4.7	24.8	22.7	-2.1	23.5	22.5	-1.0
Central	67.0	63.2	-3.8	28.8	24.7	-4.1	27.2	28.7	1.5
West	62.1	59.6	-2.5	23.9	27.8	3.9	27.2	27.6	.4
Rural	60.1	59.9	-.2	6.3	3.6	-2.7	6.3	4.6	-1.7
Low metro	53.4	45.0	-8.4	12.6	7.1	-5.5	11.8	5.7	-6.1
High metro	74.1	70.2	-3.9	11.9	15.5	3.6	14.2	17.3	3.1
Parental Education									
<HS	53.6	48.5	-5.1	5.8	4.2	-1.6	5.3	4.2	-1.1
Grad HS	63.5	56.7	-6.8	18.9	15.3	-3.6	19.0	16.3	-2.7
HS+	64.4	67.4	3.0	5.0	6.9	1.9	5.0	6.8	1.8
Grad college	71.6	67.8	-3.8	32.3	39.0	6.7	33.2	38.5	5.3

Number of items = 30

Average number of students per item in total sample: 1984 = 1958; 1986 = 2105

*Mean P is the average of item-level percent correct where item percent correct = R/(R+W+IDK) with R, W, and IDK being the weighted frequencies of right, wrong, and I Don't Know responses to the item.

Table 4-3

Changes in Performance, Sampled Frequency, and Weighted Frequency

Age 13 Reading

	Mean P*			Sampled Frequency %			Weighted Frequency %		
	1984	1986	Change	1984	1986	Change	1984	1986	Change
Total	66.7	66.8	.1	---	---	---	---	---	---
Male	64.4	65.4	1.0	50.0	49.3	-.7	50.4	49.1	-1.3
Female	69.1	68.2	-.9	50.0	50.7	.7	49.6	50.9	1.3
White	69.3	69.4	.1	72.3	61.0	-11.3	75.1	73.7	-1.4
Black	58.0	60.5	2.5	13.3	21.9	8.6	13.5	13.8	.3
Hispanic	57.3	55.3	-2.0	10.8	13.5	2.7	8.4	8.9	.5
Other	67.1	67.5	.4	3.6	3.5	-.1	3.1	3.5	.4
<modal grade	57.3	58.4	1.1	30.6	31.2	.6	30.8	32.0	1.2
Modal grade	70.9	70.8	-.1	69.1	68.3	-.8	68.7	67.6	-1.1
>modal grade	76.2	75.0	-1.2	.3	.6	.3	.5	.4	-.1
NE	68.0	68.7	.7	22.4	24.6	2.2	22.9	22.8	-.1
SE	66.8	66.7	-.1	24.7	22.5	-2.2	23.1	24.2	1.1
Central	66.2	64.5	-1.7	28.8	23.4	-5.4	27.0	24.9	-2.1
West	66.2	67.5	1.3	24.1	29.5	5.4	26.9	28.1	1.2
Rural	65.2	65.4	.2	5.7	3.9	-1.8	5.3	6.1	.8
Low metro	59.3	55.6	-3.7	10.0	12.7	2.7	8.6	8.2	-.4
High metro	74.3	73.7	-.6	11.1	11.5	.4	10.9	11.8	.9
Parental Education									
<HS	58.2	60.2	2.0	8.6	8.0	-.6	8.3	7.7	-.6
Grad HS	64.5	64.1	-.4	35.4	28.4	-7.0	35.1	30.0	-5.1
HS+	72.1	70.1	-2.0	9.9	15.2	5.3	9.9	16.0	6.1
Grad college	71.9	71.3	-.6	35.9	38.0	2.10	36.6	37.5	.9

Number of items = 24

Average number of students per item in total sample: 1984 = 2183; 1986 = 1924

*Mean P is the average of item-level percent correct where item percent correct = R/(R+W+IDK) with R, W, and IDK being the weighted frequencies of right, wrong, and I Don't Know responses to the item.

pronounced than that for female students. Yet the sampling and weighting procedures applied to both groups is the same, so that it appears very unlikely that a subgroup of low-performing male students could have been overselected and/or overweighted without the corresponding subgroup of low-performing female students being similarly affected.

Table 4-2 shows the changes in performance, sampled frequency, and weighted frequency for the age 9 students assessed in 1984 and 1986. We see again that the decline in performance (based on 30 items) is quite pervasive with declines in all but one subgroup (students with a parent who had more than a high school education). The impression here is that any shift in the sample in 1986 from 1984 is toward the higher performing subgroups, although there is a notable increase in the relative proportion of students below the modal grade.

This increase in the below-modal-grade students may account for some of the reading decline for the 9-year-old students, because if we recompute the mean P for the population as a whole using the 1986 subgroup mean P values for the grade variable but using the 1984 weighted frequencies, we get a new mean P of 62.5, cutting the drop in performance for the population as a whole in half from 3.4 to 1.7 percent points. However, this does not explain the decrease in mean P-values for students at or above the modal grade, and so the reading decline for 9-year-olds must be, at least in part, due to other causes.

Turning to Table 4-3, which addresses the changes for the 13-year-old students, we see that overall, there is no meaningful change in performance between the two years across the 24 common items. This age generally shows the same shifts in composition as ages 9 and 17.

Returning to the age 17 data, we address the issue of whether there was some error in the weighting process that tended to overweight the lower-performing students--regardless of their demographic characteristics. Figures 4-1 and 4-2 show plots, for years 1984 and 1986 respectively, of each student's LOGIST (maximum likelihood) reading score by that student's weight. It is apparent from the plots that there is little relationship between a student's weight and that student's reading score. (The correlations between reading score and weight are very low: .022 for 1984 and .017 for 1986).

We turn now to a number of characteristics of the realized 1986 sample that differed from 1984.

Time of Data Collection

Although the bridge assessments for ages 9 and 13 and the main assessment for age 17 were conducted at the same times of year in 1986 as the corresponding age class assessments in 1984, the time periods in which the assessments were conducted changed.

Figure 4-1

Plot of LOGIST Reading Proficiency Score Versus Sampling Weights for 17-Year-Old Students in 1984

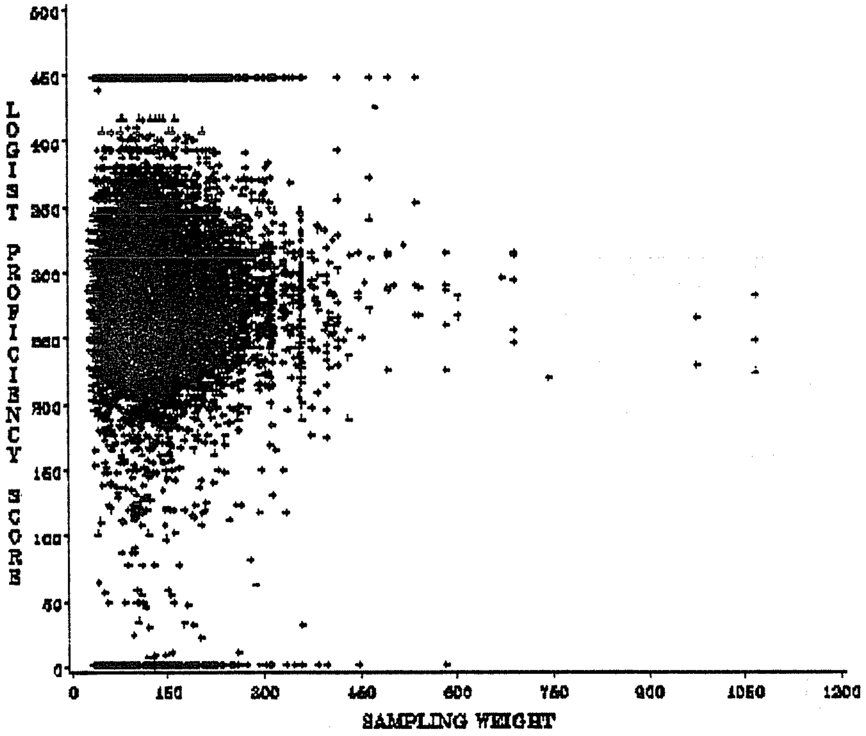
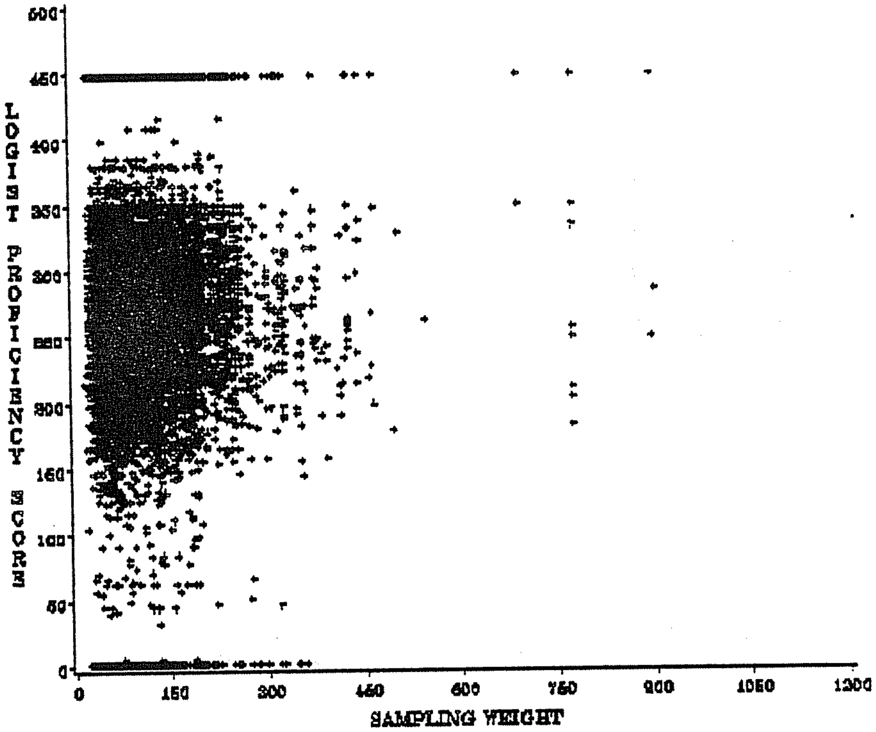


Figure 4-2

Plot of LOGIST Reading Proficiency Score Versus Sampling Weights for 17-Year-Old Students in 1986



The time periods for both years were as follows:

	1984	1986
Age 9	Jan. 2 - Mar. 19, 1984	Jan. 6 - Jan. 31, 1986
Age 13	Oct. 10 - Dec. 17, 1983	Nov. 4 - Dec. 13, 1985
Age 17	Mar. 12 - May 11, 1984	Feb. 17 - May 2, 1986

The result of these changes in time periods of assessment was that 9-year-olds and 17-year-olds tended to be assessed earlier and 13-year-olds later in 1986 than in 1984. The practical consequences of these changes is in the slight change in the average age of the students assessed between the two years. A direct way to ascertain the impact of the changes in the time periods of assessment would be to consider the parts of the 1986 and 1984 samples that were assessed over the same range of dates in both years (e.g., Jan. 6 - Jan. 31 for age 9; Nov. 4 - Dec. 13 for age 13; Mar. 12 - May 2 for age 17). If all else were equal, comparison of these subsamples would remove the effect of time changes. However, these subsamples will not necessarily be representative samples of the target populations because certain areas of the country may be assessed earlier in the assessment period than others. An indirect measure of the effect of change in time periods of assessment can be made as follows. Since the 9-and 13-year-olds were defined on a calendar year basis while the 17-year-olds were defined on an October 1 through September 30 basis, we can estimate the average age of an assessed student by considering the range of testing dates and eligible birthdates and assuming a uniform distribution of birthdates throughout the year. The results are as follows:

Average Age (in Years) of an Assessed Student

	1984	1986	Change
Age 9	9.61	9.55	-22 days
Age 13	13.37	13.38	+ 4 days
Age 17	17.03	16.98	-18 days

Consequently, the assessed 9- and 17-year-olds were, on average, slightly younger (by less than one month) in 1986 than in 1984 while the 13-year-olds were of about the same average age. Linear interpolation from the 1984 results suggest that these changes in average assessed age of 9- and 17-year-olds could only account for at most 1 scale score point.

The possibility was raised that, since the assessment of 17-year-olds began earlier in the spring in 1986 than it did in 1984, the earlier month (February) of the sampling was including students who would later drop out and so would not have been available for assessment under the old time frame. While we do not have information about the subsequent dropout status of the assessed students, we can make certain observations concerning the possibility of inclusion of excess future dropouts in the sample. First, this hypothesized inclusion will not explain the decline in performance of students of more highly educated parents nor the fact that such students are more heavily represented in the 1986 sample than in the 1984 sample, if we can assume that the propensity to drop out is negatively related to the educational level of the parents of the student. Second, the 1986 weighted estimates contain relatively fewer students from the disadvantaged inner city and relatively fewer students whose parents did not complete high school than in 1984. These subgroups of students should have a higher dropout rate and so it might have been expected, had the 1986 sample included more dropouts, that these subgroups would have had a higher representation in 1986. Third, it can be seen from Table 4-4, which compares the weighted and unweighted distributions of self-reported grades in school for the two years, that, if anything, there are relatively fewer students with the lower grades (C, D, or less). Although not all dropouts receive low grades, it is certainly true that a substantial number of drop outs have exhibited low academic performance. It is the inclusion of excessive numbers of these low-performing dropouts that would affect the reading results. Table 4-4 does not indicate such an excessive inclusion.

Table 4-4

Self-Reported Grades for 17-Year-Old Students by Year

	Unweighted Frequencies %			Weighted Frequencies %		
	1984	1986	Difference	1984	1986	Difference
A	15.3	26.7	11.4	15.1	27.1	12.0
B	40.4	41.1	0.7	39.7	41.1	1.4
C	38.3	29.7	-8.6	38.7	29.2	-9.5
D or less	6.1	2.6	-3.5	6.5	2.7	-3.8

These three observations, taken together, would appear counter to the inclusion of more (low-performing) dropouts into the sample.

Response Rates

Student response rates for 1986 were a few points lower for the 17-year-olds than they were in 1984, but increased for the 9- and 13-year-olds:

Student Participation Rate

	Age 9	Age 13	Age 17
1984	91.3%	87.3%	82.8%
1986	92.9%	89.2%	78.9%

This increase in nonresponse for age 17 is too small to explain significant changes in the reading scores. Besides, the change is in the wrong direction: Our experience has been that the nonrespondents--students who fail to attend sessions--tend to be poorer performers than the students who did attend the assessment session (Rogers, Folsom, Kalsbeek, & Clemmer, 1977). The adjustments for nonresponse tend to compensate for such differences because the adjustments are made separately, within primary sampling units, for students in the modal grade for their age, and for students below the modal grade for their age.

School cooperation rates for 1984 and 1986, by age class, were as follows:

School Cooperation Rate

	Age 9	Age 13	Age 17
1984	88.6%	90.3%	83.9%
1986	88.2%	88.0%	82.3%

The nonresponse adjustments for each age class were made separately by class of school within PSUs, using one to three classes of schools within an age class within a PSU. School nonresponse increased moderately in 1986 over that for 1984 for age classes 13 and 17, and trivially for age class 9. The difference between the response rates for age 17 seems too small to account for a measurable change in reading scores (up or down), provided we assume that the nonresponding schools within a class of school within a PSU are roughly similar in the two years. Of course, any nonresponse is undesirable, and similar and extensive efforts were made in each of the two years to achieve the highest feasible level of school cooperation.

Excluded Students

As in 1984, in 1986 students who were functionally handicapped to the extent that they could not participate in the assessment were excluded. Excluded students included students with limited English proficiency, students with behavioral disorders, and students who are physically or mentally handicapped, including those classified as Educable Mentally Retarded. The exclusion rates for the two years are:

Exclusion Rates

	Age 9	Age 13	Age 17
1984	3.8%	3.6%	3.0%
1986	3.9%	3.7%	3.4%

Consequently, relatively more students were excluded in 1986 than in 1984, a result that should lead to a rise in performance if one believes that the excludable students would be poorer performers.

* * *

In summary, it appears unlikely that the sampling and weighting procedures can be held responsible for much of the observed decline in reading performance.

Chapter 5

ATTRIBUTES OF LOW-SCORING STUDENTS

Eugene G. Johnson

This chapter examines the attributes of low-scoring students in the 1984 and 1986 assessments with the aim of determining if there are any discernible subgroups of students that showed a notable increase in the proportion of low scorers between the two years. We begin by considering demographic subgroups of the population in this search for groups of students who were more likely to be low scorers in 1986 than in 1984. Following the demographic analysis, we consider the possibility that the excess number of low scorers in 1986 is concentrated within a few schools.

Demographic Characteristics of Low Scorers

A comparison of the frequency distribution of the reading proficiency scores for 17-year-old students assessed in 1986 with the frequency distribution of reading proficiency scores for 17-year-old students assessed in 1984 (shown in Figure 2-3) revealed that the shape of the distribution of scores changed between the two assessments. In particular, the 1986 distribution had a higher concentration of lower scores than did the 1984 distribution, while the upper tails of both distributions were quite similar. The medians and quartiles of the two distributions, based on unweighted counts, and using LOGIST maximum likelihood proficiency scores, are as follows:

	Lower-quartile	Median	Upper-quartile
1984	263	288	316
1986	244	279	319

(Similar results hold for the reading proficiency plausible values). In fact, the distribution of proficiencies in 1986 appears to have a minor mode (or shoulder) at around 200 that is not present in the distribution of scores for 1984. The weighted and unweighted proportions of assessed 17-year-olds with LOGIST scores below this value of 200 differs considerably between the two assessment years:

	Total Sample Size	Unweighted % Below 200	Weighted % Below 200
1984	18879	3.96	4.11
1986	16267	9.58	9.69

Consequently, the 1986 assessment of reading produced more than twice as many "low scorers" (students with scores less than 200) as the 1984 assessment. It is of evident interest to know if this increase in the relative proportion of low scorers from 1984 to 1986 is across the board or if certain types of students are more likely to be low scorers than others.

Table 5-1 addresses this question. The first column of the table pertains to the 1984 assessment and shows the weighted percent below the threshold of 200 for the total sample as a whole as well as for students within each of a variety of demographic subgroups. The second column of the table imparts the same information for the sample of students from the 1986 assessment. Considering the first two columns, we see that, as expected, the traditionally lower-performing subgroups of the population tend to have more of their members in the below-200 group than do the higher-performing subgroups. Thus, for example, 8.7 percent of students in the Low Metro (disadvantaged-urban) STOC (size and type of community) category had scores less than 200 in 1984, but only 3.2 percent of the students in the High Metro (advantaged-urban) STOC category had scores below 200 in 1984.

Of more interest is the third column of the table, which shows the ratio of the 1986 proportion to the 1984 proportion. On looking through the table, it can be seen that the increase in the relative proportion of low scorers is fairly pervasive, since the 1986 proportion is larger than the 1984 proportion in all but one case (non-Hispanic students who speak a language other than English in the home).

Since nearly every subgroup has proportionately more members in the low-scoring group in 1986 than in 1984, we turn to considering which demographic subgroups have relatively higher proportions in the under 200 group, and which have relatively lower proportions, than the total sample. The groups with ratios greater than the overall ratio of 2.4 are those groups whose representation in the low-scoring group has increased, on a weighted basis, between 1984 and 1986 more than the total sample; those with ratios less than 2.4 are relatively less highly represented in the low-scoring group than the total sample. The results are as follows:

1) The increase for both male and female students in the low scorers group matches that of the full sample.

2) The increase for minority students is relatively less and for white students is relatively more than that of the full sample.

3) The increase for the High Metropolitan (Advantaged-urban) subgroup is relatively less than the full sample; that of the Low Metropolitan subgroup matches the sample; and that of the remainder of the STOC categories is greater than the full sample.

4) The Northeast has increased the least and the Southeast the most.

Table 5-1

Weighted Percent of Age 17 Students with LOGIST Scores Below 200 by Year and Demographic Subgroups

	1984	1986	Ratio
TOTAL	4.1%	9.7%	2.4
SEX:			
MALE	5.7%	13.2%	2.3
FEMALE	2.5%	6.1%	2.4
ETHNICITY:			
BLACK	7.4%	15.5%	2.1
HISPANIC	9.8%	19.1%	1.9
ASIAN	5.4%	7.5%	1.4
WHITE + OTHER	2.9%	7.8%	2.7
STOC:			
LOW METRO	8.7%	20.0%	2.3
HI METRO	3.2%	4.4%	1.4
OTHER	3.6%	9.7%	2.7
REGION:			
NE	4.5%	7.5%	1.7
SE	4.1%	12.5%	3.0
CENTRAL	3.3%	8.6%	2.6
WEST	4.6%	10.5%	2.3
PARENTAL EDUCATION			
UNKNOWN	10.3%	24.8%	2.4
<HS	5.7%	15.6%	2.7
HS GRAD	4.9%	12.1%	2.5
>HS	2.6%	7.7%	3.0
COL GRAD	2.6%	6.1%	2.3
ITEMS IN HOME			
0-3 of 5	10.0%	19.0%	1.9
4 of 5	4.7%	10.4%	2.2
5 of 5	2.6%	7.1%	2.7

Table 5-1
(continued)

	1984	1986	Ratio
HOMEWORK			
NONE GIVEN	6.1%	20.4%	3.3
DIDN'T DO	5.9%	18.4%	3.1
1/2 HR	3.3%	9.7%	2.9
1 HR	2.4%	7.8%	3.3
2 + HRS	2.6%	6.5%	2.4
LANGUAGE MINORITY			
YES HISP	15.2%	18.7%	1.2
YES OTHER	10.7%	7.5%	.7
NO	3.7%	9.6%	2.6
PERCENT WHITE IN SCHOOL			
0-49%	4.8%	14.5%	3.0
50-79%	4.4%	10.9%	2.5
80-100%	3.6%	7.9%	2.2
GRADE			
<MODAL	10.0%	21.4%	2.1
MODAL	2.6%	6.6%	2.5
>MODAL	1.2%	5.2%	4.3
SCHOOL TYPE			
PRIVATE	1.8%	3.1%	1.7
PUBLIC	4.5%	10.3%	2.3
GRADES			
≤D	11.4%	27.4%	2.4
C	5.6%	16.3%	2.9
B	2.3%	8.1%	3.5
A	1.7%	3.2%	1.9
TV WATCHING			
≤2 HRS	3.1%	7.8%	2.5
3-4 HRS	5.0%	9.0%	1.8
≥5 HRS	6.4%	15.9%	2.5
%SCHOOL LUNCH			
0-10%	2.9%	7.6%	2.6
>10%	4.7%	11.3%	2.4

5) There is no discernible ordering for the levels of parental education with the increase being greatest for students with parents with greater than a high school education.

6) The representation in the low-scoring group increased less than in the total sample for students with four or fewer home study items; the increase is relatively greater for the students with five items.

7) Students reporting that they did less than two hours of homework are relatively more heavily represented in the low-scoring group while the increase for students reporting two or more hours matches the total sample.

8) Language Minority students are represented in the low-scoring group at about the same rate as in 1984.

9) The increase in representation in the low-scoring group appears related to the percentage of White students in the school, with the highest increases coming from the schools with lower percentages of White students.

10) The increase is also related to the grade in school with the students in grades above the modal grade for their age having a ratio notably higher than in the total sample.

11) The increase in representation in the low-scoring group is less for students in private schools than for the total sample.

12) The increase is related to grades in that higher increases are associated with higher grades, with the exception of "A" students who have a lower increase than the sample as a whole.

13) The increase is relatively less for moderate TV watching.

14) The increase for both levels of the "% SCHOOL LUNCH" variable matches the increase in the full sample.

School-Level Analyses

In searching for explanations for the causes of the reading decline for the 17-year-old students in 1986, it is certainly reasonable to ask whether the decline might be largely associated with a relatively small number of schools. The hypothesis is that while the performance of students in the bulk of the schools in the 1986 sample was at reasonable levels, the performance of students in particular schools was lower than what would be reasonably expected due, perhaps, to factors operating differentially at the school level, such as different administration procedures or conditions or differential levels of student motivation. That is, the increase in the proportion of low scorers in 1986 is hypothesized to be concentrated in a few schools.

One way to see if part of the decline is due to factors operating at the school level is to compare the distributions of school level scores on a comparable set of items given in both of the 1984 and 1986 assessments. Because we need school-level scores, it is necessary that the set of items be presented as a group within the school. One intact set of 11 items was presented within a block of items in both years--the block containing the items in 1986 was block R4. The 11 items corresponded to two reading passages:

A Childhood Memory	with 5 items N007401 through N007405
Old Jim Bridger	with 6 items N007301 through N007306

Although the same set of items was also administered within a block in 1984, there were two changes. First, the order of the passages was reversed between the two years (an analysis of the effect of order on measured reading ability appears in Chapter 11). Second, the text of one of the distractors for the item N007401 was changed between the two years. Because this change effectively resulted in a new item, item N007401 was excluded from the school level analysis.

The school-level means of the student scores (number correct) on the ten retained items of block R4 were computed for each of the schools in which the block was presented in either the 1984 or 1986 assessments. Table 5-2 compares the weighted frequency distribution of the school level scores for the 304 schools given the block in 1984 with that of the 380 schools given the block in 1986. The table also compares the frequency distributions of the school-level scores for the schools classified as low metropolitan and for the high metropolitan schools. Figure 5-1 provides a graph of the frequency distributions for the total set of schools in the two years.

We can see from the table and graph that the distributions of scores for the two years are not the same and that the distribution for 1986 has a higher concentration of lower scores than does the distribution for 1984. The shift in the direction of lower scores is fairly smooth, with no apparent outliers. In fact, the plot resembles the plot of the frequency distributions of student-level reading proficiency scores. The frequency distributions of school-level scores for the low metropolitan and high metropolitan schools also exhibit a higher concentration of lower scores in 1986 than in 1984 but again show no apparent outliers.

Table 5-2

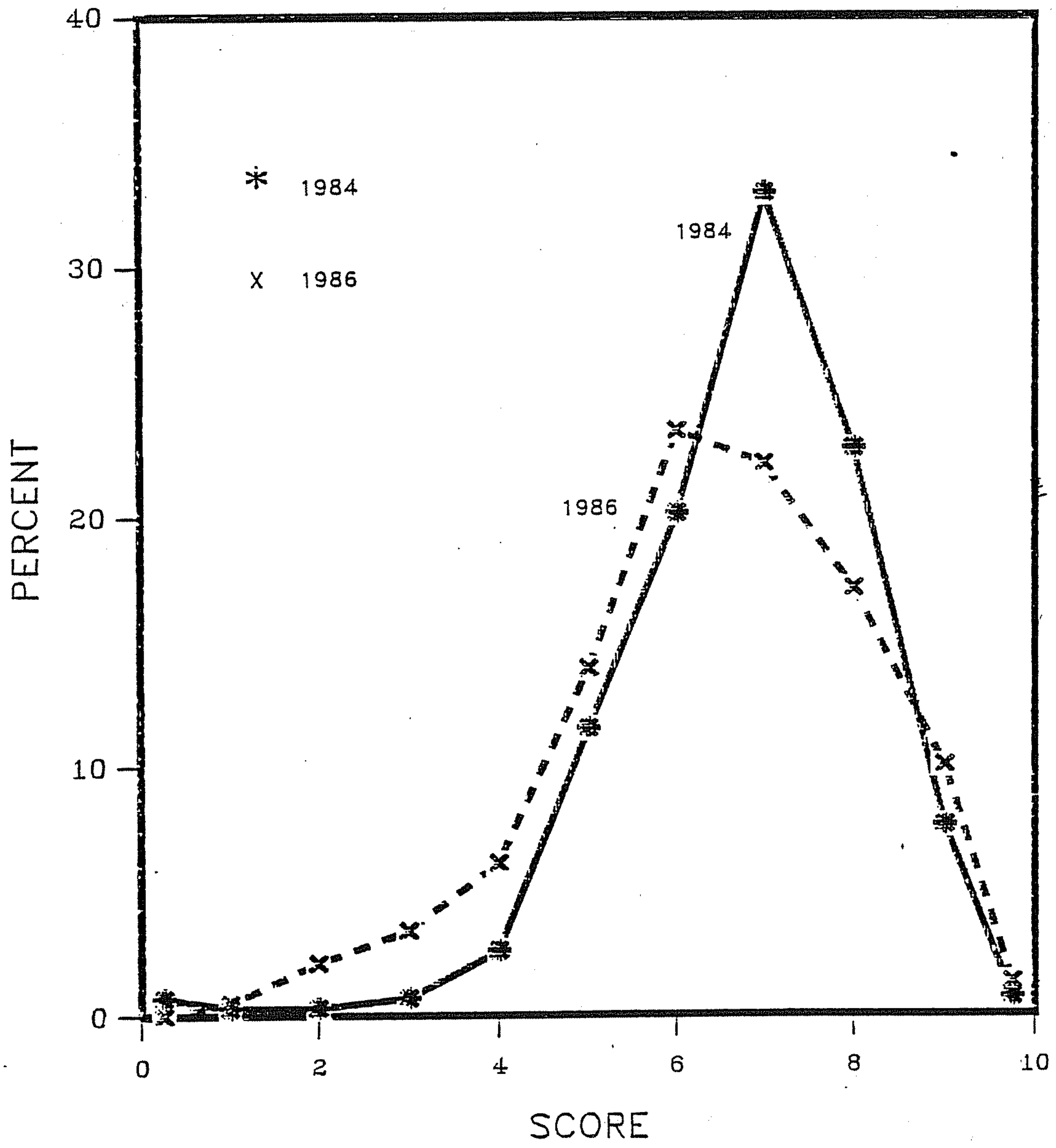
Weighted Frequency Distributions
of School-Level Scores on 10 Items from Block R4*
by Year for All Schools
and for Low Metropolitan and High Metropolitan Schools

	TOTAL		LOW METRO		HI METRO	
Score	1984	1986	1984	1986	1984	1986
9.5 - 10	0.7%	1.3%	--	--	--	2.7%
8.5 - 9.5	7,6%	10.0%	--	--	8.3%	21.6%
7.5 - 8.5	22.7%	17.1%	6.5%	9.7%	22.2%	32.4%
6.5 - 7.5	32.9%	22.1%	12.9%	6.5%	38.9%	18.9%
5.5 - 6.5	20.1%	23.4%	9.7%	19.4%	25.0%	8.1%
4.5 - 5.5	11.5%	13.9%	51.6%	25.8%	2.8%	--
3.5 - 4.5	2.6%	6.1%	16.1%	16.1%	--	13.5%
2.5 - 3.5	0.7%	3.4%	--	9.7%	--	2.7%
1.5 - 2.5	0.3%	2.1%	--	9.7%	--	--
.5 - 1.5	0.3%	.5%	3.2%	3.2%	--	--
0 - .5	0.7%	--	--	--	2.8%	--
Number of schools	304	380	31	31	36	37

* The 10 items are N007301-N007306 and N007402-N007405.

Figure 5-1

Frequency Distribution of School-Level Scores on the 10 Unchanged Items in Block R4 by Year



It would appear that if there is some school-level factor responsible for the decline in the scores, it is either operating in the bulk of the schools in 1986 or, if it is only present in a few schools, it is not restricted to schools of any particular level of student ability.

* * *

In summary, the increase in the relative proportion of low-scoring students in 1986 from 1984 is pervasive, with nearly every demographic subgroup examined having proportionately more members in the lower-scoring group in 1986 than in 1984. This suggests that the decline cannot be associated with any one of the major demographic groups. The results of this chapter suggest that neither can the decline be associated with abnormally low performance in a few schools.

Chapter 6

CONTRIBUTION OF SUBPOPULATIONS TO THE READING ANOMALY

Albert E. Beaton

To investigate the changes in sample composition and their possible effect on the reading proficiency means, we used a method called partitioning analysis. Partitioning analysis is described in Appendix A. The results are shown in Tables 6-1 through 6-6.

There are separate tables for several major categories of NAEP age 17 data: sex, race/ethnicity, region of the country, parental education, grade in school, and size and type of community. The columns in the tables contain the following information:

SUBGROUP contains the various classifications of the students.

P1 contains the proportion of students in each subgroup in the 1984 sample.

P2 contains the proportion of students in each subgroup in the 1986 sample.

DEL P contains the change in population proportion between 1984 and 1986.

XBAR1 contains the estimated average proficiency for each subgroup in 1984.

XBAR2 contains the estimated average proficiency for each subgroup in 1986.

DEL XBAR contains the difference in the average reading proficiency between 1984 and 1986.

T contains the total contribution of each subgroup to the reading decline.

PERF contains the amount that each subgroup's change in performance contributed to the estimated reading decline.

POP contains the amount that each subgroup's change in proportion of the population contributed to the estimated reading decline.

INTER contains the amount that the interaction between each subgroup's performance and proportion of the population contributed to the decline.

Perusal of the DEL P columns of these tables shows that the compositions of the samples did not change very much between 1984 and 1986. The largest shift was in the proportion of students whose parents had some post-high school education; this group grew from 49.5% to 59.5% of the population. However, we would have expected this shift to result in a rise in estimated reading proficiency.

The DEL XBAR column shows how pervasive the decline was. All subgroups of students declined, with the smallest decline being 3.5 points for the High Metro (high socioeconomic metropolitan) area. The estimated averages of several subgroups dropped by over 20 points on the NAEP reading scale.

The most telling part of these tables is in the bottom line, where the total effect of changes in performance and population are presented. In most cases, the shift in population should have resulted in an increase in the estimated average reading proficiency, but this increase is overwhelmed by the decrease in performance.

Many other partitioning analyses were also computed using selected two- and three-way breakdowns to assure that no higher-order interaction might explain the results. No such higher-order interaction was found.

Therefore, we conclude that changes in the estimated reading proficiency, not population shifts, were the major contributors to the overall estimated decline in reading proficiency.

Table 6-1

```
                                                                                      12:58:04  10/12/87  PAGE   22
NAEP  1984 AND 86 READING ASSESSMENTS                                                        F4STAT 3.25
COMPARE READING PROFICIENCY FOR AGE 17

1 WAY TABLE:  SEX

SUBGROUP         P1       P2    DEL P   |   XBAR1     XBAR2  DEL XBAR  |     - T -    |      PERF         POP       INTER

MALE          0.511    0.509   -0.002   |   282.6     269.8     -12.8  |   -7.0708    |    -6.5163     -0.5807      0.0262
FEMALE        0.489    0.491    0.002   |   293.6     285.2      -8.4  |   -3.5278    |    -4.1139      0.6034     -0.0173

TOTAL         1.000    1.000    0.00C   |   288.0     277.4     -21.2  |  -10.5986    |   -10.6302      0.0227      0.0089
```

Table 6-2

```
NAEP  1984 AND 86 READING ASSESSMENTS                                                                         12:58:04  10/12/87  PAGE   23
COMPARE READING PROFICIENCY FOR AGE 17                                                                                 F4STAT 3.25

1 WAY TABLE:  ETHNICITY

SUBGROUP        P1      P2   DEL P   |   XBAR1     XBAR2  DEL XBAR  |     - T -    |      PERF        POP       INTER

WHITE        0.774   0.780   0.005   |   293.8     283.9     -10.0  |   -6.1764    |   -7.7165     1.5941     -0.0541
BLACK        0.141   0.135  -0.006   |   265.2     251.8     -13.4  |   -3.5025    |   -1.8972    -1.6910      0.0857
HISPANIC     0.066   0.062  -0.004   |   268.0     255.3     -12.7  |   -1.8211    |   -0.8356    -1.0346      0.0491
OTH ETH      0.019   0.024   0.005   |   286.4     266.4     -20.0  |    0.9011    |   -0.3810     1.3781     -0.0961

TOTAL        1.000   1.000   0.000   |   288.0     277.4     -56.1  |  -10.5989    |  -10.8303     0.2467     -0.0153
```

Table 6-3

```
                                                                                                  12:58:04  10/12/87  PAGE   24
NAEP  1984 AND 86 READING ASSESSMENTS                                                                     F4STAT 3.25
COMPARE READING PROFICIENCY FOR AGE 17

1 WAY TABLE:  REGION

SUBGROUP          P1        P2    DEL P   |   XBAR1     XBAR2  DEL XBAR  |     - T -    |       PERF         POP       INTER

NE             0.243     0.238   -0.006   |   289.6     286.0      -3.5  |   -2.5195    |    -0.8608     -1.6792      0.0205
SE             0.222     0.212   -0.010   |   285.2     269.4     -15.8  |   -6.1708    |    -3.5221     -2.8045      0.1558
CENTRAL        0.271     0.284    0.014   |   289.1     279.7      -9.4  |    1.2480    |    -2.5453      3.9207     -0.1274
WEST           0.263     0.265    0.002   |   287.7     273.5     -14.1  |   -3.1576    |    -3.7236      0.5954     -0.0293

TOTAL          1.000     1.000    0.000   |   288.0     277.4     -42.9  |  -10.5998    |   -10.6518      0.0324      0.0196
```

Table 6-4

```
NAEP  1984 AND 86 READING ASSESSMENTS                                                              12:58:04  10/12/87  PAGE   25
COMPARE READING PROFICIENCY FOR AGE 17                                                                        F4STAT 3.25

1 WAY TABLE:  PARENTAL EDUCATION

SUBGROUP        P1       P2    DEL P   |   XBAR1     XBAR2  DEL XBAR  |      - T -    |      PERF        POP        INTER

NOT HS       0.115    0.089   -0.026   |   269.8     253.4     -16.4  |    -8.5450    |    -1.8874    -7.0884       0.4307
GRAD HS      0.348    0.277   -0.071   |   280.6     264.9     -15.7  |   -24.2486    |    -5.4580   -19.9037       1.1131
POST HS      0.495    0.594    0.100   |   299.5     289.3     -10.2  |    23.7689    |    -5.0561    29.8433      -1.0183
OTH EDUC     0.042    0.040   -0.002   |   263.0     239.5     -23.5  |    -1.5751    |    -0.9950    -0.6370       0.0569

TOTAL        1.000    1.000    0.000   |   288.0     277.4     -65.8  |   -10.5998    |   -13.3965     2.2142       0.5824
```

Table 6-5

```
                                                                                                12:58:04  10/12/87  PAGE   26
NAEP  1984 AND 86 READING ASSESSMENTS                                                                     F4STAT 3.25
COMPARE READING PROFICIENCY FOR AGE 17

1 WAY TABLE:  GRADE IN SCHOOL

SUBGROUP         P1       P2   DEL P   |   XBAR1     XBAR2  DEL XBAR  |    - T -    |     PERF        POP      INTER

< MODAL       0.222    0.218  -0.005   |   260.2     242.9     -17.3  |   -4.9604   |   -3.8348    -1.2056     0.0800
AT MODAL      0.677    0.703   0.026   |   294.7     286.0      -8.7  |    1.4891   |   -5.9196     7.6350    -0.2264
> MODAL       0.100    0.079  -0.021   |   303.9     295.5      -8.4  |   -7.1286   |   -0.8433    -6.4640     0.1788

TOTAL         1.000    1.000   0.000   |   288.0     277.4     -34.4  |  -10.5998   |  -10.5977    -0.0346     0.0324
```

Table 6-6

```
NAEP  1984 AND 86 READING ASSESSMENTS                                                                     12:58:04  10/12/87  PAGE   27
COMPARE READING PROFICIENCY FOR AGE 17                                                                               F4STAT 3.25

1 WAY TABLE:  STOC

SUBGROUP        P1       P2   DEL P   |   XBAR1    XBAR2  DEL XBAR  |     - T -    |     PERF       POP      INTER

RURAL        0.051    0.044  -0.007   |   282.6    272.0     -10.7  |   -2.4477    |   -0.5420   -1.9804     0.0747
LO MET       0.102    0.062  -0.040   |   265.5    247.9     -17.6  |  -11.6871    |   -1.7898  -10.5994     0.7020
HIGH MET     0.161    0.125  -0.036   |   299.7    297.1      -2.6  |  -11.0187    |   -0.4209  -10.6913     0.0934
BIG CITY     0.089    0.085  -0.004   |   289.2    267.1     -22.1  |   -3.0263    |   -1.9569   -1.1576     0.0883
FRINGE       0.108    0.151   0.043   |   289.6    278.3     -11.3  |   10.6789    |   -1.2217   12.3838    -0.4832
MED CITY     0.165    0.159  -0.006   |   291.3    279.1     -12.1  |   -3.5483    |   -2.0005   -1.6151     0.0673
SMALL PL     0.325    0.374   0.049   |   287.5    277.5     -10.0  |   10.4494    |   -3.2572   14.2017    -0.4951

TOTAL        1.000    1.000   0.000   |   288.0    277.4     -86.4  |  -10.5998    |  -11.1890    0.5418     0.0474
```

Chapter 7

BOOKLET FORMAT AND ADMINISTRATION

Janet R. Johnson

There were a number of administrative and format changes made between the 1984 and 1986 assessments. This chapter provides a description and tabular summary of these changes.

Table 7-1 summarizes some of the salient distinctions between the 1984 and the 1986 assessments. As can be seen in the first section of the table, the 1984 data, against which the recent data are being compared, are drawn from the BIB-spiralled section of the 1984 assessment. The BIB section of the assessment collected both age and grade data. For the reading trend comparisons only the age data were used. The comparable 1986 reading trend data for ages 9 and 13 have been based on the 1986 bridge samples. These bridge studies were conducted at approximately the same time of year as in 1984--fall for 13-year-olds and winter for 9-year-olds. For age 17, the 1986 reading trend has been based on data collected from the BIB section of the assessment.

Each student at all three ages in the 1984 BIB-spiralled assessment sessions took a single booklet containing three blocks. These booklets contained 0, 1, 2, or 3 reading blocks; the remaining blocks, if any, consisted of writing exercises. The students had to read some instructions and the exercise texts. Each student in the 1986 bridge assessments for ages 9 and 13 took a single booklet containing three blocks. The subjects were math, science, and reading and the booklets were configured as shown in Table 7-2. The same booklet was administered to an entire assessment session. The math and science parts of the booklets were paced (presented aurally using a tape recorder). The tape recorder was turned off for the reading block in each session.

For age 17, the BIB booklets in 1986 contained 0, 1, 2, or 3 reading blocks; the remaining blocks, if any, were in math, science, computer competence, or, in the case of 4 of the 97 booklets, history and literature. In 1986, the age 13 and age 17 reading blocks were identical in every respect so that the three blocks (13R1, 13R2, and 13R3) used in the age 13 bridge sample were repeated as part of the age 17 BIB reading blocks. Different students in the same session were administered different booklets.

The length of time allotted for each block changed between 1984 and 1986. In 1984 each age was given a six-minute common core of background and attitude questions followed by three subject area blocks of fourteen minutes each. At the end of each fourteen-minute interval, the students were told to move to the next block. Approximately the first two minutes of these subject area blocks were devoted to answering additional attitude questions related to the curriculum area. In 1986, the age 13 and 17 students again had six minutes to respond to the common core background and attitude questions;

Table 7-1

Comparisons for Measuring Reading Trend

Sample

	1984		1986
Age 9)	BIB	9)	Bridge
Age 13)	BIB	13)	Bridge
Age 17)	BIB	17)	BIB

Subjects Assessed and Method

	1984		1986
9)	Rdg, Writ. (BIB)	9)	Rdg, Math, Sci (pseudo-BIB)
13)	Rdg, Writ. (BIB)	13)	Rdg*, Math, Sci (pseudo-BIB)
17)	Rdg, Writ. (BIB)	17)	Rdg, Math, Sci, Computer (BIB) Competence, History**, Literature**

*The format and content of the 1986 age 13 reading blocks were identical in every respect to those used at age 17.
**Four of the 97 booklets at age 17 contained a history and a literature block combined with one reading block (13R4) as a special probe into those two subjects.

Block Timing

	1984			1986	
	Common Core Minutes	Blocks Mins.		Common Core Minutes	Blocks Mins.
9)	6	14	9)	15*	13
13)	6	14	13)	6	16
17)	6	14	17)	6	16

*The common core questions for age 9 1986 were read aloud to the students.

Dates of Assessment

	1984		1986
9)	Jan 2-Mar 19, 1984	9)	Jan 6-31, 1986
13)	Oct 10-Dec 17, 1983	13)	Nov 4-Dec 13, 1985
17)	Mar 12-May 11, 1984	17)	Feb 17-May 2, 1986

Table 7-1
(continued)

Average Session Sizes

	1984		1986
9)	Fewer than 25	9)	Fewer than 25
13)	Fewer than 25	13)	Fewer than 25
17)	Approximately 20	17)	Approximately 35

Booklet Format

	1984		1986
9)	Blue type, saddle stitched	9)	Blue type*, stapled
13)	Brown type, saddle stitched	13)	Blue type*, stapled
17)	Black type, saddle stitched	17)	Blue type*, stapled

*Slightly smaller type was used in 1986. Average line length was less than five inches in 1984 reading passages and over five inches in 1986 reading passages.

9)	"circle the letter" responses	9)	"fill in the oval" responses
13)	"circle the letter" responses	13)	"fill in the oval" responses
17)	"circle the letter" responses	17)	"fill in the oval" responses

Booklet Scoring

	1984		1986
9)	Key entered	9)	Machine scored
13)	Key entered	13)	Machine scored
17)	Key entered	17)	Machine scored

Language Minority Booklets

	1984		1986
9)	None	9)	None
13)	None	13)	None
17)	None	17)	Yes

Teacher Questionnaire

	1984		1986
9)	Language arts teacher	9)	None
13)	Language arts teacher	13)	None
17)	Language arts teacher	17)	Up to 5 teachers, identified by students

Table 7-2
1986 Bridge Booklet Configuration

	Age 9			Age 13		
Booklet 1	9R1	9M1	9S1	13R1	13M1	13S1
Booklet 2	9S2	9R2	9M3	13S2	13R2	13M3
Booklet 3	9M2	9S3	9R3	13M2	13S3	13R3

however, for 9-year-olds, the common core questions at the beginning of each booklet were read aloud to them and took 15 minutes to complete. The 9-year-olds were given 13 minutes to read and respond to the exercises in the block; the 13- and 17 year-olds were given 16 minutes. Each reading block at all ages began with additional attitude questions. The number of exercises per block was increased to allow for the amount of time allotted for each block.

(The process of developing NAEP reading objectives and items, not discussed here, is lengthy and complex. It has traditionally involved the thinking of a wide variety of persons and continues to do so. For a thorough discussion of the steps, please refer to Chapter 3 and Chapter 6 in Implementing the New Design: The NAEP 1983-84 Technical Report (Beaton, 1987) and also to Reading Objectives: 1983-84 Assessment (NAEP, 1984) and Reading Objectives 1986 and 1988 Assessments (NAEP, 1987). These publications are available from NAEP.)

As Table 7-1 indicates, the dates that the assessments were conducted varied somewhat between 1984 and 1986, as did the average number of students per session for the age 17 cohort.

The booklets differed in style and construction between the two assessments. For example, the color of type changed between 1984 and 1986. The size of the type was slightly smaller in 1986 while the length of the lines in the reading passage texts was slightly longer. Also, booklets in 1986 were designed to be machine-scorable so that students, when choosing a response to an item, were asked to fill in a scannable bubble in the booklet preceding their response choice; in 1984 students were asked to circle the letter preceding their response choice. The 1988 bridge studies will enable us to ascertain if such changes had an effect upon reading performance.

The spring 1986 assessment covered all three ages during the same period for the first time, and also incorporated a special Language Minority Probe that involved a specially tailored block of questions to be asked of Language Minority students during the main spring BIB assessment. For the purposes of this discussion, only age 17 students are of concern here since ages 9 and 13 reading data were based on the bridge samples. The language minority block followed the standard common core block of background questions and always preceded a reading and a math block in that order. This special study necessitated an additional number of sampling tasks within the school at the time of the assessment, tended to congregate designated Language Minority students together into one session, and, in general, increased administrative complexity.

Another on-site administrative complexity that was not present in 1984 was the increased number of subject area teachers to be queried. The sampling process for the selection of teachers was a multi-step affair that resulted in as many as five teachers' names being introduced to the 17-year-old students within a session before the actual assessment could begin, in order to ascertain the names of teachers to whom the teacher questionnaire would be given. The 9- and 13-year-old student data were based on the bridge study, which did not involve the Teacher Survey.

* * *

There has been some speculation as to whether some of the various changes between assessments could have accounted for the observed declines at ages 9 and 17. The 1988 bridge studies that faithfully replicate the 1984 and 1986 assessments should be able to provide us with the data necessary to answer the overall questions of the trend of American students' reading proficiency.

Chapter 8

DATABASE QUALITY CONTROL

John J. Ferris

All NAEP data were checked for accuracy and consistency. Both quality control and editing were done; this discussion deals with the quality control operation, which is an assessment of the accuracy of the data entry process.

The 1986 student data received additional attention in an effort to verify that the data were not the cause of the drop in P-values for the reading items. The quality control operation revealed that, in fact, the data entry appeared to be exceptionally accurate.

One of each booklet for each age level was selected at random as part of the basic quality control procedure for student data. The booklets in this assessment were machine-scored by a scanner. Some of the booklets, however, could not be scanned because of damage to the booklet, irregular printing, or other problems that caused the scanner to reject the document. These booklets had to be keyed by hand. Although they represented only 2.3 percent of the student data booklets, it was felt desirable to select one each of these booklets as well for quality control. In total, 213 scanned booklets and 213 keyed booklets were compared stroke for stroke with the corresponding records in the database. In each set of booklets, 42,366 keystrokes were involved.

The scanned booklets, as expected, showed the higher level of accuracy: only one error, a pickup of an erasure, was found among these 42,366 keystrokes. The keyed booklets contained a total of 11 incorrect keystrokes scattered among 11 booklets. The following table summarizes these findings in the same terms as were used to measure the accuracy of the data entry in the previous assessment.

	OBSERVED ERROR RATE	UPPER 99.8% CONFIDENCE LIMIT*
Scanned booklets	.00002	.0002
Keyed booklets	.00026	.0005

*This is to say that the probability is .998 that the true error rate for a scanned booklet does not exceed .0002 and for a keyed booklet does not exceed .0005.

As mentioned above, additional work was done in an attempt to reveal problems with the accuracy of the data. A great many booklets containing blocks of reading items were examined for any signs of irregularity or flaws that might lead to misconstruing the data; none were found. Patterns of missing data and patterns of wrong answers were listed and examined; no unusual or inexplicable patterns emerged from this analysis. Significantly, the P-values from booklets with no missing data were comparable to the P-values already determined; in other words, removing records with potentially "bad" patterns of responses had no discernible effect on the calculated P-values.

In conclusion, there is no support for a hypothesis of badly entered data in explanation of the apparent drop in reading performance shown by the 1986 NAEP student data.

Chapter 9

SCALING

Robert J. Mislevy

Inasmuch as the 1984-to-1986 reading change is the first NAEP change reported in scale scores, it is pertinent to ask whether scaling procedures were in some way responsible for the unexpected declines seen for the 9- and 17-year-old students. As reported in the other chapters of this report, however, the unexpected drops are equally evident in the item percents correct (i.e., "item-p's") of items common to the two time points. Since these values are completely independent of scaling procedures, the scaling procedures cannot be the cause.

One may yet ask whether the drops in percents-correct units are commensurate with those in scale-score units. The purpose of this chapter is to use 1984 data to calculate a simple relationship between scale-scores and item-p's, then use it to approximate scale-score means for 1986 based solely on 1986 item p's--bypassing entirely the scaling activities carried out in 1986.

The following discussion documents an approximation of NAEP reading proficiency values from NAEP assessments in 1984 and 1986 based on transformations of item percents correct. The steps of the approximation are

1) approximating the marginal probability of a correct response from a normally distributed population under the 3-parameter logistic (3PL) IRT model by a 3PL group-level IRT model (as in Mislevy, 1983);

2) treating the items presented within an age as identical, using the averages of the estimated a, b, and c parameters; and

3) treating the population standard deviations as known.

With these expedients, a closed-form solution exists for estimating the group mean.

Formulas

Suppose all items within an age are identical with parameters a, b, and c, and examinees' responses, given their θ values, follow the 3PL:

$$P(x=1|\theta) = c + (1-c)/(1 + \exp(-1.7a(\theta-b))). \qquad (1)$$

This is the probability of a correct response from a specific individual with known θ. The probability of a correct response from a person selected at random from a population in which θ has the probability density function g is then obtained as

$$P(x=1) = \int P(x=1|\theta)\ g(\theta)\ d\theta. \tag{2}$$

If g is normal with mean μ and standard deviation s, it can be shown (Mislevy, 1983) that (2) can be approximated in the form of another 3PL:

$$P \equiv P(x=1|\mu,s) = c + (1-c)/(1 + \exp(-1.7A(\mu-b))), \tag{3}$$

where

$$A = (a^{-2} + s^{-2})^{-1/2} .$$

If a, b, c, and s are known, then (3) can be solved for μ:

$$\mu = b - (1.7A)^{-1} \ln((1-P)/(P-c)). \tag{4}$$

Data

Equation 4 was applied to NAEP data in the following manner. For each age, the average a, b, and c were estimated for the trend items that linked 1984 and 1986. An average p-value was computed in each age/year for the link items; the averaging was done in the logit metric, and the average logit was converted back to a p-value for use in Equation 4. Based on 1984 results, an s of .75 on the θ scale (=37.5 on the RP scale) was assumed for all three ages.

Results

Table 9-1 summarizes computations. The columns represent the year/age combinations. The rows correspond to estimated item parameters a, b, and c, followed by the average logit percent correct and its corresponding p. The next row, logit*, is ln((1-p)/(p-c)). Next comes A, followed by the solution for μ, its conversion to the RP scale (θ*50 + 250.5), the estimated change values within age between 1984 and 1986, and the actual change values. The estimated change values are within about two points of the actual change as computed using student-level data and actual item parameters.

Conclusion

The changes in scale-score means from 1984 to 1986 predicted by a simple rescaling of item percents correct are in the same directions as, and of similar magnitudes to, the scale-score changes produced by the full scaling procedures. The scale-score declines, therefore, are consistent with item percents-correct declines. Investigations into the <u>causes</u> of those declines may focus on the simpler-to-analyze raw item response data.

Table 9-1
Summary of Computations

Computation	Year/Age					
	84/9	86/9	84/13	86/13	84/17	86/17
a	1.435	1.425	1.601	1.601	1.435	1.425
b	-0.979	-0.979	-0.026	-0.026	-0.024	-0.024
c	0.192	0.192	0.202	0.202	0.197	0.197
logit	0.673	0.502	0.816	0.843	1.415	1.149
p	0.662	0.623	0.693	0.699	0.805	0.759
logit*	-0.330	-0.134	-0.470	-0.501	-1.137	-0.847
A	0.979	0.979	1.031	1.031	0.978	0.978
μ	-0.781	-0.899	0.242	0.260	0.708	0.533
rp	211.470	205.590	262.602	263.511	285.888	277.157
estimated change		-5.881		0.909		-8.731
actual change		-5.600		2.400		-11.400

Chapter 10

ITEM DATA ANALYSES

Eugene G. Johnson

The basis for the measurement of reading achievement is the data describing, for each assessed student, the action taken to each of the items presented to that student. These actions of a student to a given item are of five general types:

1) Correct: the student selected the correct answer;

2) Incorrect: the student selected another, incorrect, answer;

3) IDK : the student responded "I Don't Know";

4) Omit : the student skipped (omitted) the item but responded (Right, Wrong, or IDK) to an item later in the block; and

5) Not reached : the student did not respond to the item or to any later item in the block of items, presumably because the student had run out of time.

While the first two categories (Correct and Incorrect) clearly impart unequivocal information about a student's reading ability, the remaining three do not since each of these categories--IDK, Omit, and Not Reached--are describing a particular kind of nonresponse to the reading item. The ultimate measure of reading ability, whether for an individual item (in the form of a percent correct) or across all items (in the form of a mean proficiency score), depends on the assumptions made about the performance characteristics of the nonresponding students.

In this report, the item-level percent correct statistic is computed as the ratio of the weighted sum of the correct responses to the weighted sum of the Correct, Incorrect, and IDK categories. The Omit and Not Reached categories are excluded from the computation, which is tantamount to saying that the students who omitted or did not reach the item, had they had enough time and the inclination to answer the item, would have responded correctly at the same rate as those students who were exposed to the item. On the other hand, the students in the remaining category of nonresponse, IDK, are assumed to have been exposed to the item but to have elected not to select any of the available answers, presumably because they did not know the correct answer and were unwilling to guess. For the purposes of estimating item percent correct in this report, these IDK choices are treated as incorrect answers to the item.

The Not Reached category is treated in the same manner in the estimation of a reading proficiency score as it was for item percent correct, but the IDK and Omit categories are treated differently. In this case, it is assumed that had the IDK and Omit choices not been available, the student would have guessed and selected one of the available valid response alternatives. That is, in the scaling process the IDK and Omit responses are treated as fractionally correct, at a proportion equal to the reciprocal of the number of valid alternatives available.

These discrepancies in treatment of the nonresponse categories are of little consequence so long as the percent of this nonresponse is small, but become important for higher levels of nonresponse. The reason for this is that the actual statistics involve both the known data (the responses to the valid alternatives) and the assumption made about the unknown data (the nonresponse categories). As the fraction of nonresponse increases, the importance of these assumptions on the final estimates also increases. Nonetheless, as long as the rates of (and reasons for) each of the types of nonresponse remain nearly constant over time, the measurement of trends in achievement will not be greatly impacted.

The problem arises when the rates of nonresponse, and/or the reasons for the nonresponse, change over time, for then the populations actually being measured are also changing over time. For example, if the rate of nonresponse is increasing over time and the additional nonrespondents are among the better-performing students, then the current population that is actually being measured (the respondents) will be of a lower overall ability than in previous years because of the exclusion of these higher-performing students.

In this section we will compare the response patterns of students in the 1986 assessment to those of students from the 1984 assessment, basing this comparison on the set of reading items common to both years.

We begin with the 39 reading items that were given to 17-year-olds in both of the last two assessments. The distributions across the 39 items of the weighted percentages of students selecting each of the five response alternatives are given in Table 10-1 for each of the two assessment years. The table also gives the distributions of the item-level differences between the two assessment years. Part (a) of the table addresses the distributions of the responses of assessed 17-year-olds in 1984. The first row of Part (a) shows the average weighted percent of students selecting the correct response alternative in 1984, this average being taken across the 39 items. The other statistics given in the first row are selected order statistics for the 39 item-level percents correct (the minimum, lower quartile, median, upper quartile and maximum). The remaining rows of Part (a) provide the equivalent information about the distributions in 1984 for each of the remaining response alternatives : Incorrect, I Don't Know, Omit and Not Reached. The item-level distributions for the students assessed in 1986 for each of the five response alternatives are given in Part (b) of the table. Finally, Part (c) gives the distributions of the differences, by item, in response activity in the two assessment years.

Table 10-1

Distributions Across Items of the Weighted Percents of Students Selecting the Correct, Incorrect, I Don't Know, Omit and Not Reached Item-Level Responses by Year for the Common Reading Items

a) Age 17: 1984 - 39 items

Type of Response	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
CORRECT	73.6	22.0	67.4	78.4	86.3	94.6
INCORRECT	18.6	2.1	8.6	13.9	22.7	71.5
I DON'T KNOW	4.1	0.7	2.0	3.0	5.0	18.7
OMIT	0.6	0.2	0.4	0.5	0.7	1.0
NOT REACHED	3.1	0.3	0.9	2.6	4.4	10.5

b) Age 17: 1986 - 39 items

Type of Response	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
CORRECT	71.1	27.6	65.1	74.9	82.6	93.2
INCORRECT	20.8	3.8	10.9	17.2	25.8	60.8
I DON'T KNOW	5.2	1.5	3.1	4.5	6.3	23.4
OMIT	0.5	0.1	0.3	0.5	0.8	1.2
NOT REACHED	2.4	0.6	0.8	1.5	3.6	7.1

c) Age 17: 1986 - 1984 item-level differences - 39 items

Type of Response	Mean	Minimum	Lower Quartile	Median	Upper Quartile	Maximum	No. of Diff ≤0
CORRECT	-2.5	-12.7	-5.5	-2.6	1.3	7.7	27
INCORRECT	2.1	-10.7	1.3	2.2	3.9	11.6	4
I DON'T KNOW	1.1	-4.4	0.4	1.1	1.8	4.7	5
OMIT	-0.1	-0.4	-0.2	-0.0	0.2	0.5	22
NOT REACHED	-0.7	-9.6	-3.2	-0.1	1.4	6.0	22

Table 10-1 shows that the percentage of students selecting the correct answer to an item has declined between 1984 and 1986 in 27 of the 39 items for an average decline of 2.5 percentage points. The table also shows that this decline is nearly matched by an increase, averaging 2.1 percentage points, in the rates of students responding but selecting an incorrect answer. In fact, the rate of incorrect answers among the respondents was larger in 1986 than it was in 1984 for 35 of the 39 items. This means that, although the response rate (the proportion of students selecting a valid answer--either correct or incorrect) has declined only slightly between the two years (from 92.2 percent in 1984 to 91.9 percent in 1986), relatively more of the respondents in 1986 selected an incorrect answer than did the respondents in 1984.

The increase in the overall nonresponse rate (from 7.8 percent to 8.1 percent) is too small at .3 percentage points to account for more than a small fraction of the decline in reading ability unless there has been a marked change in the type of (or reason for) nonresponse between the two years. While changes in the reasons for nonresponse are not directly estimable from the data, changes in the relative proportions of each of the defined types of nonresponse are. Table 10-1 shows that there indeed have been some changes in the rates of the three types of nonresponse over the past two assessments with an increase (averaging 1.1 percent) in the proportion of students selecting the I Don't Know option and a decrease (averaging .7 percent) in the proportion of students not reaching the item. The change in the omit rate between the 1984 and 1986 assessments is much less, averaging -.1 percent.

The overall picture is that, relative to the 17-year-old students assessed in 1984, the 17-year-olds in 1986 are selecting the correct answers less often, selecting a wrong answer more often, and omitting items at about the same rate. The proportions of students not reaching items have decreased in 1986 but this decrease is more than made up for by an increase in the proportions of I Don't Know responses.

We find that the changes in patterns of response are similar when we restrict our attention to certain demographic subgroups of the students, corresponding to subpopulations with markedly different characteristics in terms of achievement. Table 10-2 shows the mean percentages by year, and changes between the two years, in each of the five response categories for each of the following classifications of students:

Grade: < modal, at modal, > modal
STOC : Low Metropolitan, High Metropolitan
Parent's Education : Not Graduated High School, Graduated High School, Post High School, Graduated College.

The major difference in the changes in response patterns for the subpopulations is in degree rather than kind. In every case, the proportion of wrong answers and I Don't Know responses has gone up with the largest increases in the lower-ability subgroups (< Modal Grade, Low Metro, Not Grad

Table 10-2

Average Percents by Year and Demographic Subgroups of Students
Selecting Each of the Categories:
Correct, Incorrect, I Don't Know, Omit and Not Reached

	Correct	Incorrect	IDK	Omit	Not reached
		a) Age 17 in 1984			
TOTAL	73.58	18.77	4.14	0.40	3.11
<MODAL GRADE	59.85	28.25	6.38	0.62	4.90
AT MODAL GRADE	77.00	16.45	3.59	0.34	2.62
>MODAL GRADE	80.46	13.77	2.95	0.32	2.50
LOW METRO	61.04	25.13	4.96	0.72	8.15
HIGH METRO	77.93	14.75	3.36	0.34	3.63
PARENT:					
NOT GRAD HS	66.55	24.33	5.38	0.47	3.27
GRAD HS	71.35	20.94	4.30	0.40	3.01
MORE THAN HS	77.04	16.34	3.50	0.31	2.81
GRAD COLLEGE	78.44	14.89	3.46	0.42	2.78
		b) Age 17 in 1986			
TOTAL	71.11	20.84	5.23	0.42	2.41
<MODAL GRADE	55.07	31.47	7.91	0.68	4.86
AT MODAL GRADE	75.04	18.21	4.60	0.37	1.78
>MODAL GRADE	79.39	15.37	3.55	0.18	1.50
LOW METRO	55.40	29.32	7.50	0.92	6.86
HIGH METRO	78.61	15.29	3.52	0.43	2.14
PARENT:					
NOT GRAD HS	60.47	27.25	7.42	0.75	4.12
GRAD HS	66.80	24.55	5.67	0.57	2.40
MORE THAN HS	73.74	19.18	4.96	0.24	1.88
GRAD COLLEGE	77.48	16.56	3.93	0.33	1.70
	c)	Age 17: 1986-1984 Differences			
TOTAL	-2.47	2.07	1.09	0.02	-0.70
<MODAL GRADE	-4.78	3.22	1.53	0.06	-0.04
AT MODAL GRADE	-1.96	1.76	1.01	0.03	-0.84
>MODAL GRADE	-1.07	1.60	0.60	-0.14	-1.00
LOW METRO	-5.64	4.19	2.54	0.20	-1.29
HIGH METRO	0.68	0.54	0.16	0.09	-1.49
PARENT					
NOT GRAD HS	-6.08	2.92	2.04	0.28	0.85
GRAD HS	-4.55	3.61	1.37	0.17	-0.61
MORE THAN HS	-3.30	2.84	1.46	-0.07	-0.93
GRAD COLLEGE	-0.96	1.67	0.47	-0.09	-1.08

HS) and the smallest increases for the higher-ability subgroups (>Modal Grade, High Metro, Grad College). With one exception (students whose parents did not graduate high school), the proportions of students not reaching the items have gone down. These changes in proportions are also tied to ability level, with the largest changes occurring for the higher-ability students.

Before considering how much of the decline in the reading scores can be attributed to these changes in the patterns of response, and particularly in the patterns of nonresponse, the issue of the comparability of the rates between the two years must be addressed. The issue is that, although the 39 trend items were presented in both assessments, the groupings of those items into blocks, as well as their positions in the blocks, was different in the two years. This means that the between-year comparison of the rates of each of the types of nonresponse to a given item is confounded by changes in the position of the item.

Of the three types of nonresponse, we would expect the rate of not reaching the item to be the most affected by this confounding since this type of nonresponse is most obviously tied to the position of the item in the block. In fact, the nonresponse rate of an item due to not reaching that item depends both on the location of the item in the block, measured by how many other items precede the item, and on the characteristics of the preceding items, since a few items associated with a long passage may require a longer response time than the same number of items associated with a shorter passage. While it is difficult to disentangle the effects of context on the nonresponse rates, we can at least approximately control for the effect of the location of the items.

Let NR be the item nonresponse rate due to not reaching the item and let ILOC be the location of the item in the cognitive portion of the block (i.e., the number of items past the last background question). A plot of NR versus ILOC suggested that a reasonable description of the relationship between the not reached rate and item location would be

$$\ln(NR) = a + b*ILOC,$$

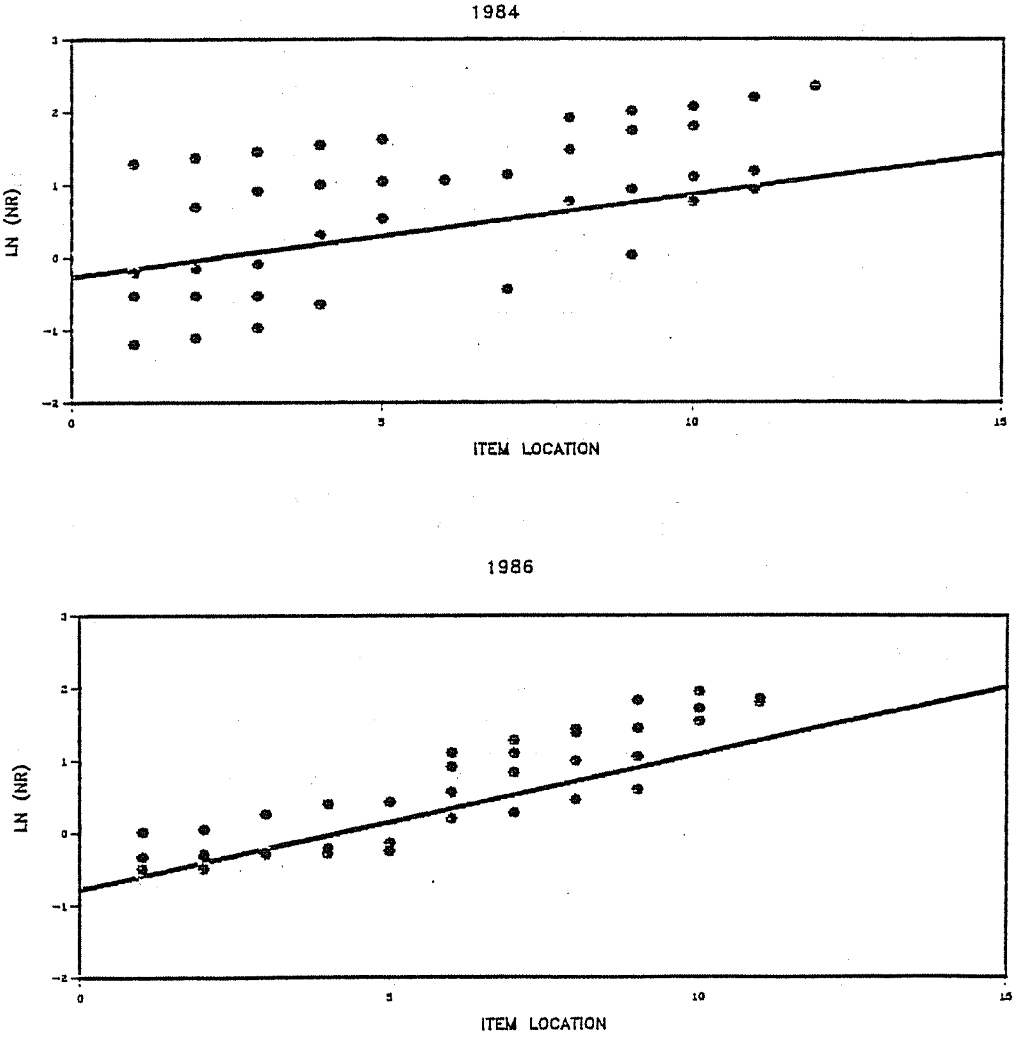
where ln is the natural log. (Plots of ln(NR) vs. ILOC, for each year, appear in Figure 10-1). This formula works fairly well for the 1986 data (R^2=.81) but less well for the 1984 data (R^2=.38). The poorer fit for the 1984 data is due to higher variability in the not reached rates at each level of the item location variable ILOC -- this variability is approximately constant over ILOC. The parameters of the fits (with standard errors in parentheses) are as follows:

1984	a= -.27 (.25)	b= .17 (.037)
1986	a= -.78 (.12)	b= .23 (.018)

The slopes of these two lines are not significantly different at any reasonable level of significance. The chief difference between the rates of nonresponse due to not reaching the item is in the overall level, after

Figure 10-1

Plot of the Log of the Not Reached Rate
Versus Item Location by Year
With Least-Square Lines



adjusting for item position the not reached rate for 1986 is around 60 percent (=exp(-.51)) of that of 1984.

The rate of nonresponse due to omitting the item or due to selecting I Don't Know is not strongly tied to the location of the item (the values of R^2 for predictions of the rates of these types of nonresponse--or their logs--by ILOC are all less than .05).

It follows that any differences in the rates of the various kinds of nonresponse between the two assessments is primarily due to something other than changes in the locations of the items. Among other possibilities are changes in the context, meaning changes in the characteristics of the preceding items, changes in administration, changes in format, and changes in motivation. Each of these possibilities is discussed in other chapters.

For the present, we will restrict our attention to estimating how much of the decline in reading proficiency scores can be attributed to these changes in patterns of response. To do this, we will employ an approximation due to Mislevy (and discussed in Chapter 9) that allows the prediction of the average value of a proficiency scale score from item proportion correct scores without actually scaling the individual responses. The prediction of a reading proficiency mean score (RP) based on a measure of average item proportion correct (P) is

$$RP = 250.5 + 50*(b - (1.7A)^{-1} \ln((1-P)/(P-c))) \qquad (1)$$

where b=.0241 and c=.197 are average item parameters based on the trend items, A=.978 is a constant specific to the age class, and P is the transformation into the proportion correct metric of the average of the logits of item p-values.

We first apply the formula to the data as they stand. The first step is the computation of item-level p-values, handling the nonresponse categories as they are handled in the scaling process so that the I Don't Know and Omit categories are given fractional credit and the Not Reached category is removed from the computation. Thus, the item-level p-value for the i^{th} item in the j^{th} year (j=15,17) is computed as

$$p_{ij} = (C_{ij} + .2*(IDK_{ij}+Omit_{ij})) / (C_{ij}+W_{ij}+IDK_{ij}+Omit_{ij}) \qquad (2)$$

where C_{ij}, W_{ij}, IDK_{ij} and $Omit_{ij}$ are the item-level proportions of correct, incorrect, I Don't Know, and Omitted responses, respectively. The constant .2 is the typical amount of fractional credit given for the Omit and I Don't Know choices and is equal (nearly) to the average value of the "guessing parameter" c.

The next step is the computation of the average of the logits of the p-values given in (2) for each of the two assessment years followed by a transformation back into the proportion correct metric to produce P. The resultant values of P are then substituted into equation (1) to produce the estimated average reading proficiency score RP. The results are as follows:

	Ave Logit	P	RP
1984	1.442	.809	286.7
1986	1.191	.767	278.6

The resulting estimates of average reading proficiency scores differ slightly from the actual values--the 1984 estimate is somewhat lower than the actual value of 288.8 and the 1986 estimate is somewhat higher than the actual value of 277.4. However, the estimates are good enough for our purposes, which are to estimate the effect of changes in nonresponse on the mean proficiency value.

We shall do this by computing a new value of RP for the 1986 data under the following assumptions:

a) the rate of nonresponse to item i in 1986 corresponds to the 1984 data, as do the relative rates of the various types of nonresponse.

b) for the respondents to an item, the proportion of correct answers to that item corresponds to the 1986 data.

That is, we compute

$$C*_{i,17} = (C_{i,17}/R_{i,17}) * R_{i,15}$$

where $R_{i,17} = C_{i,17} + W_{i,17}$ and $R_{i,15}$ is similarly defined and then compute

$$p*_{i,17} = (C*_{i,17} + .2(IDK_{i,15}+Omit_{i,15})) / T_{i,15} \qquad (3)$$

where $T_{i,15} = R_{i,15} + IDK_{i,15} + Omit_{i,15}$. We then proceed as before, computing an average logit, then a P, and finally a value for RP. The results are as follows:

Ave Logit	P	RP
1.239	.775	280.1

The effect of changing the amount and kind of nonresponse on the estimated average reading proficiency score is to raise that score by 1.5 units, reducing the magnitude of the decline in scale scores between 1984 and 1986 by around 20 percent. The bottom line is that the changes in the patterns of nonresponse might have plausibly had some effect on the scale scores but that this effect is not large enough, in itself, to explain the reading decline.

Finally, it is of some interest to consider how the patterns of response for the two assessments compare for the younger students: students of ages 9 and 13. The pertinent data appear in Table 10-3 for age 13, and Table 10-4 for age 9. The tables show that the response patterns for ages 9 and 13 differ from each other as well as from the response patterns for age 17. The major change in response patterns for the 13-year-olds is an increase in the rate of not reaching items coupled with a decrease in the proportion of correct responses and I Don't Know responses. In contrast, the proportion of I Don't Know responses increased between the two assessments for the 9-year-olds, as did the rate of not reaching items. The largest change is a decrease in the percent of correct responses of 5 percentage points.

We can estimate the potential effect of these changes in patterns of nonresponse on proficiency values by proceeding as above. We first estimate the average proficiency by age and year by computing the item-level P-values as in equation (2), converting the within year and age average of the logits of these P-values back into the proportion correct metric to produce P and then obtaining the estimate of RP via equation (1) using values of a, b, and c (appearing in Chapter 9) appropriate to the given age. We next recompute RP for 1986 under the assumption that the pattern of nonresponse was the same as in 1984 by computing the item-level P-values equation (3) and then proceeding from there. The results are as follows:

	Age 9		Age 13	
	P	RP	P	RP
1984	.68	214.20	.71	265.20
1986	.65	209.65	.71	265.20
Difference	-.03	-4.55	.00	0
1986 with 1984 nonresponse	.66	211.15	.70	263.65
Difference (1986 with 1984 nonresponse minus 1984)	-.02	-3.05	-.01	-1.55

As was the case with the age 17 estimates of RP, the age 9 and age 13 estimates differ slightly from the actual values. (The actual values for age 9 are 212.9 in 1984 and 207.3 in 1986 resulting in a decline of 5.6 units. The actual values for age 13 are 258.0 in 1984 and 260.4 in 1986 resulting in an increase of 2.4 units). However, the estimates of RP are good enough to allow for the assessment of the effect of changes in pattern of response on proficiency. The effect of changing the amount and kind of nonresponse on the estimated average reading proficiency for 9-year-old students in 1986 is to raise that value by 1.5 units from 209.65 to 211.15, reducing the magnitude of the decline by around 33 percent. The effect of changing the amount and pattern of response on the estimated age 13 1986 proficiency value is to reduce that value by 1.55 units.

The end result is that changes in the patterns of nonresponse might plausibly have some effect on the changes in scale scores for the three ages but that these effects are not large enough, by themselves, to explain more than a fraction of the reading decline for the 9- and 17-year-olds.

Table 10-3

Average Percents by Year and Demographic Subgroups of Students
Selecting Each of the Categories:
Correct, Incorrect, I Don't Know, Omit, and Not Reached

a) Age 13 in 1984

	Correct	Incorrect	IDK	Omit	Not reached
TOTAL	64.60	26.21	6.12	0.46	2.61
<MODAL GRADE	54.37	32.78	7.97	0.47	4.41
AT MODAL GRADE	69.22	23.23	5.30	0.46	1.79
>MODAL GRADE	76.17	22.69	1.13	0.00	0.00
LOW METRO	54.89	30.87	7.08	0.26	6.90
HIGH METRO	72.71	20.71	4.52	0.43	1.63
PARENT:					
NOT GRAD HS	55.32	32.09	7.81	0.38	4.39
GRAD HS	62.40	28.20	6.19	0.45	2.76
MORE THAN HS	70.17	22.68	4.52	0.26	2.37
GRAD COLLEGE	70.38	22.87	4.75	0.49	1.51

b) Age 13 in 1986

	Correct	Incorrect	IDK	Omit	Not reached
TOTAL	63.76	26.17	4.83	0.26	4.98
<MODAL GRADE	53.68	31.40	6.09	0.29	8.53
AT MODAL GRADE	68.77	23.63	4.19	0.24	3.17
>MODAL GRADE	72.78	18.35	5.36	0.00	3.50
LOW METRO	50.35	31.56	7.37	0.33	10.39
HIGH METRO	71.62	22.25	3.13	0.06	2.94
PARENT:					
NOT GRAD HS	55.63	28.81	6.78	0.36	8.42
GRAD HS	60.78	28.51	5.13	0.34	5.23
MORE THAN HS	68.21	24.23	4.19	0.16	3.21
GRAD COLLEGE	68.84	23.81	3.42	0.21	3.72

c) Age 13: 1986 - 1984 Differences

	Correct	Incorrect	IDK	Omit	Not reached
TOTAL	-0.84	-0.04	-1.29	-0.20	2.37
<MODAL GRADE	-0.69	-1.38	-1.88	-0.18	4.12
AT MODAL GRADE	-0.45	0.40	-1.11	-0.22	1.38
>MODAL GRADE	-3.39	-4.34	4.23	0.00	3.50
LOW METRO	-4.54	0.69	0.29	0.07	3.49
HIGH METRO	-1.09	1.54	-1.39	-0.37	1.31
PARENT:					
NOT GRAD HS	0.31	-3.28	-1.03	-0.02	4.03
GRAD HS	-1.62	0.31	-1.06	-0.11	2.47
MORE THAN HS	-1.96	1.55	-0.33	-0.10	0.84
GRAD COLLEGE	-1.54	0.94	-1.33	-0.28	2.21

Table 10-4

Average Percents by Year and Demographic Subgroups of Students Selecting Each of the Categories:
Correct, Incorrect, I Don't Know, Omit and Not Reached

a) Age 9 in 1984

	Correct	Incorrect	IDK	Omit	Not reached
TOTAL	60.57	23.32	10.13	0.56	5.42
<MODAL GRADE	42.83	29.61	16.86	0.96	9.73
AT MODAL GRADE	66.37	21.28	7.92	0.43	4.00
>MODAL GRADE	82.98	15.03	1.48	0.19	0.32
LOW METRO	48.58	29.16	12.89	0.88	8.50
HIGH METRO	70.67	17.72	6.59	0.78	4.24
PARENT:					
NOT GRAD HS	50.30	31.32	11.34	0.43	6.62
GRAD HS	60.10	25.02	9.09	0.58	5.21
MORE THAN HS	60.76	24.98	8.54	0.97	4.75
GRAD COLLEGE	68.78	21.01	6.09	0.40	3.72

b) Age 9 in 1986

TOTAL	55.48	24.03	11.86	0.50	8.13
<MODAL GRADE	41.58	27.72	18.34	0.66	11.70
AT MODAL GRADE	62.55	22.17	8.52	0.42	6.33
>MODAL GRADE	75.50	13.29	4.34	0.97	5.91
LOW METRO	39.49	29.20	18.45	0.81	12.04
HIGH METRO	66.11	20.94	7.35	0.34	5.26
PARENT:					
NOT GRAD HS	43.41	28.58	17.73	0.69	9.60
GRAD HS	51.54	27.69	11.62	0.54	8.62
MORE THAN HS	63.15	23.70	7.00	0.58	5.57
GRAD COLLEGE	62.95	22.69	7.37	0.39	6.59

c) Age 9: 1986 - 1984

TOTAL	-5.09	0.71	1.73	-0.06	2.71
<MODAL GRADE	-1.25	-1.89	1.48	0.30	1.97
AT MODAL GRADE	-3.82	0.89	0.60	-0.01	2.33
>MODAL GRADE	-7.48	-1.74	2.86	0.78	5.59
LOW MODAL	-9.09	0.04	5.56	-0.07	3.54
HIGH METRO	-4.56	3.22	0.76	-0.44	1.02
PARENT:					
NOT GRAD HS	-6.89	-2.74	6.39	0.26	2.98
GRAD HS	-8.56	2.67	2.53	-0.04	3.41
MORE THAN HS	2.39	-1.28	-1.54	-0.39	0.82
GRAD COLLEGE	-5.83	1.68	1.28	-0.01	2.87

Chapter 11

BLOCK AND BOOKLET ANALYSES FOR AGE 17

Rebecca Zwick

In 1986, NAEP assessment booklets consisted of three blocks of cognitive items, as well as a block of background and attitude items.[1] At age 17, there were 35 assessment booklets[2] that contained at least one of six possible blocks of reading items. A series of analyses was conducted to determine whether the apparent drop in reading proficiency at age 17 was associated with certain item blocks or booklets. In particular, it was of interest to determine whether reading results were affected by the position of the reading block within a booklet or by the subject area of the accompanying blocks.

Table 11-1 shows, for each of the 35 assessment booklets, the estimated percent of 17-year-olds in 1986 with reading proficiency levels less than 200 (Basic) on the 1984 scale. (These are weighted estimates, based on the first of the five imputed values associated with each respondent. Estimates computed without sampling weights were very similar. The analysis reported here produced results that differed somewhat from those described in Chapter 5, which were based on maximum likelihood estimates of proficiency. Also, note that the number of students for whom imputed values are available is larger than the number of students for whom maximum likelihood estimates are available.) The number of students who received each booklet is also given, as well as the booklet composition. For instance, Table 11-1 shows that Booklet 7 consists of two reading blocks (R6 and R4), followed by a computer competence block (C1).

Before discussing the variation across booklets, it is important to compare the overall findings with those from 1984. Combining results across booklets, it is estimated that in 1986, about seven percent of 17-year-olds had proficiency values less than 200. The estimated percents of students below 200 ranged from 3.2 to 12.1 across booklets. A similar analysis of the BIB sample at age 17 in 1984 yielded an overall estimate of less than two percent below 200, with estimated percents ranging from 0 to 3.8 across 56 booklets.

Clearly, the percents below 200 for 1986 were both higher and more variable than those from 1984. Further analyses were conducted to explore

[1] An exception is the booklet administered to the Language Minority sample, which consisted of a reading block, a mathematics block, and a special block of attitude items, in addition to the ordinary set of background and attitude items. The age 17 Language Minority booklet is not included in the analyses described here.

[2] excluding the Language Minority booklet.

Table 11-1

Estimated Percent of 17-year-olds in 1986 with Reading Proficiency Below 200[a]

Booklet Number	Booklet Composition[b]	Number of Respondents	Estimated Percent Below 200
7	R6, R4, C1	270	5.5
9	R4, M5, M7	270	8.8
13	M11, R3, M10	267	8.0
14	R2, M4, M2	297	6.3
21	C6, R2, C1	254	3.3
22	C5, C3, R6	262	7.4
24	M6, R5, C3	281	4.5
26	R5, R2, C4	268	4.3
28	R2, C5, S4	278	4.8
33	C5, C2, R4	280	11.8
34	M8, R1, S5	285	6.5
35	C2, C4, R6	273	7.6
36	M1, R2, M9	270	4.0
38	C1, C3, R1	281	6.7
42	R3, C2, M7	279	3.2
47	S1, S11, R1	286	5.6
50	R1, R3, C5	286	3.4
51	S7, S8, R6	290	5.4
54	C4, C6, R1	285	5.3
61	R1, R5, R4	288	3.3
62	S2, S4, R5	281	5.5
68	S3, R6, M3	291	7.8
75	C6, C5, R5	272	5.7
77	C1, C2, R5	287	5.2
79	R4, C6, S1	276	6.4
81	R3, R4, R2	272	6.2
85	R5, R6, R3	269	7.7
86	S5, S9, R3	272	4.7
88	S10, S6, R4	266	12.1
89	R6, R1, R2	279	4.8
90	C4, C3, R3	279	7.9
92	R4, H1, L1	1964	5.6
93	H2, R4, L2	1931	8.2
94	L3, R4, H3	1968	7.9
95	L4, H4, R4	1961	8.8
Total		16418	6.8

[a] in the 1984 metric

[b] C = computer competence, H = history, L = literature, M = mathematics, R = reading, S = science.

the reasons for the differences among the 1986 booklets. The two booklets with the largest percents below 200 both included reading block R4 as the third of the cognitive blocks. In booklet 33, R4 occurred after two computer competence blocks and in booklet 88 after two science blocks. To determine how reading results are affected by the position of a block and the nature of the blocks that accompany it, a further analysis was conducted: For each of the six reading blocks administered at age 17 in 1986, the average percent correct was computed separately for each of the following cases: (a) the reading block in question was the first of the three cognitive blocks in the booklet, (b) the reading block was the second of the three cognitive blocks, following another reading block, (c) the reading block was the second block, following a block in a subject area other than reading, (d) the reading block was the third block, following two other reading blocks and (e) the reading block was the third block, following two blocks in an area other than reading. (In all but one booklet in category e, the preceding blocks were either two computer competence blocks or two science blocks. See Table 11-1.) All 35 booklets fit into one of these five categories, since there were no booklets in which a reading block was in the third position, following one reading and one non-reading block. The results of this analysis are shown in Table 11-2. The only reading block that shows substantial variation across categories in the average percent correct is R4. For R4, the average percent correct is about 72 when the block is in the first position (N=2510), compared to about 67 percent when the block is in the third position, following two non-reading blocks (N=2507.) In no other case is the difference between the a and e categories more than two percent. It is possible that R4, which is the only block with more than one reading passage over a page long, is more susceptible to position and context effects.

Although it appears from Table 11-2 that context and position effects did not, in general, have a major impact, an analysis was conducted to determine how much the drop in item percents correct between 1984 and 1986 would be reduced if the students who received category e booklets were excluded. For the 39 items that were administered to 17-year-olds both in 1984 and 1986, percents correct were obtained for the 1984 sample, for the entire 1986 sample, and for the 1986 sample, excluding category e booklets. Results are shown in Table 11-3.[3] Based on the full samples, the average drop in percents correct was 3.3; the exclusion of category e booklets had only a tiny impact, reducing the drop to 3.0.

In summary, these analyses showed that only for one reading block (R4) did performance seem to be substantially affected by the position of the

[3] Although the analyses of this chapter are based on 17-year-olds only, lists of percents correct for reading trend items for the bridge samples at ages 9 and 13 are given in Tables 11-4 and 11-5 for comparison purposes.

Certain of the items changed slightly between 1984 and 1986. These items are as follows: N002801 (ages 13 and 17) and N007401 (age 17)--slight change in one distractor; N003101 (ages 13 and 17)--typographical error in stem; N008201-N008205 (ages 13 and 17)--typographical error in passage.

block within a booklet and the content area of the accompanying blocks. Furthermore, even when students who received the most disadvantageous configuration of item blocks in 1986 (i.e., one reading block following two non-reading blocks) were excluded, the drop in reading performance between 1984 and 1986 was virtually unaffected.

Table 11-2

Average Percents Correct (Weighted) for 1986 Reading Blocks
Age 17

	Reading Block											
	R1		R2		R3		R4		R5		R6	
	Wtd. %	N	Wtd. %	N	Wtd. %	N	Wtd. %	N	Wtd. %	N	Wtd. %	N
a Position 1	87.2	574	61.4	575	73.7	551	72.1	2510	68.3	537	71.9	549
b Position 2 after reading	86.6	279	61.8	268	73.2	286	71.6	542	68.4	288	70.4	269
c Position 2 after other	84.6	285	61.1	524	70.1	267	69.6	3899	69.4	281	71.8	291
d Position 3 after 2 reading	----	---	59.8	551	72.1	269	70.1	288	----	---	----	---
e Position 3 after 2 other	86.8	852	----	---	71.9	551	67.3	2507	68.6	840	70.1	825

NOTE

(1) mutually exclusive and exhaustive

(2) e: except in one core, preceding blocks were 2C or sS

Table 11-3

Percents Correct[a] for Reading Trend Items for Age 17 BIB Sample

NAEP Item ID	1984	1986	1986-1984	1986[b]	1986[b]-1984
N001501	96.7	94.3	-2.4	93.5	-3.2
N001502	89.2	84.8	-4.4	84.0	-5.2
N001503	92.4	89.6	-2.8	89.5	-2.9
N001504	90.4	87.2	-3.2	87.3	-3.1
N002001	78.1	72.3	-5.8	72.0	-6.1
N002002	81.4	75.7	-5.8	75.2	-6.3
N002003	86.3	83.4	-3.0	82.6	-3.8
N002801	94.9	92.8	-2.0	92.9	-2.0
N002802	96.6	93.9	-2.7	93.6	-3.0
N003001	47.2	38.8	-8.3	38.8	-8.3
N003003	22.4	30.2	7.8	30.2	7.8
N003101	94.0	87.4	-6.6	88.7	-5.3
N003102	88.7	85.7	-3.0	85.9	-2.8
N003201	91.6	88.2	-3.4	89.3	-2.3
N003202	85.6	81.3	-4.3	80.6	-5.0
N003203	74.0	70.4	-3.6	72.2	-1.8
N003204	83.9	79.8	-4.1	79.1	-4.8
N004601	70.7	68.5	-2.2	69.2	-1.4
N004602	80.3	80.0	-0.3	80.6	0.3
N004603	88.2	86.4	-1.7	87.2	-1.0
N005001	39.6	41.8	2.2	41.8	2.2
N005002	51.2	55.0	3.7	55.0	3.7
N005003	31.6	29.0	-2.7	29.0	-2.7
N007301	87.1	74.5	-12.6	74.7	-12.4
N007302	63.9	62.5	-1.4	63.0	-0.9
N007303	78.5	70.6	-7.9	71.5	-7.0
N007304	77.4	68.7	-8.8	69.9	-7.6
N007305	65.9	59.6	-6.3	60.7	-5.2
N007306	76.9	70.4	-6.4	71.7	-5.2
N007401	63.0	67.0	3.9	67.6	4.6
N007402	85.1	83.6	-1.5	84.3	-0.8
N007403	80.0	75.6	-4.3	76.4	-3.6
N007404	76.0	73.9	-2.1	74.9	-1.1
N007405	41.1	40.7	-0.4	41.0	-0.1
N008201	92.7	86.2	-6.5	86.2	-6.5
N008202	78.2	75.5	-2.7	75.5	-2.7
N008203	85.5	83.5	-2.0	83.5	-2.0
N008204	86.8	80.0	-6.8	80.0	-6.8
N008205	86.0	81.3	-4.7	81.3	-4.7
Mean	76.4	73.1	-3.3	73.3	-3.0

[a] Percent correct = R/(R+W+IDK), where R, W, and IDK are the weighted frequencies of right, wrong, and I Don't Know responses.

[b] Excluding students who received a reading block in the third position, following two non-reading blocks.

Table 11-4

Percents Correct[a] for Reading Trend Items for Age 9 Bridge Sample

NAEP Item ID	1984	1986	1986-1984
N001501	81.5	77.1	-4.4
N001502	54.0	51.9	-2.1
N001503	69.2	66.3	-2.9
N001504	61.2	58.9	-2.3
N002001	31.6	34.5	3.0
N002002	38.2	38.0	-0.2
N002003	42.8	42.8	0.0
N002801	56.5	50.7	-5.8
N002802	58.0	57.3	-0.7
N003101	55.0	53.2	-1.8
N003102	37.2	38.1	0.9
N004101	68.3	59.9	-8.4
N008601	68.1	61.9	-6.2
N008602	58.9	54.1	-4.8
N008603	65.6	59.0	-6.6
N008901	70.3	71.2	0.8
N008902	72.3	71.7	-0.6
N009401	76.1	73.6	-2.5
N009801	93.7	91.7	-2.0
N010201	87.6	81.3	-6.3
N010301	85.7	83.4	-2.3
N010401	72.3	64.6	-7.7
N010402	37.7	38.5	0.8
N010403	26.2	22.6	-3.6
N010501	82.8	76.4	-6.4
N010502	68.2	65.4	-2.8
N010503	81.0	76.3	-4.7
N010504	72.3	57.9	-14.4
N013301	83.6	85.4	1.7
N014201	69.3	61.1	-8.3
Mean	64.2	60.8	-3.4

[a] Percent correct = R/(R+W+IDK), where R, W, and IDK are the weighted frequencies of right, wrong, and I Don't Know responses.

Table 11-5

Percents Correct[a] for Reading Trend Items for Age 13 Bridge Sample

NAEP Item ID	1984	1986	1986-1984
N001501	93.9	96.0	2.1
N001502	79.3	78.6	-0.6
N001503	85.0	89.0	4.0
N001504	81.5	82.1	0.5
N002001	63.4	67.0	3.6
N002002	68.3	67.2	-1.1
N002003	74.5	76.0	1.5
N002801	86.3	89.2	2.8
N002802	90.4	91.2	0.8
N003001	31.2	25.2	-6.0
N003003	9.9	9.8	-0.1
N003101	85.3	83.5	-1.9
N003102	75.8	78.2	2.5
N004601	58.8	59.4	0.6
N004602	70.2	66.3	-3.9
N004603	82.9	76.9	-6.0
N005001	22.6	24.2	1.6
N005002	36.1	44.8	8.7
N005003	22.1	21.3	-0.8
N008201	86.9	86.2	-0.7
N008202	65.4	63.5	-1.9
N008203	78.7	79.5	0.8
N008204	76.5	74.7	-1.8
N008205	76.4	74.7	-1.7
Mean	66.7	66.8	0.1

[a] Percent correct = R/(R+W+IDK), where R, W, and IDK are the weighted frequencies of right, wrong, and I Don't Know responses.

Chapter 12

THE CHALLENGER HYPOTHESIS

Albert E. Beaton

On January 26, 1986, the American space program suffered a severe setback when the Challenger spacecraft exploded shortly after takeoff, killing its crew, including a New Hampshire elementary school teacher. The accident occurred at 11:39 A.M., Eastern Standard Time, during a space launch that appeared on live national television. All America was distressed by this tragedy.

At the time of the accident, NAEP was assessing the 9-year-old students who were in the 1986 trend sample. The hypothesis addressed here is whether or not the Challenger accident could have resulted in sufficient disruption of the assessment activities or the emotional responses of the children so as to lower their performance on NAEP.

If this hypothesis were true, we would expect that the disruption would affect the students on the day that the accident occurred, the next day, and, perhaps, for the rest of the week. We would expect a substantially greater number of low scorers during that time period. Time zone differences would also explain why the eastern region of the country declined less, since these students would have heard the news later in their school day than students in other time zones.

Our data do not support this reason for the decline in estimated reading proficiency. To investigate this hypothesis, we organized the NAEP results for the 9-year-old student sample by date of testing. The results are shown in Table 12-1. This table does not show any particularly unusual number of low-scoring students on the day of the accident or on the following days. We, therefore, conclude that the Challenger accident did not substantially affect the NAEP results.

We know of no other external event that might have affected NAEP results.

Table 12-1

NAEP 1986 Age 9 Bridge Sample
Reading Proficiency Scores* by Date of Administration

Date	1986	Wtd N	Percent Below 100	Percent Below 200
Mon	1/6	37627	4.37	38.85
Tue	1/7	247957	7.57	41.14
Wed	1/8	657563	10.51	50.68
Thu	1/9	617180	3.94	35.47
Fri	1/10	130353	2.28	27.73
Mon	1/13	172819	7.84	51.15
Tue	1/14	769781	7.98	44.01
Wed	1/15	933525	5.11	40.00
Thu	1/16	685917	4.97	38.43
Fri	1/17	262460	5.84	36.25
Sat	1/18	75262	8.33	42.14
Mon	1/20	149673	4.07	29.86
Tue	1/21	422720	3.45	40.08
Wed	1/22	783421	6.92	41.59
Thu	1/23	841407	6.03	27.80
Fri	1/24	287559	8.33	48.60
Mon	1/27	168437	6.21	39.88
Tue	1/28	823264	5.40	36.14
Wed	1/29	488895	4.97	42.61
Thu	1/30	343123	7.75	48.23
Fri	1/31	336649	6.43	44.40
Total :		9235591	6.20	40.04

* Maximum likelihood estimates of reading proficiency were used.

Chapter 13

SUMMARY AND CONCLUSIONS

Albert E. Beaton

The quest for finding a clear, correctable problem that explains the estimated decline in reading proficiency of 9- and 17-year-olds has failed. We remain unconvinced that declines of the magnitude estimated have actually occurred, although smaller declines may well have. We continue to believe that publishing the 1986 reading data without corroborating evidence of a major decline would be irresponsible.

The quest has brought about, however, a very thorough and intensive review of the entire assessment process. The purpose of this chapter is to summarize what we have learned and explains the next steps in exploring the mystery associated with the 1986 reading data.

The intensive review has led us to reject several of the hypotheses that we explored. We conclude that:

- The population from which NAEP sampled did not change enough to produce a sharp decline in reading proficiency and that the NAEP sample is a good representation of the population.

- The existing quality control procedures assure us that the student data in the NAEP files accurately reflect the actual student responses.

- The estimated decline is not an artifact of the procedures used to develop NAEP's new reporting scales.

The results from exploring the other hypotheses are not as clear and we are left with the conclusion that some of the seemingly minor changes in design and execution may have had, singularly or jointly, a substantial effect on the estimated proficiency of the students. For example, the change in the type size or the change to machine scorable responses may have had an effect on performance, but we are unable to estimate the size of the effect from the available data.

Therefore, we have designed an experiment to gather more data about the changes in the design and execution on NAEP. The experiment is included in the NAEP 1988 data collection which is now in progress.

The experiment is as follows. At each of the three age levels on which the trend data are based, three samples of students have been drawn:

- 1984E samples. These samples duplicate the 1984 assessment as closely as possible. They are defined in the same way as was the equivalent 1984 sample, administered copies of the same booklets that were used in 1984, and the dates of testing and administrative procedures are as close as possible to those used in 1984.

- 1986E samples. These samples duplicate the 1986 sample as closely as possible, using the same sample definitions, booklets, administrative procedures, and, as closely as possible, the same dates of testing.

- 1988 samples. These samples are from the NAEP 1988 assessment which is in progress. In 1988, reading, writing, civics, and history are being assessed. The administration of these samples will proceed as designed. The design for the 1988 assessment is described in Realizing the Nation's Report Card: A Proposal in Response to Grant Announcement CFDA 84.199 (ETS, 1987).

In principle, these samples are randomly equivalent and have the same average reading proficiency except for sampling error, which is estimable. Comparing these samples will allow us to estimate the joint effects of the changes in design and execution.

When the data become available, we will estimate the reading proficiency of each sample separately and then compare results. We would expect one of the following outcomes:

- The three sets of samples may result in similar estimates of the distribution of reading proficiency. In this case, we would conclude that the effect of the changes in design and execution of NAEP were indeed minor and that a large decline in reading proficiency did in fact occur in 1986. Whether the decline was peculiar to 1986 or continued into 1988 would depend on whether the three similar sets of data collected in 1988 are closer to the actual 1984 or 1986 distributions.

- The estimated distributions of reading proficiency from the 1986E samples may be lower than the estimated distributions from the 1984E samples at ages 9 and 17 but similar at age 13. In this case, we would conclude that the estimated decline in 1986 was indeed a function of the changes in design and execution of NAEP. From comparison of the distributions of reading proficiency, an equation could be developed to adjust the 1986 scores for their disadvantage, and this equation then applied to the 1986 samples to give improved estimates of student proficiency in 1986.

  Comparison of the 1988 samples with the 1984E and 1986E samples will measure the effect of any changes introduced in 1988. If substantial effects occur, they can be measured and the results adjusted to be comparable to past trend data.

We do not know at this time how much the changes in design and execution affected the NAEP results, or if they are affected at all. It is quite possible, however, that the results will show some changes due to design and execution but also some change in the performance of students. This experiment gives us an opportunity to fairly estimate that change, first in the 1986 procedures and also in 1988.

APPENDIX A

PARTITIONING ANALYSIS

Albert E. Beaton

In many research situations, it is of interest to compare average values computed from different samples. For example, the average SAT score declined annually for a number of years, and it is of interest to establish whether the decline is attributable to changes in the population of students choosing to sit for the examination or to differences in the performance of the student populations. Another example is the National Assessment of Educational Progress (NAEP) where fluctuations in average performance may be due to either changes in the mix of students in the nation's schools or differences in student proficiency. Partitioning analysis was designed to address such questions by allocating the differences between means to components attributable to population shifts, to changes in performance, and to the interaction of population and performance.

* * *

Let us assume that there are two samples (t=1,2) of subjects which are measured on some variable x. There are N_t subjects in each sample. The averages of the samples are denoted x_t. The difference between the two samples $D=x_2-x_1$ is the statistic to be examined.

Now, let us assume that both samples can be separated into the same set of K mutually exclusive and exhaustive categories. The number of subjects in category k (k=1,2,...,K) for a sample is N_{tk} $(\Sigma_k N_{tk}=N_t)$. The average values for the samples may be written

$$x_1=p_{11}x_{11}+p_{12}x_{12}+\ldots+p_{1K}x_{1K} \quad \text{and}$$

$$x_2=p_{21}x_{21}+p_{22}x_{22}+\ldots+p_{2K}x_{2K}$$

where $p_{tk}=N_{tk}/N_t$ and x_{tk} is the average value of x for the kth category of the t-th sample. The difference between these means may be written

$$D=(p_{21}x_{21}-p_{11}x_{11})+(p_{22}x_{22}-p_{12}x_{12})+\ldots+(p_{2K}x_{2K}-p_{1K}x_{1K})$$

Letting the differences between the proportions in the K cells be $d_{pk}=p_{2k}-p_{1k}$ and the differences between the averages of x in the cells be $d_{xk}=x_{2k}-x_{1k}$, D may be written

$$D=(p_{11}+d_{p1})(x_{11}+d_{x1})-p_{11}x_{11}+(p_{12}+d_{p2})(x_{12}+d_{x2})-p_{12}x_{12}+\ldots$$
$$+(p_{1K}+d_{pK})(x_{1K}+d_{xK})-p_{1K}x_{1K}$$

and after simplification

$$D=x_2-x_1=\Sigma_k(p_{1k}d_{xk}+d_{pk}x_{1k}+d_{pk}d_{xk})$$

The difference between the two means may therefore be partitioned into 3K parts of the following forms,

$\delta_{1k}=p_{1k}d_{xk}=p_{1k}(x_{2k}-x_{1k})$ which represents the difference in the average values attributable to changes in performance.

$\delta_{2k}=d_{pk}x_{1k}=(p_{2k}-p_{1k})x_{1k}$ which represents the difference in the average values attributable to changes in population. And,

$\delta_{3k}=d_{pk}d_{xk}=(p_{2k}-p_{1k})(x_{2k}-x_{1k})$ which represents the interaction between performance and population.

These components of D are completely additive.

An example of partitioning analysis may show the usefulness of this technique. Table 6-3 of Chapter 6 contains a summary of the NAEP data collected on two national samples of 17-year-olds. The first sample was collected in 1984 (t=1) and the second sample in 1986 (t=2). Both samples were measured in reading. The results are shown by four regions (K=4) of the country and for the country as a whole.

The columns of Table 6-3 contain the following:

- subgroup: the grouping into which the mean difference is partitioned. In this case, the grouping is the four regions of the country.
- p1: the proportion of the sample in each subgroup in 1984. We note that the sample is distributed quite equally among the regions.
- p2: the proportion of the sample in each subgroup in 1986.
- DEL P: the differences between the proportions in p1 and p2. We note that there is little change in population distribution between the two years, although the

Northeast and Southeast compose a slightly smaller segment of the population.

- XBAR1: the average reading value for each subgroup in 1984. The TOTAL line presents the national average for the first sample which is an average of the regional means weighted by the subgroup size. This national average is not affected by the choice of partitioning categories.

- XBAR2: the average reading value for each subgroup in 1986. The TOTAL line is computed in the same way as that line for XBAR1.

- DEL XBAR: the differences between the average values in XBAR2 and XBAR1. (The value in the TOTAL line is not easily interpretable.)

- T: the total contribution of a subgroup to the average difference. The value in the TOTAL line is the sum of the values in this column and is algebraically equal to the difference between the two national means.

- PERF: the contribution of each subgroup due to change in performance, i.e., δ_{1k}. The value in the TOTAL line is the simple sum of the elements in the column.

- POP: the contribution of each subgroup due to change in population, i.e., $\delta 2k$. The value in the TOTAL line is the simple sum of the elements in the column.

- INTER: the contribution of each subgroup to the interaction of performance and population, i.e., $\delta 3k$. The value in the TOTAL line is the simple sum of the elements in the column.

From inspection of the partitioning table, we see that:

1) The change in average performance between 1986 and 1984 was 277.4-288.8=-11.4 points on the reading proficiency scale.

2) The component of the mean difference due to change in performance is -10.6518, somewhat less than the -11.4 mean difference change that is being investigated. This figure shows that, if the second sample were distributed into subgroups in exactly the same way as the first but members of the subgroups differed in performance as they did in the two samples, then the changes in performance in the individual cells would result in a mean difference of -10.6518.

Looking at individual elements of the PERF column, we see that the performance of all areas of the country contributed to the decline, with the largest contribution in the West (-3.7236) and the smallest in the Northeast (-0.8608).

We note that the total performance change can be viewed as the difference between two standardized means, where the first sample is considered a "standard" sample and the second sample is adjusted to that standard.

3) The component of the mean difference due to population shift is +0.0324, which is at the bottom of the POP column. This indicates that, if students in subgroups continued to perform as they did in 1984 but population shifts occurred as in these samples, then the second national sample would have had a slightly larger mean than the first. Clearly, this does not support the hypothesis that population shifts around the country contributed to the decline in reading proficiency values.

   The elements of the POP column show how population shifts in various regions of the country contributed to the mean difference.

4) The component of the mean difference due to interaction is 0.0196, which is at the base of the column labelled INTER.

5) The elements of the column labelled T show how the various regions contributed to the mean decline. The elements of the T column are the sums of the corresponding elements in the PERF, POP, and INTER columns.

From this partitioning analysis, we would judge that the mean difference is due to changes in performance of students in all areas of the country and that population shifts do not explain the decline.

APPENDIX B

Memorandum for: MR. BEATON

Subject: Reading Performance on NAEP

Date: June 4, 1987

From: W. B. Schrader

At a meeting on April 29, in which Eugene Johnson, Janet Johnson, Rebecca Zwick, and you participated, I learned that a marked shift in percentage correct occurred on common reading items between 1984 and 1986 for 9-year-olds and 17-year-olds but not for 13-year-olds. The comparison at age 17 was based on two BIB samples and at age 9, the comparison was based on a BIB sample and the corresponding bridge sample. Thus, the nature of the evidence makes it improbable that the bridging process or the BILOG analysis are relevant to the question of why the performance shift occurred. Moreover, item analysis results for the common items provide no reason to question the scoring keys.

It occurred to me that if, somehow, a mismatch had occurred between a set of items and the key used to score those items for a subsample of the 17-year-olds, a drop in the percentage correct would arise for those items in the set for which the keyed response was different in the key actually used and the key which should have been used for that set of items. At your suggestion, I met with Norma Norris, Alfred Rogers, and David Freund. We reviewed the various operational stages including printing, assembling the test booklets, scanning, and creating the response file. There seemed to be no way in which a mismatch between items and the identified response in the data file could have been introduced and gone undetected in the subsequent operations. I learned that Jim Ferris had analyzed the percent correct for items included in block R4 for each of the 11 books in which block R4 was included. The results of this analysis offer no support for the hypothesis that a mismatch between item and key occurred for these booklets. It is a matter of some interest that items 7301 through 7306, based on a reading passage in block R4, show the following differences in percent correct when the "Weighted Phonebook" method is used: -12, +1, -8, -8, -6, and -7. Except for item 7302, these differences are substantially higher than the usual difference between years.

One additional way of checking the mismatch hypothesis calls for examining the pattern of responses to each item. If the hypothetical mismatching occurred, there should be a noticeably higher percentage of students choosing one of the "incorrect" options. I learned that Jim Ferris had already reviewed the response patterns and had found that they were reasonable. He had no objection to my taking a second look. I did so for age 17 and found no support for the mismatch hypothesis.

The results for mathematics and science shown in your report "NAEP Year 17 (1985-86): Preliminary Reading Results," dated May 4, 1987, indicate that the selection of schools and students and the self-selection of participating schools and students do not provide a promising source of explanation for the unusual results for reading in 1986.

A printout of preliminary BILOG results as of 5/11/87, made available to me by David Freund, indicates that the differences for 9-year-olds between 1984 and 1986 are noticeably smaller than the differences for 17-year-olds. In evaluating these differences, it would be helpful if the standard error of the difference between means were estimated, particularly for differences in which bridging is involved.

The reading results for age 17 can only be regarded as puzzling. It appears that omits and not reacheds for 1984 and 1986 are so similar as to make a hypothesis of mis-timing or difference in motivation implausible. The only additional analysis that I can think of would be to study within-school means and standard deviations. I understand from Eugene Johnson that values for each student analogous to scores have been calculated. I think that it might be illuminating to compare the shape of the distribution of school means for 17-year-olds in 1986 with that for 1984. Similar comparisons based on the distribution of with-school standard deviations, mean omits, and mean not reacheds should also be worth considering. It is not clear what course of action would be appropriate if aberrant results were found for certain schools. On the other hand, evidence that distributions of these four variables were similar when the distribution for 1984 is compared to the corresponding distribution for 1986 would be relevant to the question of whether administrative procedures or student motivation in particular schools might have contributed to the unusual results for 1986.

APPENDIX C

WESTAT

Memorandum

February 16, 1988

To: Morris Hansen

From: Renee Slobasky

As you requested, Nancy Caldwell and I have reviewed our thinking about factors in the field administration that might have affected the 1986 reading scores of the 9-year-olds and 17-year-olds. It seems to us that the age groups must be discussed separately, since there were differences in the design specifications and resulting administrative procedures by age group.

Assessment of the 9-year-olds

For this age group, as for the 13-year-olds, the organization for the field administration of the 1986 bridge assessment was very similar to 1984. We employed a small group of 16 supervisors each responsible for 4 PSU's. In both years, these supervisors conducted the assessment of the 13-year-olds in the fall and then the 9-year-olds in the winter. Eleven of the sixteen supervisors worked both years, as did the two additional backup supervisors.

It is our understanding that the bridge sessions were used to measure trends for the 9-year-olds and 13-year-olds and that the reading scores of the 13-year-olds did not decline. Further, we understand that the math scores for the 9-year-olds seem to be acceptable and math was assessed at the same time as reading. These factors would seem to indicate that the implementation of the assessments was not a major problem in the decline of 9-year-old reading scores.

The content and timing of the booklets varied from 1984 to 1986. Based on the administration of the booklets by our field staff, the following observations were made:

> Differences in the contents of the booklets. In 1984 and prior years, the booklets contained reading and another similar subject such as writing, literature or art. In 1986, for the first time, the bridge books for the 9-year-olds and 13-year-olds combined reading with mathematics and science (the books for the 17-year-olds combined reading with a variety of different subjects).

In 1984, when reading was paired with writing, our field staff observed that it was difficult to get the students to spend the appropriate time on the writing portion. This was particularly true in the BIB spiral sessions where different books have different sections and it is difficult to make sure that students are working in the correct section. Hence, it is possible that students spent more of their allocated time on reading. With the change in paired subjects in 1986, students may not have been able to spend more than their allocated time on reading.

Differences in the type of administration. In 1984, the sections of the BIB spiral books were timed, but within sections the questions were self-administered, so there was some opportunity for students to finish a section and switch among other sections despite the timing. For the 9-year-olds and 13-year-olds in 1986, the math and science sections were paced by a tape, and the reading section was self-administered as in 1984. So, the students may have been more fully occupied with the math and science sections and may not have had more than the allocated time available to switch back to the reading section.

Length of the background sections. The amount of time devoted to the background block was increased for the 9-year-olds in 1986 but did not change for the other age groups. Our field staff reported that many of the 9-year-olds were frustrated by the amount of time spent on the block of background questions in 1986 (about 15 minutes versus 6 minutes in 1984). Perhaps their concentration or motivation for the reading section was affected.

Assessment of the 17-year-olds

In 1986, the design of NAEP was changed so that the spring assessment, which had involved 17-year-olds only, was expanded to include the main NAEP assessments of the 9-year-olds and 13-year-olds as well. Therefore, the spring 1986 assessment was much larger than any previous spring assessment in terms of the numbers of schools and students involved. The field work was more difficult because all three age groups were being assessed. In addition, a Language Minority Component was added that increased the complexity of the in-school sampling and the booklet distribution procedures.

In recognition of the increased size of the spring assessment, we increased supervisory staff from 18 to 54 and increased home office staff to monitor the field work. This meant that about half of the supervisors did not have previous NAEP experience. Senior home office staff, however, remained unchanged throughout the 1984 and 1986 field administrations. Through quality control observations of the supervisors work in the field, problems were identified and corrective action taken when necessary.

It seems unlikely that the changes to the field organization that resulted from the design changes to the 1986 assessments could have caused the widespread drop in reading scores. In addition, there should have been some effect on math and science scores since these subjects were assessed by the same supervisors in the same schools. It is our understanding that the math scores of the 17-year-olds do not show the drop shown by the reading scores.

Another difference in the assessment of the 17-year-olds is that session size increased from 1984 to 1986. This was not true for the 9-year-old and 13-year-old bridge sessions used to measure trends.

Session size increased for two reasons. First, the maximum allowable sample size in a school was increased from 200 to 250. Second, it was our experience in 1984 that many schools with large numbers of students to be assessed requested or insisted that all students be assessed at the same time to minimize disruption of the school day. With the BIB spiral design, where the books are self-administered rather than tape administered, it is possible to assess large groups of students at one time. In 1986, to maintain high levels of school cooperation, more large sessions were permitted, at the insistence of the schools. Large sessions are more difficult to control and require more monitoring and supervision than small sessions.

We understand that ETS has done some investigation of reading scores and session size and has not found differences in the reading scores of students assessed in small versus large sessions.

There were other changes in the design specifications as discussed in the ETS report. For example, the teacher survey procedures were more complicated in 1986 than in 1984. We received reports from the field that these procedures were at times confusing to the students and some sessions were initially disrupted as a result.

A change in booklet design reported by field staff as confusing to the students related to the numbering of sections. The various sections of the spiral books, which had been numbered consecutively (1, 2, 3, 4) in 1984, were not numbered in 1986. This also made it more difficult to monitor that students were moving correctly from section to section.

REFERENCES

Beaton, A. E. (1987). Implementing the new design: The NAEP 1983-84 technical report. (NAEP Report 15-TR-20) Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1987). Realizing the nation's report card: A proposal in response to Grant Announcement CFDA84.199. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1985). The reading report card: Progress toward excellence in our schools. (NAEP Report 15-R-01) Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1983). Item response models for grouped data. Journal of Educational Statistics, Winter 1983, 8(4), 271-288.

National Assessment of Educational Progress. (1984). Reading objectives: 1983-84 assessment. (No. 15-RL-10) Princeton, NJ: Educational Testing Service.

National Assessment of Educational Progress. (1987). Reading objectives: 1986 and 1988 assessments. (No. 17/19-R-10) Princeton, NJ: Educational Testing Service.

Rogers, W. T., Folsom, Jr., R. E., Kalsbeek, W. D., & Clemmer, A. F. (1977). Assessment of nonresponse bias in sample surveys: An example from National Assessment. Journal of Educational Measurement, 14(4), 297-311.