ED 322 216 TM 015 464

AUTHOR	Beaton, Albert E.; Zwick, Rebecca
TITLE	The Effect of Changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly. Revised.
INSTITUTION	National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	ETS-17-TR-21
PUB DATE	Feb 90
GRANT	OERI-G-008720335
NOTE	248p.; Other contributors to this report are Kentaro Yamamoto, Robert J. Mislevy, Eugene G. Johnson, and Keith F. Rust.
AVAILABLE FROM	National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, Princeton, NJ 08541-0001.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Academic Achievement; Educational Change; *Educational Trends; Elementary School Students; Elementary Secondary Education; Item Analysis; *National Programs; *Reading Tests; Secondary School Students; *Standardized Tests; Statistical Analysis; *Testing Programs; Trend Analysis
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

Results of new research into the anomalous results of the 1986 reading portion of the National Assessment of Educational Progress (NAEP) are reported. The original analysis of the 1986 data indicated that the estimated performance level of 9- and 17-year-old students had dropped dramatically since 1984, whereas the performance of 13-year-olds had increased very slightly. Since it was unlikely that such large changes could have taken place in such a short time, the results were not presented to the general public and publication of the results was suspended until further research into their accuracy could be completed or until other corroborating evidence for the declines could be found. To this end, the design of the 1988 NAEP was modified, and new data were collected and analyzed. The main finding from the 1988 data analyses is that the changes in assessment booklets and procedures that were introduced in 1986 had a substantial and unpredictable effect on the estimates of performance. Although many of the same items were used in both the 1984 and 1986 assessments, student performance on these items differed substantially when the items were administered in different contexts. Discovered differences between the 1986 and 1988 assessments were used in a design to estimate the reading performance of students in the 1986 sample. The new estimates show that reading performance declined slightly in 1986 at all age levels as compared to 1984. The declines, however, are not statistically significant. The new data also indicate a rebound in 1988 from the 1986 levels to about the same level of performance exhibited in 1984. The report contains 8 chapters by Beaton, Zwick, Yamamoto and Mislevy detailing the research and analysing the data. Appendices include various types of statistical and methodological information. Twenty-five graphs and 33 data tables are included. (TJH)

THE NATION'S REPORT CARD

THE EFFECT OF CHANGES IN THE NATIONAL ASSESSMENT:

Disentangling

DISENTANGLING THE NAEP 1985-86 READING ANOMALY

Albert E. Beaton • Rebecca Zwick

In collaboration with—
Kentaro Yamamoto • Robert J. Mislevy • Eugene G. Johnson • Keith F. Rust

with a Foreword by John W. Tukey

REVISED FEBRUARY 1990

This report, No. 17-TR-21, can be ordered from the National Assessment of Educational Progress, Educational Testing Service, Rosedale Road, Princeton, New Jersey 08541-0001.

The contents of this booklet were developed un[illegible]nt from the Department of Education. However, those contents do not necessarily reflect the policy of the Departme[illegible] education, and you should not assume endorsement by the Federal Government.

The work upon which this publication is based was performed pursuant to Grant No. G-008720335 of the Office of Educational Research and Improvement.

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, *ETS*, and ETS® are registered trademarks of Educational Testing Service.

THE EFFECT OF CHANGES IN THE NATIONAL ASSESSMENT: DISENTANGLING THE NAEP 1985-86 READING ANOMALY

Contents

Contents (continued)

ACKNOWLEDGMENTS

This report was produced through the efforts of a team of talented and dedicated professionals. Our profound appreciation goes to all of them. Those who have made especially significant contributions are acknowledged below.

We are particularly indebted to the members of NAEP's Design and Analysis Committee--John Carroll, Robert Glaser, Bert Green, Sylvia Johnson, Robert Linn, Ingram Olkin, Tej Pandey, Richard Snow, and John Tukey--for their invaluable advice and suggestions.

We are grateful to the National Center for Education Statistics of the Office of Educational Research and Improvement, and in particular to Emerson Elliott, David Sweet, Gary Phillips, and Eugene Owen for their support, help, and suggestions. We also appreciate the worthwhile comments provided by the group of anonymous reviewers appointed by the National Center.

Special thanks go to the ETS internal review committee of William Angoff, Gregory Anrig, Nancy Cole, Garlie Forehand, Eugene Johnson, Ann Jungeblut, Archie Lapointe, Samuel Messick, Robert Mislevy, and Ina Mullis for their important contributions and their encouragement. Special thanks also go to reviewers from Westat, including Morris Hansen, Keith Rust, and Renee Slobasky.

In addition, we are indebted to Juliet Shaffer and Fritz Drasgow, two specially appointed external reviewers, for their detailed and thoughtful comments. Kent Ashworth, John Barone, David Freund, Jo-Ling Liang, Terry Salinger, and Kentaro Yamamoto deserve special thanks for their careful reviews and constructive observations.

Essential to the report are statistical programming and the tabular and graphic presentation of NAEP data, which were carried out through the conscientious efforts of David Freund, Maxine Kingston, Edward Kulick, Jo-Ling Liang, Michael Narcowich, Kate Pashley, and Minhwei Wang.

We also express our special gratitude to Debbie Kline, who supervised the preparation of the report and provided valuable editorial assistance. Kent Ashworth and Jan Askew coordinated the production of the publication.

FOREWORD

For decades I have been uncomfortable about behavioral science's failure to take the defects of measurement as seriously as seemed appropriate to one initially trained as a chemist. Appearances of such failures varied from highly quantitative matters like at least thinking about re-expressing observed numbers, to quite qualitative matters like thinking about possibilities of surrogacy. The areas where one kind or another of more thoughtful or more detailed approaches seem to have been relatively common include large-scale psychological and educational testing, market-basket selection for economic indices, and weighting aspects of surveys based on probability samples. Even so, if we had been asked five or more years ago about educational testing, we should have admitted less worry than for other fields, but not *no* worry.

The most difficult--and to some of us the most important--task of NAEP, to assess changes in what well-defined populations of young people can do, requires much more care than traditional educational testing, whether in a single class, or in a single school, or in a nationwide program. Asking about individuals may well require less precision than asking about population means (or medians), since standard errors for individuals are $\sqrt{n}$ times as large (and $\sqrt{n}$ may easily be 50 or more), and thus often dominate measurement uncertainty. Asking about performance, even population performance, at a single time can also be sloppier, since the importance of small deviations along a scale is limited when one does not know what specific values on a scale mean. This need not be true of asking about short-term change, since it may not be too hard to find relatively solid, widely separated anchor points. (We do not need precise knowledge of scale behavior between anchor points, since population measures of change will average individual values spread over a wide range--perhaps including the whole range between the anchor points--and, consequently, will be relatively little affected by different scale shapes--variances, of course, may be affected, but are always estimable, at least in large part.) For these reasons, NAEP has had, since its earliest years, a much greater need for care in measurement than other educational testing programs.

In the report that follows, Beaton spells out very clearly the tension between keeping things the same, to improve measurement of change, and changing them, to take advantage of new innovations. In NAEP's earliest years, in part because of inherent caution among the members of its technical advisory committees, this tension was resolved by strong pressure for keeping everything as nearly the same as seemed possible--though the results reported in Chapter 8 could lead one to wonder whether enough was held constant.

The "changing of the guard," when responsibility for NAEP was transferred to the Educational Testing Service, was marked by "new brooms" and an influx of more conventional psychometric attitudes. Since the "old guard," though concerned with possible effects of moving items, had not really emphasized any concept of "item in context," it is not surprising that the "new guard" succumbed to the superstition that behavior of a well-specified item was not materially influenced by context. Or that the "new guard" was rather more willing to make changes in other detailed aspects of test administration.

We cannot fail to admire the firmness and wisdom shown by the present management of NAEP (assisted by its advisors) when the "1986 reading anomaly" was first detected. For not only was reporting held up (until results could be better understood) but changes were made in the rapidly approaching 1988 assessment to make it possible to gather relevant data that allowed comparisons never before possible.

I can think of no comparable instance of a behavior-measurement program where nearly so careful an analysis was undertaken. I had a hand in an external examination of Alfred C. Kinsey's *Sexual Behavior in the Human Male* (cp. Cochran, et al., 1954) and a worm's-eye view of the external examination of the Coleman Report, but these efforts were not comparable in nature, scope, or detail with the examination discussed in the present account. Specifically,

a) they did not have the advantage of planned supplemental samples--of planned experiment,

b) they were, by far, not as detailed or careful, and

c) they were conducted by outsiders, rather than by those responsible for the data planning, gathering, and analysis.

While we are still far from any ultimate quality of measurement, the analysis in the body of this report, and the lessons that are recognized as having been learned, are a major step forward. As a result, NAEP, in the years ahead, may be the first instance of behavioral measurement of which we may all be proud as a matter of careful measurement.

This does not mean that the challenges to NAEP are ended. It is easy, and I hope ultimately helpful, to put down here a couple of examples of future challenges, so long as we do not consider them to be either exhaustive or representative:

a) Objective selectors and item constructors believe that reporting subscales is important. Is this only a matter of monitoring whether, at some future date, a subscale or two might show distinctive behavior? Or does NAEP measure the subscales separately enough so that we can routinely learn from their separate reporting? (Or are there other subscales that NAEP does measure separately enough which could be reported?)

b) How should performance in different areas--such as mathematics, reading, and science--be compared? Can anything useful be done without bringing in relative expectations? (Could NAEP broaden its focus to gather information on relative expectations, not only of professionals, but of the general public?)

The lessons learned from the study of the reading anomaly are pointed out throughout the whole account that follows. They are focussed most sharply, however, in the Epilogue (Chapter 9). I feel I can best serve the hurried reader by quoting a few of the sharpest and clearest statements:

a) "When measuring change, do not change the measure." (page 165)

b) "The tension between continuity and change is not unique to educational measurement." (page 166)

c) ". . . no measurement is perfect, especially the measurement of changes over time." (page 167)

d) "The identification of technological limitations always presents a challenge for methodological improvement." (page 168)

To these must be added the general principle, of which (b) is a specific consequence:

e) THE BEST WAY TO MEASURE CHANGE IS rarely TO MEASURE TWO LEVELS AS BEST WE CAN SEPARATELY, AND THEN SUBTRACT ONE NUMBER FROM THE OTHER.

The Educational Testing Service, its NAEP scientists, and the authors of this report deserve hearty congratulations from all of us for bravery, insight, stick-to-itiveness, and care in inquiry, and for clarity and honesty of exposition.

JOHN W. TUKEY
Princeton, New Jersey
December 13, 1989

THE EFFECT OF CHANGES IN THE NATIONAL ASSESSMENT:

DISENTANGLING THE NAEP 1985-86 READING ANOMALY

EXECUTIVE SUMMARY

Albert E. Beaton

Since 1969, the National Assessment of Educational Progress (NAEP) has reported what students in American schools, both public and private, know and can do. Over the years, NAEP has assessed student proficiency in reading, writing, mathematics, and science as well as a number of other subject areas. Reading proficiency has been assessed six times--in 1971, 1975, 1980, 1984, 1986, and 1988--and will be assessed again in 1990. NAEP has focused on measuring educational trends, and its long-term trend reports have been based on the assessment of the proficiency of carefully selected national probability samples of 9-, 13-, and 17-year-old students. NAEP has become a respected indicator of progress in American education.

Maintaining the integrity and credibility of NAEP requires the development and careful execution of a complex assessment design and, ultimately, sound professional judgment. The original analysis of the 1986 reading trend data showed anomalous results. The estimated performance level of 9- and 17-year-old students had dropped dramatically from 1984, whereas the performance of 13-year-olds had increased very slightly. Since it was deemed unlikely that such large changes could have taken place in such a short time, it was decided not to present the results to the general public and to suspend publication of the results until further research into the accuracy of the results could be completed or other corroborating evidence for the declines could be found. To collect such additional evidence, the design of the 1988 National Assessment was modified, and the new data have now been collected and analyzed. The purpose of this report is to present the results of this new research into the anomalous results of the 1986 reading assessment.

The main research finding from the study of the 1988 data is that the changes in assessment booklets and procedures that were introduced in 1986 had a substantial and unpredictable effect on the estimates of performance. Although many of the same items were used in both the 1984 and 1986 assessments, student performance on these iters differed substantially when the items were administered in different contexts. The additional data gathered in the 1988 assessment allowed the study of the differences between assessment systems, and the differences that were found were used to reestimate the reading performance of students in 1986.

The new estimates show that reading performance declined slightly in 1986 at all age levels from the 1984 levels. The declines, however, are not statistically significant; that is, estimated declines of this magnitude might have occurred through random variation even if there had been no actual changes in student performance between 1984 and 1986. The new data also show a rebound in 1988 from the 1986 levels to about the same level of performance that was exhibited in 1984.

The research into the anomalous data can be further verified in 1990. The 1990 assessment will produce additional data that will be available to check the accuracy of the modified estimates and to investigate ways to improve the estimates of error in all assessments.

Chapter 1

INTRODUCTION

Albert E. Beaton[1]

The National Assessment of Educational Progress (NAEP) is an ongoing, congressionally mandated survey that has been designed to measure what students in American schools know and can do. Since 1969, NAEP has been measuring trends in student performance in many academic subject areas, including reading, writing, mathematics, and science. NAEP's long-term trend reports are based on carefully selected national probability samples of 9-, 13-, and 17-year-old students in American schools, both public and private. NAEP is the only regularly conducted survey of educational achievement at the elementary, middle, and high school levels.

As in any long-term project measuring change, there is a tension between measuring trends in education, which implies maintaining continuity with NAEP's past objectives and measurement procedures, and introducing the best new curriculum concepts and measurement technology, which implies making changes from past assessments. In the 1984 and 1986 assessments[2], a number of innovations were introduced into NAEP in order to improve the measurement and reporting of educational proficiency. It was expected that the new technology

[1] Jo-Ling Liang produced the figures in this chapter.

[2] Although assessments are conducted in school years, this report will refer to assessments by the second year only. For example, the 1985-86 assessment will be referred to as the 1986 assessment.

would not produce results directly comparable to previous assessments, so safeguards were also introduced to protect relevance of the data collected in previous assessments. The safeguards included special "bridge" samples that were assessed in the same way as in past assessments and were intended to form bridges between the new and old assessment technologies. However, even with these safeguards, the difficulty involved in maintaining accurate trend measures while introducing innovations became apparent when the 1986 NAEP reading trend data were analyzed. This report is the story of the changes that were made and the effects they had on the estimation of student performance; it also describes some of the lessons learned about measuring trends, which should provide a valuable contribution to assessment and psychometric research.

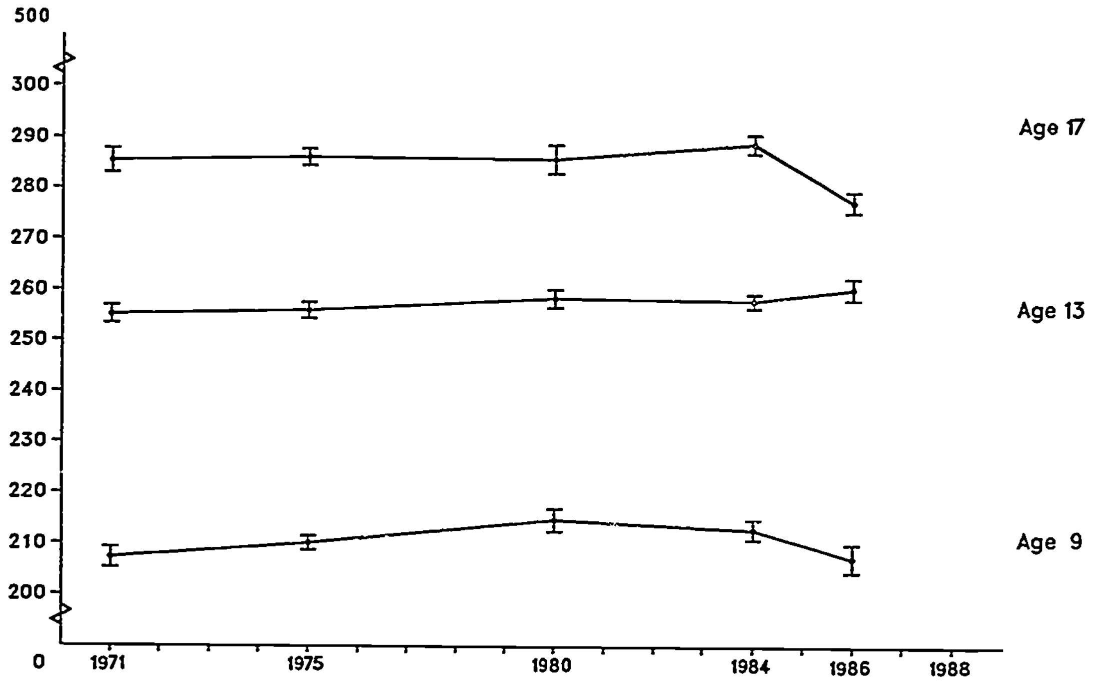
A major innovation in the 1984 and 1986 assessments was the introduction of the NAEP scales. NAEP scale scores can range from 0 to 500, but typically fall between 100 and 400. The scale scores may be interpreted as estimated scores on a hypothetical 500-item test with certain idealized properties. Using item response theory (IRT), the NAEP scales are developmental in the sense that 9-, 13-, and 17-year-old students are reported on the same subject area scales, and their proficiencies can be compared. In 1986, subscaling was introduced in mathematics and science so that proficiency in different parts of a subject area (e.g., algebra and physical science) could also be reported. Scale anchoring was introduced to report what students at a particular score level knew and were able to do that students scoring at lower levels could not. A full description of the NAEP scales and the technology used in NAEP can be found in Beaton (1987, 1988b). A general discussion of the issues in NAEP scaling are in Mislevy (1988).

The bridge samples were essential for estimating performance in past assessments on the new NAEP scales. In 1984, reading and writing were assessed. In the 1984 reading assessment, samples of students were assessed using the newer technology and randomly equivalent samples of students were assessed using the same technology that NAEP had used in previous assessments. Using the fact that the randomly equivalent samples were in principle equivalent in reading proficiency, except for sampling error, the results from previous assessments were projected onto the new NAEP reading scale. Data from past writing assessments were not projected onto the new (non-IRT) writing scale. In 1986, mathematics and science were both assessed using equivalent samples, one using the new technology and the other using traditional NAEP practices, and then the data from previous assessments were projected onto the new mathematics and science scales. Reading was also assessed in 1986, but this time using only the newer technology, since the change in technology had already been bridged in 1984. Although the general technology of the 1984 and 1986 reading assessments remained the same, some seemingly minor modifications of the booklets and administrative procedures did occur, and there was no bridge sample with which to measure their effect.

When they were first produced, the NAEP estimates of the reading proficiency of students in American schools in 1986 appeared anomalous in the judgment of the NAEP project staff and its technical advisors. Thus, the publication of trend results from the 1986 assessment was suspended; instead, the trend results were subjected to further investigation and documented in a technical report (Beaton, 1988a). The anomalous estimates are shown in Figure 1.1. The average reading proficiency estimates for 1986 indicate very sharp declines at ages 9 and 17 from the estimates for the 1984 students but no

Figure 1.1

Estimated Average Reading Performance, 1971 - 1988
with Anomalous Results for 1986
(and standard errors*)



*Bands extend from two estimated standard errors below to two estimated standard errors above the mean. Appendix A (p. 171) gives a summary of which modifications of reading scale results are used in the tables and figures in this report.

Estimated Average Reading Performance, 1971-1988
with Anomalous Results for 1986

	Weighted Reading Proficiency Means and Standard Errors		
Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
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)

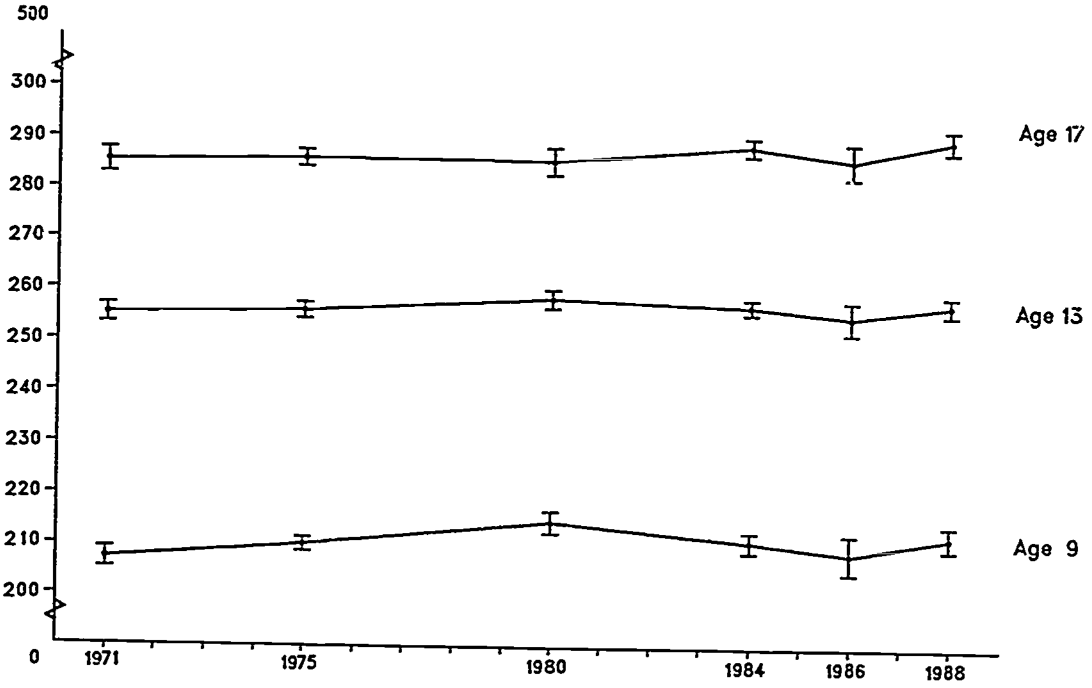
corresponding declinc--in fact, a slight rise--in average reading proficiency at age 13. The data suggested that the reading proficiency of the 1986 students was substantially more variable at all three age levels than in past assessments, with the result that more students were estimated to be at both very high and very low levels of reading proficiency. Such substantial changes in reading proficiency were considered extremely unlikely to have occurred over a two-year period without being noticed and reported by the teaching profession. Therefore, it was recommended that these results should not be used for estimating trends in American education until supported by corroborating evidence.

The purpose of this report is to present a detailed technical explanation for modified estimates of the trends in reading performance for the years 1971 to 1988, including the 1986 results. Substantial new evidence has been collected and, after a reanalysis of the reading trend data that included additional data from the 1988 assessment, the estimated long-term trends in student reading proficiency have been modified. The modified reading trend estimates, extended to 1988, are shown in Figure 1.2.

The modified trend estimates suggest that the average reading proficiency of students declined slightly at all three age levels from 1984 to 1986 and that the 1988 students rebounded to about the same averages as their 1984 counterparts. These new trend estimates show similar declines at all three age levels in 1986, not the steep declines that appeared in the first runs of the data at ages 9 and 17. The variances in student performance are now reasonably similar over the several assessment years. The remaining apparent decline in 1986, although slight and not statistically significant, and the apparent rebound in 1988 are not fully understood. Consequently, the

Figure 1.2

Modified Results
Reestimated Average Reading Performance, 1971 - 1988
(and standard errors*)



*Bands extend from two estimated standard errors below to two estimated standard errors above the mean. Appendix A (p. 171) gives a summary of which modifications of reading scale results are used in the tables and figures in this report.

Modified Results
Weighted Reading Proficiency Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
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	211.0 (1.0)	257.1 (0.7)	288.8 (0.9)
1986 Adjusted	208.6 (1.9)	255.0 (1.6)	286.0 (1.7)
1988	211.8 (1.2)	257.5 (0.9)	290.1 (1.1)

questionable 1986 reading proficiency estimates are not presented to the general public in *The Reading Report Card, 1971 to 1988* (Mullis & Jenkins, 1990).

The investigation into the anomalous 1986 reading results is not yet complete. The analysis of the data collected during the 1988 assessment has resulted in improved procedures for estimating trends in educational performance. In the 1990 assessment, more data be will collected so that the effectiveness of the newer methods can be tested in actual practice. The results of the further investigation will be published when they are complete.

* * *

Recommending that the publication of timely results be postponed is a serious matter. The decision to withhold results hinges upon contrasting the possible harm of publishing erroneous results against the possible consequences of failing to publish correct results. For example, the sharp decline in performance at age 17 might not have been an accurate representation of true changes in student performance but rather the result of flawed assessment procedures, such as errors in assessment booklets, sampling procedures, field administration, data processing, or scaling. On the advice of the NAEP Design and Analysis Committee[3], the design, administration, and analysis of the 1986 assessment were carefully reviewed by ETS/NAEP staff and a full report, *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988a) was prepared and published. The NAEP Assessment Policy Committee and the staff of the National Center for Education Statistics (NCES) of the U.S.

[3] The Design and Analysis Committee 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. Barbara Shapiro served as an observer for the NAEP Assessment Policy Committee.

Office for Educational Research and Improvement (OERI) concurred with the Design and Analysis Committee's advice to suspend publication of the 1986 reading trend results until more information was available to support the apparent decline. The OERI also commissioned an independent Technical Review Panel to prepare a thorough review of NAEP technology, and this panel has published its report (Haertel, 1989).

Fortunately, the early discovery of the 1986 reading anomaly and the consequent postponement of the publication of the 1986 reading trend results allowed enough time not only to review NAEP procedures thoroughly but also to modify the following assessment in 1988 in order to collect additional, corroborating data. In this detailed review of NAEP procedures, a number of hypotheses about the reason for the decline were investigated; these hypotheses are summarized in Chapter 2 of this report. Most of the hypotheses were rejected as very unlikely to have caused such precipitous declines in estimated student performance, but some could not be accepted or rejected because additional information was needed. In particular, in 1986 there were some seemingly minor changes in the administrative procedures, assessment booklets, and timing from the assessment in 1984. It also had to be considered that a combination of causes, not one cause alone, could explain the anomalous results (Hedges, 1989). Since there was still a possibility that the 1986 assessment results did not represent true change in student performance, the prudent course was to suspend publication until corroborative information was available.

At the time the anomalous results were discovered, the 1988 assessment had already been designed but not yet implemented, and it was early enough in the assessment cycle to modify the 1988 assessment design in order to gather some essential explanatory information. The modifications of the design are detailed in Chapter 3. Briefly, in 1988 both the 1984 and 1986 assessment procedures, booklets, and timings were administered to separate, randomly equivalent samples of 1988 students. To reproduce the 1986 assessment procedures precisely required administering some mathematics and science questions in 1988, even though these subject areas were not included in the original 1988 NAEP design. The pertinent data have now been collected and analyzed, and the data from 1986 as well as data from previous assessments have been subjected to further analyses and investigation. The mathematics and science data have been analyzed for comparison to those collected in 1986.

A summary of the recent investigations and their effect in modifying the estimated trend results are presented in Chapter 4 of this report. Subsequent chapters contain the details of the studies that contributed to the trend modifications.

* * *

The new data and analyses explain a good part, but not all, of the anomalous estimates of reading performance. Thus, the correctness of the decision to postpone presenting the 1986 trend results to the general public has been generally supported. Many individuals and agencies participated in the decision to suspend broad dissemination of the results until the questions about their validity could be satisfied, despite the obvious difficulties in missing an important publication deadline. Such careful professional judgment is essential to the continued integrity and credibility of NAEP.

Although the delay in presenting results was unfortunate, the 1986 reading anomaly has important lessons for future assessments and for other educational measurement programs as well; indeed, there are valuable lessons for the public in its perception of the rǝsults of any survey or poll. We would be remiss in not reporting what we have learned as well as reporting the modified results. More will be said about this elsewhere in this report.

One overall lesson stands out: *When measuring change, do not change the measure*. Precise implementation of this dictum is, of course, impossible in actual practice. In fact, NAEP has modified its measurement instruments by rearranging and reformatting assessment exercises since it began measuring trends.

When ETS became the NAEP grantee in 1983, it introduced item response theory (IRT) into NAEP in order to fulfill in an efficient manner NAEP's primary goal of reporting to the public what students in American schools know and can do. It is important to stress that the introduction of IRT technology did *not* cause the anomalous results; however, IRT could not compensate for the format and context changes either. Under the assumptions of IRT, test items have characteristics that are invariant in different contexts, and this property has been widely publicized and valued in the psychometric literature (see Chapter 6). Assuming this property and following past NAEP practice, the 1986 assessment booklets included many 1984 items, but placed them in different contexts. The results of the ana'yses of the data from the redesigned 1988 NAEP have demonstrated that, contrary to accepted assumptions, the item context substantially affected the behavior of these items. This effect is shown to be the major contributor to the 1986 reading anomaly.

Although they are slightly less easy to see and more difficult to isolate, the same measurement changes affect the proportion of students who respond correctly to individual items and thus also affect the average percentage of correct responses to a group of items. The adoption of IRT procedures in NAEP did not cause the anomalous results, but rather dramatized the effect of the measuring instrument on the perception of the phenomenon measured. It should be noted that the inferences of IRT are valid given the truth of the assumptions, but the assumptions may not be true; they are assumptions about the state of nature, not natural laws. In most IRT applications that compare individual students' scores or changes over time, violations of the assumptions may result in inaccuracies small enough to be ignored, since the inaccuracies are typically less than the error of measurement. However, changes in format and context that may be considered negligible when comparing individuals may *not* be negligible when comparing differences among subpopulations over time (see Chapter 8 of this report). In the particular case of NAEP, the effects of changes in measurement were apparently larger than the trend effects that were being measured. Thus, maintaining identical instruments is critical when looking for small differences.

This important lesson has led to an improvement in the 1990 NAEP design that has also been proposed for future assessments. In the 1990 design for long-term age trends, the trends in proficiency will be estimated using identical assessment booklets, administrative procedures, and timings as in the last assessment in the same subject area. In other words, each subject area for which trends are reported will duplicate as closely as possible the previous assessment with which it is to be compared, including booklets printed from the same plates, identical instructions for administration, and precise replications of definitions for target populations. In the future, if new assessment instruments are developed, they will be used to estimate long-term age trends only after they have been administered in two different assessments and their relationship to the previous trend instruments has been firmly established. This design improvement has the important effect of separating the part of NAEP used for trend estimation from the part of NAEP used to prepare detailed estimates of the proficiencies of the current students in American schools[4].

As mentioned earlier, NAEP has experienced a continuing tension between the need to retain comparability with the past and the need to be innovative in assessing the curriculum that is currently valued and taught. The new design separates the two NAEP functions, with separate samples dedicated to each. The trend samples are required to replicate past assessments as closely as possible; the cross-sectional samples are free to be innovative, introducing new objectives, new items, and new technologies. If curriculum

[4] In 1990, the long-term age trend samples will duplicate past measurement procedures as closely as possible and will use identical assessment booklets. The 1990 main NAEP samples will be used for short-term grade trends in reading. These samples will be BIB-spiraled; that is, the items will be divided into seven blocks, and three of these blocks will be placed in each assessment booklet using a balanced incomplete block (BIB) design. In this way, each item block will appear with each other block in one booklet, and each student will be asked to respond to only three of the seven blocks of items. Three to four of the seven blocks at each age/grade level will be identical to those used in the 1988 assessment, although they may be administered in conjunction with new reading blocks. It should also be noted that because the designs for future assessments will have to be consistent with the legal requirement that half of the items be publicly released, NAEP may have to differ somewhat from the ideal of maintaining identical measuring instruments, but adequate samples to estimate the effects of such differences will be maintained.

changes become so substantial as to render the trend samples obsolete, and if the new, cross-sectional samples stabilize, it may be possible to replace the older trend data with newer, more relevant trend information. In any case, any innovation introduced into the cross-sectional assessment will be linked to the trend scale, when this prove3 possible, and then moved into the trend portion of the assessment when this is desirable and the links to the trend data have been fully r 'tablished.

Chapter 2

SUMMARY OF EAKLIER RESEARCH ON THE READING ANOMALY

Albert E. Beaton

The discovery of the anomalous trend estimates during the analysis of the 1986 data started a flurry of activity to identify their causes. At first, it was assumed that some data processing error--such as a bug in a computer program--would explain the unusual results, so a concerted effort was made to examine the computer systems from data entry to final trend estimation. When no such error was found, we investigated more complex reasons for the anomaly. Finally, when no conclusive reason was found for the sharp changes in estimated performance, the 1988 NAEP design was modified to collect new data that we hoped would explain the anomaly.

This chapter presents a summary of the investigations into the reading anomaly that took place *before* the new explanatory data were collected. These studies are not reported in detail here since they have already been fully reported in *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988a); the reader is directed to that report for technical details. This previous report on the reading anomaly has been thoroughly reviewed and discussed by the Technical Review Panel on the 1986 Reading Anomaly (Haertel, 1989). The Technical Review Panel performed additional investigations into the anomalous results, and presented the results of these studies in its report.

Eight general classifications of hypotheses were investigated in the original study, relating to population and sample, measuring instruments, administrative changes, quality control, scaling, items, booklets and blocks[1], and others.

Population/Sample Hypotheses

The first set of hypotheses revolved around the possibility that either the composition of the populations of students that NAEP assesses or the actual NAEP samples from these populations of students had changed in some substantial way that would result in sharp declines in performance. 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 be likely to result in a decline in average proficiency. Also, it was important to assure that the samples that were actually drawn were not unusual and were representative of the intended student populations within the range of sampling variability.

Detailed study showed no reason to believe that the NAEP sample was not representative of the nation's students. First, we do not know how to produce a better estimate of the numbers of in-school 9-, 13-, and 17-year-old students, since the sampling weights produced using the present NAEP

[1] A block is a timed portion of an assessment that contains assessment items and/or background and attitude questions. In the 1984 and 1986 assessments, each assessment booklet contained a common block, which included background and attitude questions, and three variable blocks, which included mostly subject area exercises but also some background or attitude questions. Variable blocks are assigned to booklets using a balanced incomplete block design, and the booklets are placed in a random sequence (spiraled) so that students in an assessment session receive different booklets. A student is allowed about an hour to respond to a booklet. The timings and contents of the blocks are discussed in Chapter 3.

technology are poststratified[2] using information from the Census Bureau, the Current Population Survey, and the NAEP samples themselves. In any case, the small differences in the NAEP estimates of population sizes since 1984 could not have had a major effect on the average student performance.

E. G. Johnson (1988a) described the NAEP sampling process. There was no substantial difference in the percentage of students excluded from NAEP because of limited English proficiency, behavioral disorders, or physical or mental handicap. He did not find any reason to believe that a substantial change in the dropout rate for 17-year-olds occurred.

Johnson (1988b) also investigated the attributes of low scorers to see if there was an unusual increase in low scores in any discernible subgroup of students. He found, however, that the proportion of low scorers increased in all major subgroups of students, not merely in one or two. Johnson also examined the data to determine whether the decline was concentrated in a few schools and concluded that it was not.

Although the changes in population sizes were slight, Beaton (1988a, Chapter 6) investigated whether the slight changes could have a major effect on the estimates of reading performance, and concluded that they could not have substantially affected the overall trend estimates. In fact, the evidence showed that the decline in proficiency at age 17 was pervasive, occurring in all of the groups for which NAEP has traditionally reported results. In fact, estimated proficiencies declined for

- boys and girls, with the decline for boys somewhat larger than that for girls;
- all racial and ethnic groups;

[2] See Appendix B.

- all regions of the country, 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.

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

Measuring Instrument Hypotheses

It was thought that if the populations and samples did not explain the reading anomaly, then, perhaps, the measurement instruments would. Accordingly, we investigated several hypotheses about the assessment forms. We found that there were a number of seemingly minor differences in the assessment forms used in NAEP between 1984 and 1986. These changes are documented in detail by J. R. Johnson (1988); some of the changes are also presented in the next chapter of this report.

We had no reason to doubt the validity of the NAEP 1986 reading assessment *as a measure of reading proficiency*. The separately timed and scored blocks of assessment items may be considered as short tests, and as such they were subjected to standard item analyses. The item analysis statistics were comparable to those that occur in tests of similar length. The student number-right scores on various assessment blocks were correlated. For age 17, the median correlation, as well as the range of correlations, among the reading blocks and between the reading, mathematics, science, and computer competence blocks is shown in Table 2.1. The reading blocks contained fewer items than the blocks in other curriculum areas; however, the

Table 2.1

Correlations Among NAEP Blocks
1986 Assessment, Age 17

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

average reliability of individual reading items, which was estimated using the Spearman-Brown prophecy formula, is estimated to be greater than in other curriculum areas.

Although there was no reason to doubt that the 1986 assessment measured reading, the changes in the assessment instruments did lead to the suspicion that the 1986 assessment measured reading differently, in a way not fully comparable with past assessments. Before the physical changes in the measuring instruments were made, they were judged by professional staff to be so minor as not to affect the students' responses in a substantial way. 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. Since we could not be sure that these minor changes did not have a major effect, we therefore could not reject the hypothesis that the changes in the assessment instruments produced the sharp declines in estimated performance.

Administrative Changes Hypotheses

We also investigated hypotheses about the administration of the assessment in the field. Perhaps some changes in procedure had affected student performance. In fact, a number of changes were made in the administrative procedures; these are discussed by J. R. Johnson (1988) and summarized in Chapter 3 of this report. 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. Investigation of the reading results by the size of the assessment session showed no reason to suspect that changes in session size had a substantial effect on the results for the 17-year-olds. Another change involved the time of assessment for 9-year-olds, which changed from January 2 to March 9 in 1984 to January 6 to January 31 in 1986. Investigation of this change did not seem to explain a change of the magnitude found for 9-year-olds. In addition, using field observations, the NAEP subcontractor Westat, Inc., (Slobasky, 1988) reviewed administrative procedures used in 1984 and 1986, but failed to find changes that were considered likely to have affected only reading for the 9- and 17-year-olds.

We could not be sure, however, that seemingly minor changes in the design specifications and resulting procedures did not have an effect; therefore, we could not reject the hypotheses that administrative changes might have affected estimated student performance.

Quality Control Hypotheses

A logical possible source of the apparent decline was inaccuracy in the data processing. NAEP data were already subject to strict quality control procedures (see Beaton, 1987), but to assure independently the accuracy of the data, we selected a copy of each type of booklet at random and confirmed that student responses in the assessment booklets were accurately recorded in the database. A study of the database, described by Ferris (1988), showed it to be very accurate. An external consultant, Dr. W. B. Schrader (1988), also reviewed this process and found no basis for questioning the database or the scoring keys. Computations of proportions passing various items were done by several programs, and the results were in agreement.

We therefore concluded that gross errors in the database and major computational errors could be ruled out as explanations of the decline. We could not, however, completely rule out the possibility that minor errors did occur in 1986 or that other errors occurred in previous years.

Scaling Hypotheses

NAEP uses a complex process to estimate the distribution of reading proficiency. We therefore investigated hypotheses that the anomalous decline was an artifact of the scaling process that was used to develop and equate the NAEP scales.

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, described by Mislevy (1988), 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 ruled out the scaling process as the cause of a substantial part of the decline in reading proficiency.

Item-level Hypotheses

Another set of hypotheses involved the responses to assessment items. Several questions were pursued about individual items. 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 particular items? These hypotheses were examined by E. G. Johnson (1988c).

In summary, there was neither one nor a few items that behaved differently enough from past assessments to affect the entire results. In general, the 17-year-old students were less likely than those of previous years to respond correctly to an item, more likely to respond incorrectly or to select "I don't know," and slightly less likely to omit or not reach items. These changes in the "I don't know," omitted, and not-reached rates were found to contribute little to the decline. The decline seemed, therefore, to be associated with performance in general rather than a few unusual items.

We therefore rejected the hypotheses about one or a few aberrant items.

Booklet and Block Hypotheses

A number of hypotheses were developed about the item blocks, assessment booklets, and the context in which they were administered. We hypothesized that a student might respond differently to a reading exercise when the exercise is placed in a booklet with mathematics or science exercises. The effect of changing the context and position within a booklet of reading blocks was studied, and the results were reported by Zwick (1988a).

The study showed that, in most cases, the context and position of the block had a small effect on reading performance. There was some evidence that reading performance was adversely affected in reading blocks that followed two nonreading blocks, but even when the booklets containing this mix of blocks were removed, the sharp decline in estimated reading proficiency remained.

We therefore felt at the time of these analyses that the mixture of blocks within booklets did not contribute in a major way to the anomalous results. The question of the placement of items within blocks could not be studied without the additional data discussed in Chapter 3.

Other Hypotheses

When none of the preceding theories about the 1986 assessment seemed to give an adequate explanation for the declines in estimated reading proficiency, a number of other hypotheses were explored. Two examples follow.

The external event hypothesis. We looked for some event in 1985 or 1986 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. We felt that this tragedy might have affected the students emotionally, thereby influencing their performance. The study of this hypothesis is discussed by Beaton (1988a, Chapter 12). To investigate it, 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 1984 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.

The examination of these hypotheses did not seem to explain the reading anomaly.

* * *

In summary, these first investigations of the reasons for the 1986 reading anomaly were inconclusive. Although a number of possible explanations for the estimated decline in reading proficiency were discredited, there was insufficient information in the data to discredit a number of others. It was clearly 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 therefore modified the 1988 sample to collect data that could lead to a clarification of these issues. These changes are discussed in the next chapter.

Chapter 3

THE REDESIGN OF THE 1988 ASSESSMENT

Albert E. Beaton[1]

Although the research summarized in Chapter 2 rejected beyond a reasonable doubt several hypotheses about the 1986 reading anomaly, several other hypotheses remained viable, inasmuch as sufficient information was not available either to confirm or to reject them. In particular, there were changes in the assessment booklets and administrative procedures. Although these changes had been believed to be minor and unlikely to have a major effect on student performance, there was no way to establish the magnitude of the effect if, indeed, any effect did exist. In order to estimate the effect, the design of the 1988 NAEP was modified, as described below.

The 1988 assessment had been designed to assess performance in reading, writing, civics, and U.S. history. Assessments in mathematics or science were scheduled for 1990. As in past assessments, the design encompassed students enrolled in American schools, both public and private, at ages 9, 13, and 17, and, for some purposes, overlapping samples of fourth-, eighth-, eleventh-, and twelfth-grade students. The design of the entire 1988 assessment will be discussed in the 1988 technical report; this chapter will detail only the parts of the design that are relevant to investigating the reading trend estimates that were obtained in 1986.

[1] The tables in this chapter were produced by Jo-Ling Liang.

Before discussıng the redesign of the 1988 assessment, it is important to understand the similarities and differences between the 1984 and 1986 student samples that were used in estimating the reading trend. In both 1984 and 1986, NAEP was implementing the new design (see Messick, Beaton, & Lord, 1983) that had been proposed to improve its efficiency and usefulness. The 1984 assessment introduced many design changes, but still more were desirable. The new technology was introduced in the "main" NAEP samples. In these samples, an important change was made in the definition of the age categories, and thus different samples of students had to be assessed for estimating trends. NAEP had traditionally defined age categories differently for ages 9 and 13 than for age 17. For the main NAEP 1986 samples, uniform age definitions were used, changing the 1986 population of 9-year-olds to mostly third graders and the population of 13-year-olds to mostly seventh graders, instead of fourth and eighth graders respectively as in the past. Other important changes were also introduced into the main 1986 assessment.

Because such changes would have destroyed comparability with the past and thus the ability to estimate trends, separate samples, called "bridge samples," were assessed in 1986 at ages 9 and 13. In these bridge samples, all age populations were defined exactly as in past assessments. Consequently, the new data from these bridge samples were presumed to be comparable to the data from past assessments. Since the age definition of 17 year-olds did not change in 1986, it was felt that the main NAEP sample could be used for the measurement of trends for that age without the addition of a separate sample.

As discussed in the previous chapter, some differences did occur between the 1984 assessment and the portions of the 1986 assessment used for trend estimation, and it was presumed at the time that these differences were minor and would not have a noticeable effect on the assessment results. Similarities and differences between the 1984 and 1986 samples that were used for trend analysis are compared in the second and third columns of Tables 3.1 to 3.3. These tables are adapted from a table in a chapter by J. R. Johnson (1988), which also contains more detailed information about the differences between the 1984 and 1986 assessments.

The age definitions for trend samples in 1984 and 1986 are comparable. The estimation of reading trends used student populations defined by age only, since, before the 1984 assessment, only age populations were sampled and thus no long-term trend data are available by grade.

The 1984 students used for estimating the reading trend between 1984 and 1986 were assessed using BIB (Balanced Incomplete Block) spiraling[2] at all age levels. The purpose of BIB spiraling is to allow us to administer a large pool of items without a heavy burden on any individual student while retaining the ability to estimate the interrelationship between each pair of items. BIB spiraling requires developing and administering a set of assessment booklets so that most students in an assessment session receive different booklets, although various pairings of subsets of the items appear across booklets. To form the booklets, the items are organized into equally timed "blocks," each representing a subset of the entire item pool. These blocks are permuted so

[2] In 1984, two randomly equivalent samples were selected at each age level. In order to maintain continuity with past NAEP practices, reading and writing were assessed in one sample using matrix sampling and tape recorded administration, as in the past. In the other sample, reading and writing were assessed using BIB spiraling. The two measurement systems were equated. The 1984 reading scale means were obtained by weighting the BIB and paced means in inverse proportion to their squared standard errors and then summing.

Table 3.1

Comparison of Data Used to Measure Reading Trend, Age 9

Characteristics	1984 Sample Used for Trend	1986 Sample Used for Trend	1988 Bridge to 1984	1988 Bridge to 1986
Description of Sample	Age subsample of main NAEP sample	Age-only sample	Age subsample of bridge data set	Age-only sample
Modal Grade	4	4	4	4
Curriculum Areas	Reading, writing	Reading, mathematics, science	Reading, writing	Reading, mathematics, science
Sample Size (the number of students with reading scale values)	16,799	6,932	3,782	3,711
Age Definition	Calendar year Jan.-Dec. 1974	Calendar year Jan.-Dec. 1976	Calendar year Jan.-Dec. 1978	Calendar year Jan.-Dec. 1978
Method of Assessment	Printed	Mathematics and science--paced audiotape Reading--printed	Printed	Mathematics and science--paced audiotape Reading--printed
Dates Assessed	Winter 1984 1/2 - 3/19	Winter 1986 1/6 - 1/31	Winter 1988 1/4 - 3/11	Winter 1988 1/4 - 3/11
Time--Common background block	Approximately 15 minutes (questions were read aloud to students)[a]	Approximately 15 minutes (questions were read aloud to students)	Approximately 15 minutes (questions were read aloud to students)	Approximately 15 minutes (questions were read aloud to students)
Time--Cognitive block	14 minutes 28 minutes for each of the double-length blocks (U, V, W)	13 minutes	14 minutes 28 minutes for the double-length block V (reading and writing)	13 minutes
Number of Reading Blocks Administered	12	3	10	3
Booklet Printing and Binding	Blue ink, saddle-stitched	Blue ink[b], stapled	Blue ink, saddle-stitched	Blue ink, stapled
Response Mode	Circle letter	Fill in oval	Circle letter	Fill in oval
Scoring Method	Key-entered	Machine-scanned	Key-entered	Machine-scanned
Teacher Questionnaire	Language arts teacher was identified by students	None	Language arts teacher was identified by students[c]	None

[a] For the first three weeks of the assessment, six minutes were allowed for students to complete the background items. Because the students did not understand the questions, background items were read to the students for the remainder of the assessment.

[b] 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. The 1988 bridge booklets were duplicated from the corresponding assessment years.

[c] No teacher data were collected for this sample.

Table 3.2

Comparison of Data Used to Measure Reading Trend, Age 13

Characteristics	1984 Sample Used for Trend	1986 Sample Used for Trend	1988 Bridge to 1984	1988 Bridge to 1986
Description of Sample	Age subsample of main NAEP sample	Age-only sample	Age subsample of bridge data set	Age-only sample
Modal Grade	8	8	8	8
Curriculum Areas	Reading, writing	Reading[a], mathematics, science	Reading, writing	Reading, mathematics, science
Sample Size (the number of students with reading scale values)	17,535	6,200	4,005	3,942
Age Definition	Calendar year Jan.-Dec. 1970	Calendar year Jan.-Dec. 1972	Calendar year Jan.-Dec. 1974	Calendar year Jan.-Dec. 1974
Method of Assessment	Printed	Mathematics and science--paced audiotape Reading--printed	Printed	Mathematics and science--paced audiotape Reading--printed
Dates Assessed	Fall 1983 10/10 - 12/17	Fall 1985 11/4 - 12/13	Fall 1987 10/12 - 12/18	Fall 1987 10/12 - 12/18
Time--Common background block	6 minutes	6 minutes	6 minutes	6 minutes
Time--Cognitive block	14 minutes 28 minutes for each of the double-length blocks (U, V, W)	16 minutes	14 minutes (no double-length block in these booklets)	16 minutes
Number of Reading Blocks Administered	12	3	10	3
Booklet Printing and Binding	Brown ink, saddle-stitched	Blue ink[b], stapled	Brown ink, saddle-stitched	Blue ink, stapled
Response Mode	Circle letter	Fill in oval	Circle letter	Fill in oval
Scoring Method	Key-entered	Machine-scanned	Key-entered	Machine-scanned
Teacher Questionnaire	Language arts teacher was identified by students	None	Language arts teacher was identified by students[c]	None

[a] format and content of the 1986 age 13 reading blocks were identical to those used at age 17.

[b] 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. The 1988 bridge booklets were duplicated from the corresponding assessment years.

[c] No teacher data were collected for this sample.

Table 3.3

Comparison of Data Used to Measure Reading Trend, Age 17

Characteristics	1984 Sample Used for Trend	1986 Sample Used for Trend	1988 Bridge to 1984	1988 Bridge to 1986
Description of Sample	Age subsample of main NAEP sample	Age subsample of main NAEP sample	Age subsample of bridge data set	Age subsample of bridge data set
Modal Grade	11	11	11	11
Curriculum Areas	Reading, writing	Reading, mathematics, science, computer competence, history[a], literature[a]	Reading, writing	Reading, mathematics, science, history
Sample Size (the number of students with reading scale values)	18,984	16,418	3,652	3,715
Age Definition	Oct. 1966-Sept. 1967	Oct. 1968-Sept. 1969	Oct. 1970-Sept. 1971	Oct. 1970-Sept. 1971
Method of Assessment	Printed	Printed	Printed	Printed
Dates Assessed	Spring 1984 3/12 - 5/11	Spring 1986 2/17 - 5/2	Spring 1988 3/14 - 5/13	Spring 1988 3/14 - 5/13
Time--Common background block	6 minutes	6 minutes	6 minutes	6 minutes
Time--Cognitive block	14 minutes 28 minutes for each of the double-length blocks (U, V, W)	16 minutes	14 minutes (no double-length block in these booklets)	16 minutes
Number of Reading Blocks Administered	12	6	10	6
Average Session Size	Approximately 20	Approximately 35	Approximately 20	Approximately 35
Booklet Printing and Binding	Black ink, saddle-stitched	Blue ink[b], stapled	Black ink, saddle-stitched	Blue ink, stapled
Response Mode	Circle letter	Fill in oval	Circle letter	Fill in oval
Scoring Method	Key-entered	Machine-scanned	Key-entered	Machine-scanned
Teacher Questionnaire	Language arts teacher was identified by students	Up to 5 teachers were identified by students	Language arts teacher was identified by students[c]	Up to 5 teachers were identified by students[c]

[a] Four of the 97 booklets at age 17 contained one history block, one literature block, and one reading block (13R4).

[b] 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. The 1988 bridge booklets were duplicated from the corresponding assessment years.

[c] No teacher data were collected for this sample.

that each block appears paired with each other block in some booklet. In both 1984 and 1986, each student received a common block containing background questions and three subject matter blocks containing assessment exercises in a specific subject area and a small number of background and attitude questions.

The implementation of the BIB spiraling differed somewhat between 1984 and 1986. In 1984, both reading and writing were assessed. Accordingly, students received some combination of reading and writing blocks. Some students in an assessment session received three reading blocks, others two reading and one writing block, others one reading and two writing blocks, and still others received three writing blocks with no reading blocks at all. The 1986 design called for estimating trends in reading, mathematics, and science. Therefore, students in the samples intended for measuring trends were administered booklets that contained items from these three areas. Consequently, although many items were administered in both the 1984 and 1986 assessments, the context in which they were administered differed; the items were arranged differently within blocks and reading was administered with subject areas other than writing.

Another possibly important difference between 1984 and 1986 was the timing. In the 1986 assessment, an effort was made to increase the pool of items that could be administered. Accordingly, the time allowed to complete a subject area block was increased at ages 13 and 17 from 14 to 16 minutes. In order to improve the responses to background questions and to minimize the fatigue of the 9-year-olds, the blocks for the 9-year-olds were reduced in length to 13 minutes. To make these changes, items were rearranged and new blocks were formed. The number of items within a block was altered to allow the student about the same amount of time per item in the two assessments.

Estimating item response time, however, is only approximate, and, as will be shown late-, the number of students reaching the last items in the blocks was reduced. Since the reading items within the blocks were rearranged, an item that was near the beginning of a block in 1984 and reached by nearly all students might be near the end of a block in 1986 and not reached by a large proportion of students.

A number of other well-intentioned changes were also incorporated into the 1986 assessment. For example, to speed up the reporting process, machine-scorable books, which had been used in assessments prior to 1984, were reinstated. The format of the assessment books was made more pleasing to the eye. The number of 17-year-olds in an assessment session was increased in order to reduce the burden of several sessions on the participating high schools. The time of year in which data were collected from the 1986 sample of 9-year-olds was restricted for operational efficiency. A special study of language minority students was also administered along with the 1986 assessment.

Since the 1984 and 1986 samples were measured somewhat differently, changes in student performance are confounded with changes in measurement procedure. The 1986 reading anomaly made it no longer tolerable to assume that the change due to measurement procedure was small enough to be ignored. Unfortunately, without further information, the effect of the change in the measurement procedure could not be estimated and removed from the trend estimates on the basis of the 1984 and 1986 data alone. Therefore, new data had to be collected to estimate the effect of the changes in the measurement procedure.

* * *

In order to report the trend results as quickly as possible, appropriate data were needed to distinguish between differences resulting from student performance and differences resulting from measurement improvements. To measure the effect of each of the several 1986 improvements in measurement procedure, as well as the interactions among them, would require a very complex research design and then a very large data collection effort, an effort so large that the 1988 reporting would also be likely to be delayed. Instead, the 1988 assessment design was enlarged in such a way that the net effect of all changes in measurement procedure could be estimated, although the effect of each individual change could not. Using this net effect, it is possible to study the overall effect of measurement changes on the 1984 and 1986 reading proficiency data.

The general strategy for the redesign was to collect two samples of data from the population of students at each age level. One of the samples at each age level would be measured using the 1984 booklets and procedures and the other using the booklets and procedures of 1986. Although the data would be collected in the 1988 assessment, the measurement systems of the 1984 and 1986 trend assessments would be duplicated as closely as possible. Since the pairs of samples were to be selected from the same 1988 populations, their estimated distributions of reading proficiency should be identical, except for sampling error. If the estimated distributions differed by more than could reasonably be expected from the sampling process, then the differences in the estimated distributions could be attributed to changes in the measuring systems.

The revised design included the following samples for use in distinguishing between changes in performance and measurement:

The bridge-to-1984 samples. The 1988 assessment had been designed to assess reading, writing, civics, and U.S.. history. The design already included special bridge samples for estimating trends in reading and writing at all age levels because the 1988 NAEP overall design also introduced several changes from past assessment practices. The decision had already been made to bridge reading and writing performance back to 1984, since there had been a full assessment in both these subjects in that year. The unusual and unexpected results in the 1986 reading assessment resulted in redoubling the effort to make the measurement in the bridge to 1984 as close as possible to an exact re-creation of the 1984 measurement. The students in these samples were given copies of reprints of selected 1984 booklets, and the assessment was administered using the 1984 administration procedures, including the same block timings. The measurement difference between this 1988 sample and the 1984 sample was thus minimized.

The bridge-to-1986 samples. In order to duplicate the 1986 methodology, the 1988 design was modified by adding an additional sample at each age level. These samples were selected from the same age populations that were used in the previous trend analyses and, of course, the same populations as in the samples from the bridge to 1984. Selected booklets that were used by the 1986 trend samples were administered to these 1988 samples, duplicating as closely as possible the 1986 administrative procedures, including the timing.

Although all age populations are defined in exactly the same way for the bridge-to-1986 sample as for the bridge-to-1984 sample, the measurement system for the bridge-to-1986 sample differs at different age levels whereas the measurement system for the bridge-to-1984 sample does not. Duplicating the differences in the 1986 samples used for estimating trends, the assessment procedures for the bridge to 1986 at ages 9 and 13 differed from the procedures at age 17.

At ages 9 and 13, the trend samples in 1986 were given reading items along with mathematic. and science items. The reading data were to be compared to the BIB spiraled data collected in 1984. In the previous assessments with which the 1986 mathematics and science data were to be compared, the measurement had been administered using a tape recorder to pace students uniformly through the assessment items. In 1986, the same samples of students were used for estimating trends in reading, mathematics, and science. To accommodate the differences in procedure, the trend data in 1986 were collected using a pseudo-BIB design that attempted to blend the BIB spiraling of 1984 with the paced administration of previous assessments.

To do this, each student was administered one block of items from each subject area. The mathematics and science items were individually paced using a tape recorder, as in past mathematics and science assessments. The recorder was turned off when the reading block was administered, and the reading block was timed as a single unit. Each of the trend reading blocks was, therefore, administered as a single unit in a manner similar to the 1984 assessment, but the tape recorder was turned on for the blocks of mathematics and science items. Since selected 1986 booklets were administered in the same way to the 1988 bridge-to-1986 samples at ages 9 and 13, the result was that some mathematics and science data were collected in 1988 as a byproduct of the reading anomaly study, although such data were not part of the original assessment design.

Since the definition of 17-year-olds did not change in the main part of the 1986 assessment, the definition of the 17-year-old population remained comparable to all previous assessments, and so the pseudo-BIB design was deemed unnecessary at this age level. The 1986 reading trend for 17-year-olds was based on the main NAEP sample, which was BIB-spiraled as in 1984. Students in this sample were administered some combination of reading, mathematics, science, computer competence, U.S. history, and literature blocks. (The estimates of trends in mathematics and science were based on separate samples that were paced through the items using a tape recorder, as in their comparison samples; there were no estimates of trends in computer competence, U.S. history, or literature since these were newly developed.) Therefore, some of the BIB-spiraled booklets from 1986 containing reading blocks were selected for administration in 1988 to the bridge-to-1986 sample, and the 1986 procedures were duplicated as closely as possible. Since reprinting exact 1986 booklets was required, and most of the 1986 BIB booklets containing reading blocks also contained mathematics and science blocks, the 1988 bridge-to-1986 sample also includes samples of 17-year-olds who were assessed in portions of the 1986 mathematics and science materials.

* * *

To summarize, the analyses in this report are based primarily on four samples of data at each of the three age levels. The samples are:

- the 1984 main NAEP sample, collected during the 1984 assessment;

- the 1986 reading trend sample, collected during the 1986 assessment;
- the bridge-to-1984 sample, collected during the 1988 assessment; and
- the bridge-to-1986 sample, collected during the 1988 assessment.

The properties of these samples have been summarized in Tables 3.1, 3.2, and 3.3.

Comparing the bridge-to-1984 and the bridge-to-1986 samples is of methodological interest, since both samples were drawn from the same student populations, at the same time, and are thus identical in principle in reading proficiency. Any differences in estimated reading proficiency must, therefore, be attributable to the differences introduced by changing the measurement procedures and to those inherent in random sampling. The major part of estimating the effect of measurement procedures is comparing the estimates from the two randomly equivalent bridge samples.

However, it should be noted that exact duplication of procedure is impossible in practice and a few compromises had to be made. For example, since it was considered important to have the two 1988 bridge samples comparable to each other, the bridge-to-1986 sample for 9-year-olds was assessed between January 4 and March 11, 1988, although, as noted above, the age 9 trend assessment in 1986 occurred in January only. Also, at age 17, it was not feasible to assess the bridge-to-1986 students in sessions as large as those in 1986. However, earlier research had shown that the number of students in an assessment session did not have a substantial effect on performance at age 17.

Under the assumption that these assessment forms measured reading in the same way in 1988 as when the identical forms were last used, the comparison of the bridge-to-1984 data with the actual 1984 data is of substantive interest, since it estimates the trend in reading proficiency *in the metric of the 1984 assessment technology*. Likewise, the comparison of the bridge-to-1986 data with the actual 1986 data is of substantive interest, since it estimates the trend in reading proficiency from 1986 to 1988 *in the metric of the 1986 assessment technology*. The next chapter will give an overview of the results from these comparisons.

Chapter 4

OVERVIEW OF RESULTS

Albert E. Beaton
Rebecca Zwick
Kentaro Yamamoto[1]

The data collected in 1988 from the augmented NAEP design, which was described in the last chapter, have now been analyzed. With these data, the ETS/NAEP staff continued its research into explaining the 1986 reading anomaly and obtained improved estimates of reading performance in 1986. The improved estimates were shown in Figure 1.2 of Chapter 1. The research has led us to conclude that the changes in the reading assessment instruments and administrative procedures that were introduced between 1984 and 1986 had a major effect on the 1986 estimates of reading proficiency. A summary of this research is shown in the next section of this chapter, which describes the effect of changes in measurement procedures. The following section summarizes how the data collected during the 1988 assessment were used to improve the 1986 estimates of reading proficiency.

It should be noted that all survey results are subject to error, and NAEP uses the best available technology to estimate the standard errors for the statistics that it publishes. As assessment technology matures, and as new insights into the application of existing technology appear, there is an

[1] The figures in this chapter were produced by Jo-Ling Liang and David Freund.

opportunity to improve the estimated values of past and present surveys. In addition to adjusting the 1986 results for the effects of changes in item context and administration procedures, we have taken the opportunity to improve estimates of student performance wherever possible, although the sizes of the changes were trivially small. The next section of this chapter summarizes all of the improvements in reading proficiency estimates that were made between the publication of *The Reading Report Card: Progress Toward Excellence in Our Schools* (1985) and this report.

For completeness, the results of the analyses of newly available data on mathematics and science proficiency are presented. These data were collected as a byproduct of the modification of the NAEP design to include reading samples that were measured in the same way as in 1986. Although these data do not meet the usual standards of a full NAEP assessment for a subject area, they were analyzed in hopes of generating alternate hypotheses for the reading anomaly, but did not seem to do so.

Finally, this chapter presents its conclusions and a discussion of continuing research.

The Effect of Changes in Measurement Procedures

The redesigned 1988 assessment permitted an estimate of the effects of the changes in the measurement instruments and administrative procedures between 1984 and 1986. At each grade level, two randomly equivalent samples of students were assessed, the assessment of the bridge-to-1984 sample duplicating as closely as possible the 1984 assessment system and that of the bridge-to-1986 sample duplicating the 1986 methodology. Since both sets of samples came from identical populations, the population distributions of reading proficiency is in principle the same for the pairs of samples at each age level. If there were no differential effect due to measurement procedure, sample estimates of these distributions would be the same, except for sampling error; since the variance of the sampling error is estimable, any differences between the samples that are excessive in light of the sampling error must be due to the measurement process. Thus, the effects due to changes in the measurement process could be estimated by comparing the estimated distributions of reading proficiency for the pairs of samples.

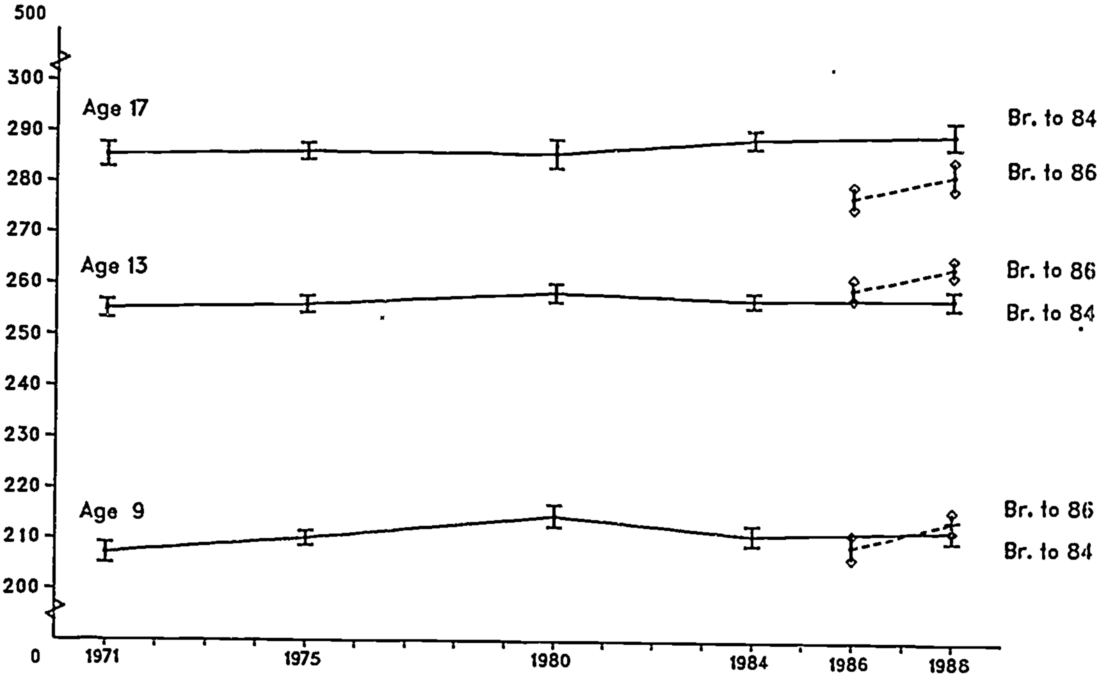
The comparisons between the estimates of the distributions of reading proficiency for the pairs of age-equivalent samples showed substantial differences. Figure 4.1 shows these results graphically. The solid lines show the estimated trend at each age level from 1971 to 1988, omitting the point for 1986, since it differed in measurement procedure. These trend lines[2] are what we would have estimated if there had been no assessment in 1986 and thus no anomalous 1986 data.

The dotted lines show where the estimated trend from 1986 to 1988 would have been (1) if the reading anomaly had been ignored, (2) if the unmodified 1986 data had been used for trend estimation, and (3) if the data from the 1988 bridge-to-1986 data had been used to estimate reading proficiency in 1988.

[2] The trend line in Figure 4.1 contains the same estimates for 1971 to 1984 as Figure 1.2. The 1988 estimates in Figure 4.1 used a conditioning model to maximize the comparability between the two 1988 bridge samples; in Table 4.1 and Figure 1.2, the conditioning model maximized comparability between the 1984 and bridge-to-1984 data. The differences between the two sets of estimates are less than 0.3 points for age-level means. The figures in Figure 1.2 are used in the most recent reading trend report (Mullis & Jenkins, 1990). See Footnote 4 in Chapter 5 of the present report.

Figure 4.1

Reading Scale Results
1971 - 1988*



* Standard errors of means are approximately 1.0. Bands extend from two standard errors below to two standard errors above the mean. Appendix A (p. 171) gives a summary of which modifications of reading scale results are used in the tables and figures in this report.

Weighted Reading Proficiency Means and Standard Errors

	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
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	211.0 (1.0)	257.1 (0.7)	288.8 (0.9)
1986	208.9 (1.2)	259.4 (1.0)	277.4 (1.1)**
1988 Br. to 84	212.1 (1.1)	257.5 (0.9)	289.9 (1.3)
1988 Br. to 86	214.0 (1.0)	263.7 (0.8)	281.9 (1.4)

** Standard error differs from column 4 of Table 4.1 because of a change in jackknife methodology.

In Figure 4.1, there are two points in 1988 at each age level, one representing the estimate made from the bridge-to-1984 sample and the other from the bridge-to-1986 sample, and the differences between the pairs of points are estimates of the differences attributable to the changes in measurement and administrative procedures. The graph shows these differences in the context of the other changes that have been observed since NAEP began measuring reading trends. The estimated effects differed by age level:

- At age 9, the estimated average reading proficiency score was *slightly higher* (1.9 points) for the 1988 bridge to 1986 than for the bridge to 1984.

- At age 13, the estimated average reading proficiency was *noticeably higher* (6.2 points) for the 1988 bridge to 1986 than for the bridge to 1984.

- At age 17, the estimated average reading proficiency was *substantially lower* (8.0 points) for the 1988 bridge to 1986 than for the bridge to 1984.

Except for the 9-year-old students, the differences due to changes in measurement procedure are larger than the changes in reading proficiency since NAEP first assessed reading.

Not only was the average reading proficiency affected by the changes in measurement procedure, but the variance was also affected. The estimated distributions of reading proficiency from the 1988 bridge-to-1986 data had larger variances at ages 13 and 17 than the variances estimated from the bridge-to-1984 samples. The distributions of reading proficiency estimated from the 1988 bridge-to-1984 and bridge-to-1986 samples are shown in Figures 4.2, 4.3, and 4.4. These distributions were estimated before the common-population equating discussed in the next section. The estimated distributions are reasonably similar at age 9, show the tendency for higher

Figure 4.2

Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 9

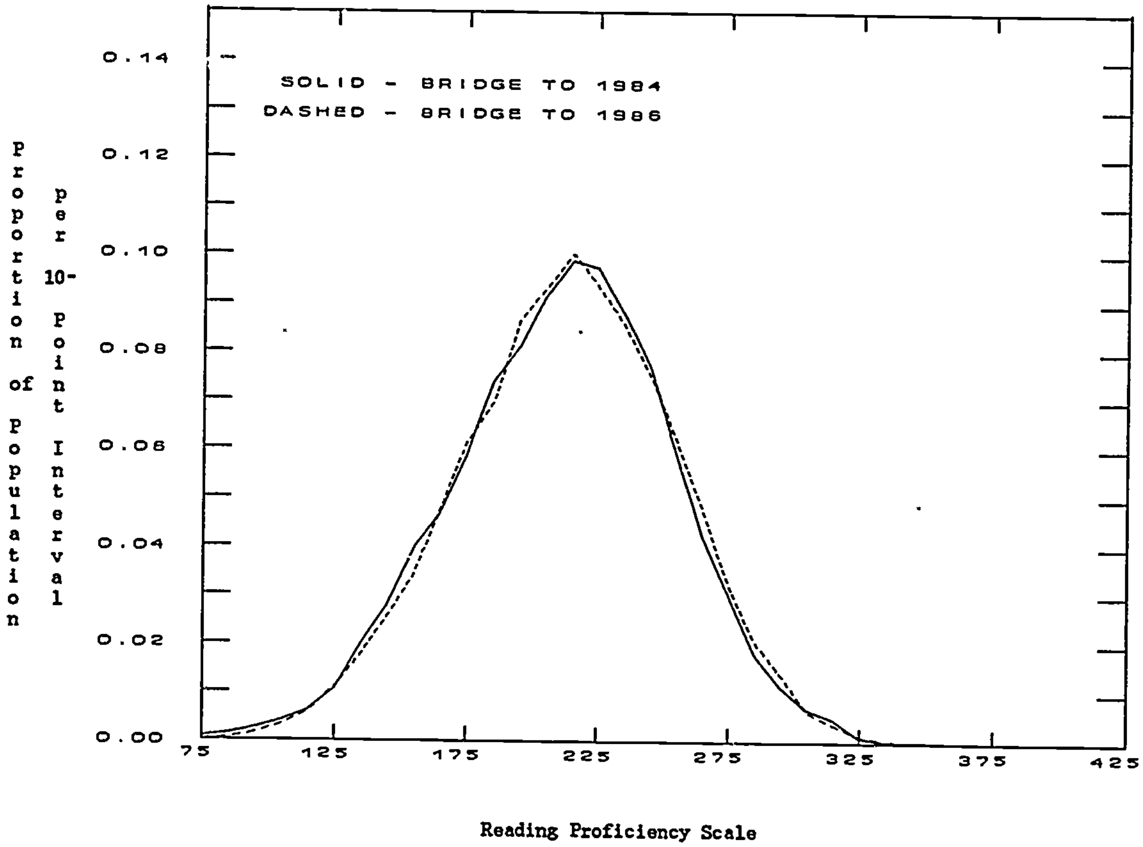
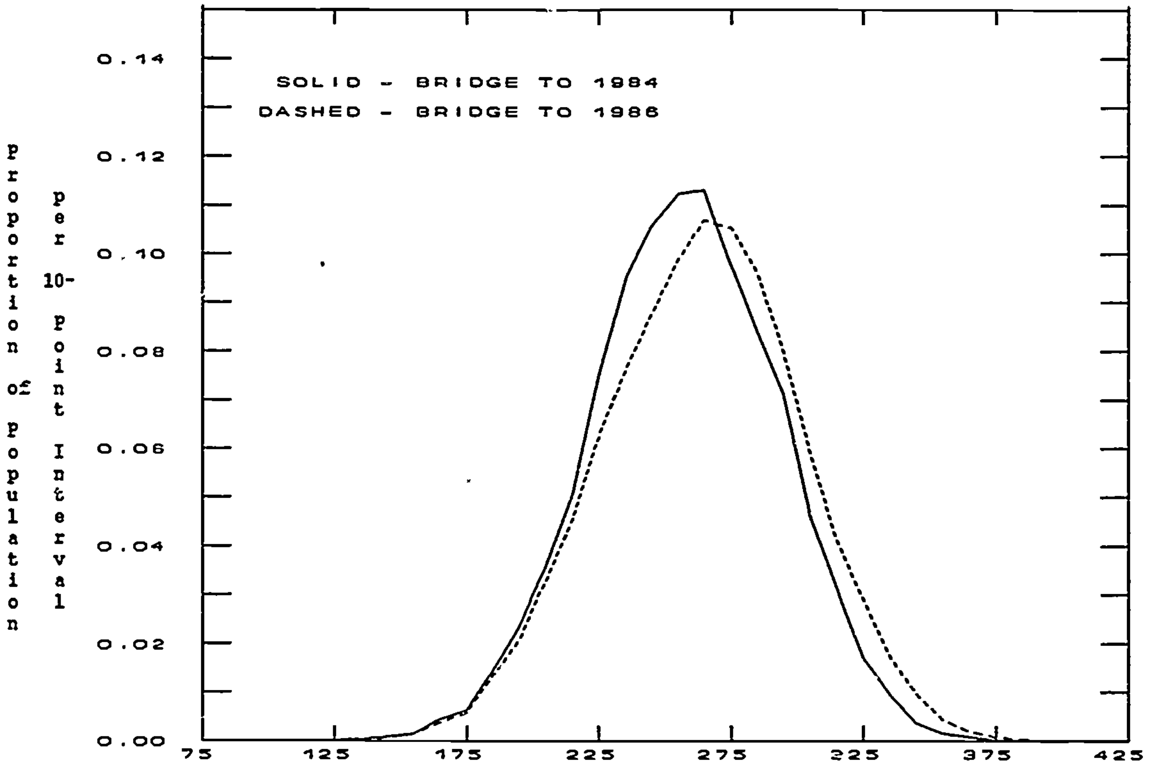


Figure 4.3

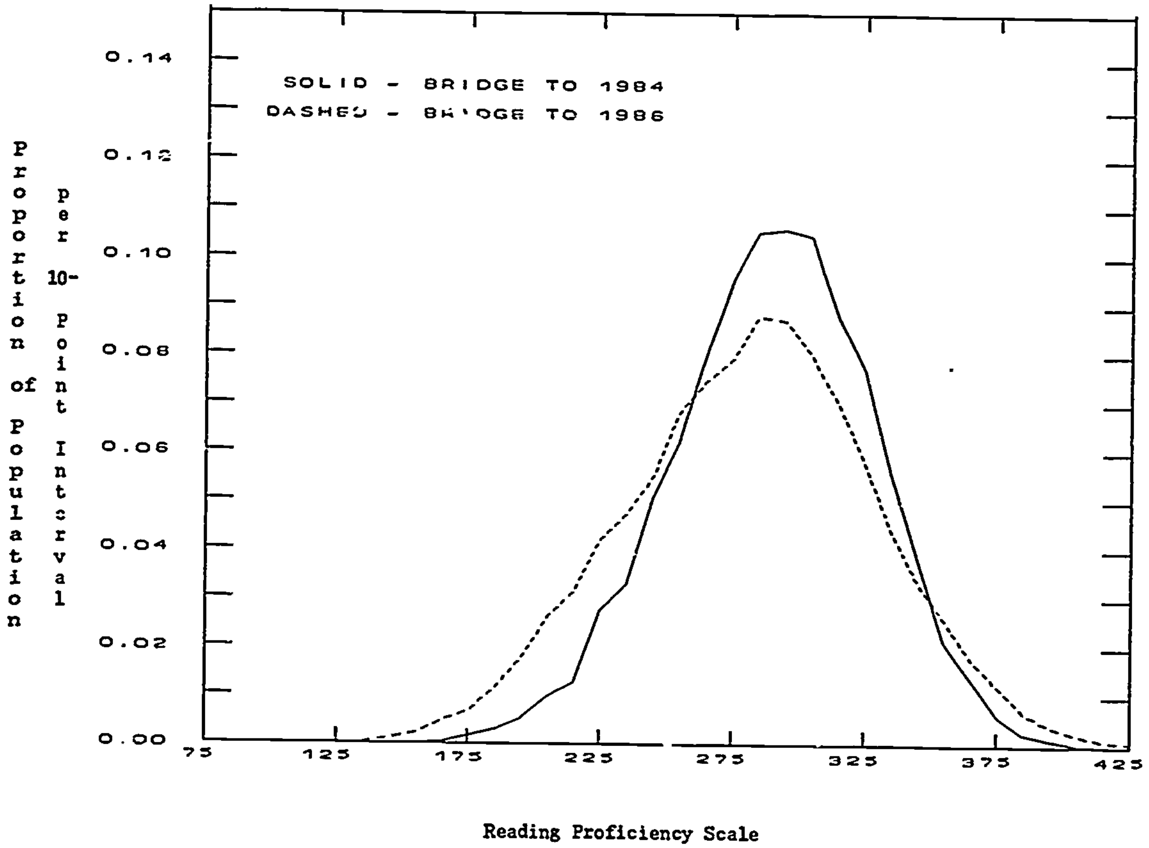
Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 13



Reading Proficiency Scale

Figure 4.4

Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 17



scores and increased variance at age 13, and the vastly increased percentage of lower scores along with a slight increase in high scores for the 17-year-olds.

Clearly, the differences due to measurement changes are substantial and unacceptable. The measurement changes are also reflected in the responses to specific items; the differences in the proportion of students correctly answering individual items showed similar results. The differences are present in the basic item data and are not, therefore, attributable to the IRT scaling technology.

Reestimating Reading Proficiency in 1986

The original procedure for equating the scales developed in 1984 and 1986 rested on the assumption that an item would function in the same way in different contexts. The availability of common populations made it possible to do another type of equating between the two different measurement procedures. Since the two 1988 bridge samples represented randomly equivalent populations, it was possible to use common-population equating methods, equating the distributions of proficiency without reliance on the consistency of parameters for the common items. In this way, the item parameters of the items common to both samples were allowed to vary as appropriate for the different contexts. This alternate approach resulted in a satisfactory equating of the bridge-to-1986 samples to the bridge-to-1984 samples.

Thus, the relationship between the reading proficiency measurements from the 1984 and 1986 assessment technologies were developed by equating the results from the two randomly equivalent 1988 samples, the bridge to 1984 and the bridge to 1986. Assuming that the relationship between the 1984 and 1986 forms had not changed, the 1986 reading results were transformed into the metric of the 1984 assessment. The 1986 results transformed into the 1984 metric were used as the modified estimates of reading proficiency in 1986. The processes of equating and transformation between the two assessment forms are presented in Chapter 6.

Although the more precise equating procedures described in Chapter 6 were used, the general concept of transforming the 1986 reading proficiency estimates into the metric of the 1984 assessment can be thought of more simply in graphic terms by returning to Figure 4.1. We know *a priori* that the pairs of points in 1988 are from identical populations of students, which are thus identical in reading proficiency, except for sampling error. Assuming that the average sampling error is close enough to zero to ignore, the observed differences between the pairs of points is due to the measurement methodology. Although a nonlinear transformation[3] was actually used, moving the 1988 bridge-to-1986 points linearly so that they coincide with the equivalent bridge-to-1984 points is a simple way to adjust for the effect of measurement technology. Moving the actual 1986 points by the same amounts approximately transforms them into the 1984 reading metric.

Modifications of the Reading Trend Lines

There has been a major improvement in the reading proficiency estimates for 1986, which was discussed in the last section, and three minor improvements in the general estimates of trend, which resulted from other improvements in assessment technology. The effects of these improvements on the reading trend lines are presented in Table 4.1. These improvements are

[3] See Figures 6.8 to 6.10 in Chapter 6.

Table 4.1

Effects of Various Changes on Trend Values

Age 9

				------- Change Resulting from -------			
Year	Reading Report Card Estimate [a]	Addition of Conditioning Variables-1985 [b]	Technical Report on Reading Anomaly Estimate [c]	Context Adjustment	Weights Adjustment	Change in Conditioning Model	Modified Estimate
1971	207.2 (1.1)	0.1	207.3 (1.0)				207.3 (1.0)
1975	209.6 (0.7)	0.6	210.2 (0.7)				210.2 (0.7)
1980	213.5 (1.1)	1.3	214.8 (1.1)				214.8 (1.1)
1984	213.2 (0.9)	-0.3	212.9 (1.0)		-1.9		211.0 (1.0)
1986			207.3 (1.4)	-0.3		+1.6	208.6 (1.9)
1988							211.8 (1.2)

Age 13

				------- Change Resulting from -------			
Year	Reading Report Card Estimate [a]	Addition of Conditioning Variables-1985 [b]	Technical Report on Reading Anomaly Estimate [c]	Context Adjustment	Weights Adjustment	Change in Conditioning Model	Modified Estimate
1971	253.9 (1.1)	1.3	255.2 (0.9)				255.2 (0.9)
1975	254.8 (0.8)	1.2	256.0 (0.8)				256.0 (0.8)
1980	257.4 (0.9)	1.1	258.5 (0.9)				258.5 (0.9)
1984	257.8 (0.6)	0.2	258.0 (0.7)		-0.9		257.1 (0.7)
1986			260.4 (1.1)	-4.4		-1.0	255.0 (1.6)
1988							257.5 (0.9)

Age 17

				------- Change Resulting from -------			
Year	Reading Report Card Estimate [a]	Addition of Conditioning Variables-1985 [b]	Technical Report on Reading Anomaly Estimate [c]	Context Adjustment	Weights Adjustment	Change in Conditioning Model	Modified Estimate
1971	284.3 (1.2)	1.1	285.4 (1.2)				285.4 (1.2)
1975	284.5 (0.7)	1.6	286.1 (0.8)				286.1 (0.8)
1980	284.5 (1.1)	1.3	285.8 (1.4)				285.8 (1.4)
1984	288.2 (0.9)	0.6	288.8 (0.9)				288.8 (0.9)
1986			277.4 (1.0)	+8.6			286.0 (1.7)
1988							290.1 (1.1)

[a] From The Reading Report Card (1985, p. 65).

[b] Additional conditional variables were added in 1985.

[c] From The NAEP 1985-86 Reading Anomaly: A Technical Report (Beaton, 1988a, p. 7).

summarized here in the same order as presented in Table 4.1 and discussed in more detail below.

- a *minor* improvement by increasing the number of variables used in the conditioning process, which affected the reading trend estimates for the 1971 to 1986 assessments at all age levels

- a *major* improvement ' the 1986 estimates of reading performance at all three age lev _s due to adjusting for the effect of the changes in measurement instruments and administrative procedure

- a *minor* improvement in the 1984 estimates of reading performance at ages 9 and 13 that resulted from a reanalysis of the 1984 sampling weights

- a *miror* improvement in the 1986 estimates of reading performance at ages 9 and 13 resulting from a change in the conditioning model

Table 4.1 is presented in three parts, one for each of the age populations assessed. The first column of each part of this table is the year in which the assessment took place.

The second column is the former trend estimate, which is taken from *The Reading Report Card: Progress Toward Excellence in Our Schools* (1985). For each age population, the estimated average reading performance is recorded for each year that reading was assessed up through 1984. The estimated standard error for the average is given in parentheses.

The third column of the table contains the effects of adding conditioning variables to the psychometric model. Conditioning is a process by which estimates of proficiency distributions can be improved by incorporating student background variables as well as item responses, assuring consistent estimates of population parameters if the conditioning model is accurate. The conditioning process, which comprises one phase of proficiency estimation, is described by Mislevy (1988) and its application in NAEP is described in other chapters in the 1986 NAEP technical report (Beaton, 1988b). When the 1984 data were analyzed, the computer program limited the number of variables that could be conditioned, and thus the conditioning process was restricted largely to basic demographic variables (e.g., region, race/ethnicity, and sex). In 1985, programming capacity was expanded. In order to assure the comparability of all assessment results, all reading proficiency results for 1971 through 1984 were reconditioned using an extended model, which included more conditioning variables. The extended conditioning model was also used in all analyses of 1986 data and some analyses of 1988 data (see Footnote 2 and Chapter 5). The largest effect of extending the conditioning model was a 1.6 point increase for the 1975 sample of 17-year-olds.

The fourth column of Table 4.1 contains the reading trend estimates that were reported in *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988a). This column contains the estimated average reading performances (and their estimated standard errors in parentheses) for students at each age level. Since the reconditioned estimates discussed in the previous paragraph were already available, they were used in this report. No estimate of 1988 proficiency was available at the time the technical report wa. published. It was primarily these trend estimates that signalled the anomaly, resulting in the further investigations.

The next three columns in the table present changes in the reading trend estimates that occurred between the publication of *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988a) and this report.

The column labeled "Context Adjustment" contains the estimated effects of the changes in measurement instrumentation and in administrative procedure on the 1986 estimates, which were discussed in the previous two sections. This adjustment was based on the common-population equating of the bridge-to-1986 samples to the bridge-to-1984 samples. Using common-population equating instead of common-item equating allowed for the possibility that the items common to the 1984 and 1986 assessment forms were functioning differently, and that form and administration changes may have had a different impact at each age level. In the equating process, a linear function was determined to match the first two moments of the estimated reading proficiency distributions of the bridge-to-1986 to the bridge-to-1984 samples at each age level. These equating functions implied a set of transformations for the 1986 item parameters. Using these transformed item parameters, the 1986 data were adjusted to derive the modified 1986 results. The adjustment procedure and its effects are described in detail in Chapter 6. It is noteworthy that the effects on average reading performance vary by age level, with a trivial negative effect (-0.3 points) at age 9, a larger ₋gative effect (-4.4 points) at age 13, and a very large positive effect (+8.6 points) at age 17.

The sixth column, labeled "Weights Adjustment," shows the changes resulting from a reanalysis of the 1984 weights. Historically, NAEP has defined the ages of 9- and 13-year-olds on an October-to-September basis and the age of 17-year-olds on a calendar-year basis, and it is necessary to continue these definitions to maintain trends. For the computation of the 1984 sampling weights, a common algorithm for poststratification was used across the three age levels. Upon review of the weighting procedures, it was noted that the 1984 estimate of the percentage of 9-year-old students in fourth grade and 13-year-old students in eighth grade could be improved and made consistent with other assessment years by applying a different algorithm, and so, in 1989, this algorithm was applied to the 1984 data at ages 9 and 13. This modification led to a decrease of 1.9 points at age 9 and 0.9 points at age 13. The details of sampling and weighting procedures are described in Appendix B. The details of the adjustment shown in this column are given in Appendix C. Except for the previous columns of this table and Figure 1.1, all results in this report are based on the modified weights. This modification affects only the two younger age levels in 1984.

The sixth column, labeled "Change in Conditioning Model," shows the effect of a second change in the conditioning model at ages 9 and 13. In 1986, the trend and cross-sectional data were conditioned together at all age levels. Differences in age definition between the trend and cross-sectional samples changed modal grades for 9- and 13-year-olds and resulted in a less-than-optimum estimate of the effect of a student being above, at, or below the usual grade for his or her age. To improve the estimates, the 1986 trend data were conditioned separately in a 1989 analysis. The details of this modification are described in Chapter 6. The change in the conditioning model affected only ages 9 and 13 in 1986, since the age of the 17-year-old population had been defined in the same way for both the trend and cross-sectional samples. The larger effect was a 1.6 point increase in the estimated average of the 9-year-olds.

The final column contains the net effect of the revised estimates of the average proficiencies (and their standard errors) as depicted graphically in Figure 1.2. These estimates incorporate the four changes described above.[4] As mentioned in Chapter 1, the revised estimates indicate a slight decline in

[4] Appendix A (p. 171) gives a summary of which modifications of reading scale results are used in the tables and figures in this report.

reading proficiency at each age level in 1986 and a rebound in 1988. However, at each age level, the estimated 1986 decline from 1984 is between two and three points on the NAEP reading scale and is not statistically significant. Also at each age level, the estimated 1988 level is essentially the same as the 1984 level, with the largest being a (nonsignificant) 1.3 point gain at age 17.[5]

The effect of the changes in measurement systems seems to have explained most, but not all, of the anomalous 1986 estimates of reading proficiency. Although the slight dip in proficiency at each age level are not individually statistically significant, the fact that all three ages show such similar results leaves some concern that another unknown factor also slightly affected the 1986 reading data.

Mathematics and Science Results

As mentioned in the last chapter, the redesign of the 1988 assessment resulted in the additional collection of some data on student performance in mathematics and science. The data were analyzed in the hope that they might shed some light on the 1986 reading anomaly. In this section, we will explore the implications.

Before proceeding, several important differences between the reading assessment and the mathematics and science assessments in 1988 should be noted. In 1988, the measurement of mathematics and science was done using the same assessment booklets and procedures as were used in the 1986 assessment. As previously noted, the 9- and 13-year-old students were assessed in

[5] See *The Reading Report Card, 1971 to 1988: Trends from the Nation's Report Card* (Mullis & Jenkins, 1990) for a detailed discussion of reading proficiency between 1984 and 1988.

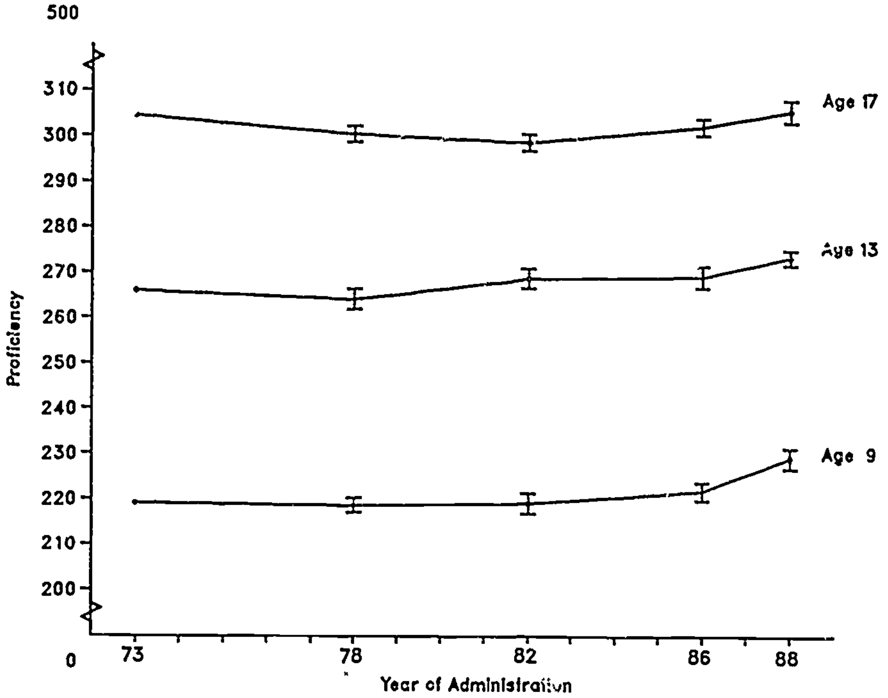
mathematics and science assessments using a tape recorder as in 1986 and in all past assessments, but the 17-year-old students were assessed by BIB spiraling (without a tape recorder), as in 1986. NAEP included only one set of samples for which mathematics and science were assessed; therefore, the effect of different forms could not be investigated. What could be investigated was whether or not the student populations appeared to improve or decline in estimated performance in mathematics and science between 1986 and 1988.

Before discussing these results, it should be noted that the NAEP scales in different subject areas are not directly comparable. The reading, mathematics, and science scales are arbitrarily set, and neither a scale point on one scale nor the differences between two scale points on that scale should be compared to those of another scale. The fact that the mathematics average is always higher than the science average at age 17 does not imply that the average 17-year-old student knows more mathematics than science; the scales are not comparable. We do believe, however, that *the directions (and perhaps magnitudes) of change* in scale performance are somewhat comparable, and we wished to see if changes in performance in mathematics and science between 1986 and 1988 were in the same direction as the estimated increase in reading performance. The magnitude of the changes could be expected to be different on the different scales.

Estimates of the trends in performance for mathematics and science are shown respectively in Figuies 4.5 and 4.6. To show the amount of change between 1986 and 1988 in the context of the average performances that have been estimated in past assessments, the trend lines include performance estimates from all previous assessments in each subject area. The years for

Figure 4.5

Trend of Proficiency Scale Means for Mathematics
1973 - 1988*



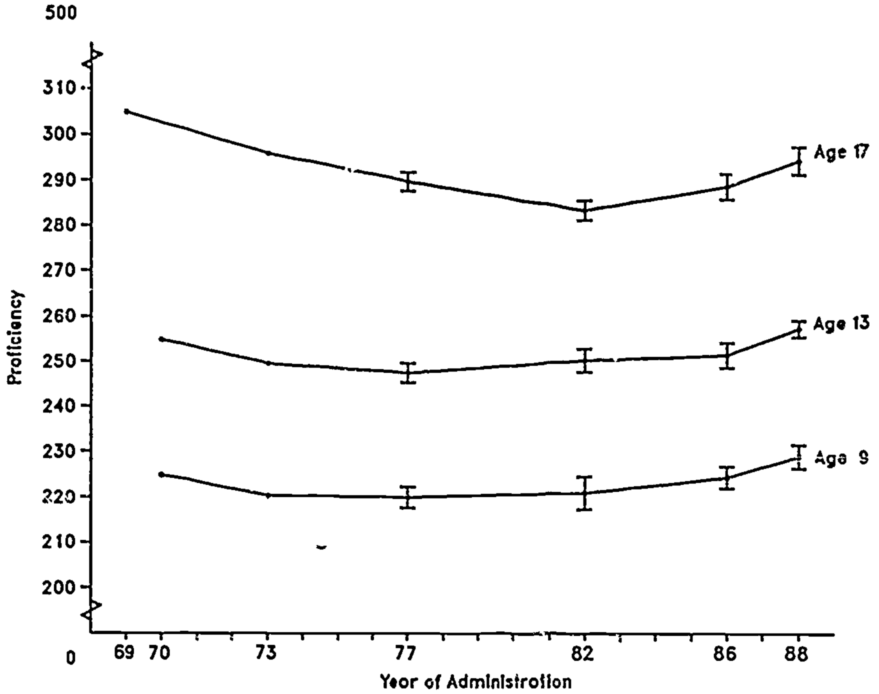
* 1973 results were interpolated for this plot. Bands extend from two standard errors below to two standard errors above the mean.

Mathematics Scale Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
1973	219.1*	266.0*	304.4*
1978	218.6 (0.8)	264.1 (1.1)	300,4 (0.9)
1982	219.0 (1.1)	268.6 (1.1)	298.5 (0.9)
1986	221.7 (1.0)	269.0 (1.2)	302.0 (0.9)
1988	229.0 (1.1)	273.3 (0.8)	305.4 (1.2)

Figure 4.6

Trend of Proficiency Scale Means for Science
1969 - 1988*



* 1969, 1970, and 1973 results were interpolated for this plot. Bands extend from two standard errors below to two standard errors above the mean.

Science Scale Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
1969	--	--	304.8*
1970	224.9*	254.9*	--
1973	220.3*	249.5*	295.8*
1977	219.9 (1.2)	247.4 (1.1)	289.6 (1.0)
1982	220.9 (1.8)	250.2 (1.3)	283.3 (1.1)
1986	224.3 (1.2)	251.4 (1.4)	288.5 (1.4)
1988	228.9 (1.3)	257.3 (0.9)	294.2 (1.5)

which estimates are available are shown on the abscissa; these years differ for the two subject areas because they were not usually assessed together until 1986 and 1988. The figures show the estimated average performance on the mathematics and science scales for the different age levels. The details of the analyses that led to these estimates are described in Chapter 7.

What is at first most striking in these graphs is that estimated changes in average performance in both mathematics and science are similar to the changes in average reading performance between 1986 and 1988. In reading, the modified trend lines, shown in Figure 1.2, show a slight increase in average reading performance from 1986 to 1988 at all age levels; in both mathematics and science, a similar slight rise also occurs at all age levels. Thus, the estimates are consistent for the trend lines in all subject areas.

The fact that the nine (three subject areas by three age levels) estimated changes between 1986 and 1988 are consistent is not truly surprising, however, since the nine estimated changes are not independent. At each age level, the changes in reading, mathematics, and science are based on one sample of students that was assessed in all three subject areas in 1986 and another sample that was similarly assessed in 1988[6].

The evidence from the mathematics and science data, therefore, did not suggest any other reasons for the anomalous reading data. But, for the

[6] At ages 9 and 13 in both 1986 and 1988, exactly the same students were used for measuring trend in all subject areas. At age 17 in 1986, the student samples were partially overlapping, with individual students assessed in one, two, or three subject areas. The 1988 samples of 17-year-olds were also partially overlapping, using six BIB-spiraled booklets: one booklet contained three reading blocks, two booklets contained one reading block and two mathematics blocks, two booklets contained one reading block and two science blocks, and the final booklet contained one reading, one mathematics, and one science block.

reasons cited above, and since analysis of these data might have suggested other hypotheses, they are included in this report.

Conclusions and Continuing Research

These investigations into the reasons for the apparently anomalous reading data in 1986 have resulted in modified estimates of reading performance. The reading trend lines do not now seem anomalous, and the 1986 estimates, now slightly lower than the 1984 estimates at each age level, are within the boundaries that could be expected from the random sampling process if there had been no reading proficiency changes in the student populations. Although the results are now reasonable, they are not conclusive. More research should be done to assure that no other major factors affect the accuracy of assessment results.

The major contributor to the unusual 1986 results was the effect of changes in the measurement system, which in this case included changes in assessment context and administrative procedures. The present research shows that these changes had a substantial and unpredictable effect on reading proficiency estimates. The 1988 assessment included randomly equivalent samples in which the different measurement systems were used, and so equal population equating instead of equal item-parameter equating could be used to equate the two measurement systems. The equal population equating resulted in the trend line modifications that make the reading trend seem reasonable.

The 1990 assessment will collect additional data that may shed more light on the 1986 reading results. As in 1988, the 1990 assessment will contain two randomly equivalent samples, one of which will be measured using the 1984 measurement system and the other using the 1986 measurement system.

The analyses of the 1988 equivalent samples that resulted in the modified trend estimates, which were reported above, can be repeated using the new 1990 data. The stability of the equal population equating process can thus be estimated. The stability of item parameters within a particular measurement system can also be further investigated.

Chapter 5

ANALYSES OF 1988 READING BRIDGE DATA

Rebecca Zwick[1]

Overview

As described in Chapter 3, a bridge to 1984 and a bridge to 1986 were included in the 1988 assessment, each incorporating test booklets and administration procedures that replicated as closely as possible those of the corresponding assessment year. The analysis of data from these bridges was expected to shed further light on the causes of the anomalous reading results in 1986. Three types of comparisons are discussed in this chapter: (1) comparison of the 1988 bridge to 1984 with the 1988 bridge to 1986, (2) comparison of the 1984 assessment with the 1988 bridge to 1984, and (3) comparison of the 1986 assessment with the 1988 bridge to 1986. The first of these comparisons has the most direct bearing on the anomaly; the second two comparisons yield estimates of changes in reading proficiency that are unconfounded with changes in the assessment instruments and conditions. Comparisons are given both in terms of item percents correct and in terms of reading scale values. Table 5.1 lists the samples on which this chapter is based, along with the number of students, number of reading blocks, and time of testing. Sampling procedures for the bridges are described in Appendix B.

[1] David Freund provided statistical programming, with assistance from Minhwei Wang and Kate Pashley. David Freund and Jo-Ling Liang produced the figures in this chapter. Robert Mislevy provided consultation on scaling.

Table 5.1

NAEP Samples Used in 1988
Investigation of 1986 Reading Anomaly*

	Sample	Number of Students (Scaled Results)	Number of Reading Blocks **	Time of Testing
1984				
Age 9	Age subsample of spiral	16,799	12	Jan. 2 - March 19, 1984
Age 13	Age subsample of spiral	17,535	12	Oct. 10 - Dec. 17, 1983
Age 17	Age subsample of spiral	18,984	12	March 12 - May 11, 1984
1986				
Age 9	Bridge to 1984	6,932	3	Jan. 6 - Jan. 31, 1986
Age 13	Bridge to 1984	6,200	3	Nov. 4 - Dec. 13 1985
Age 17	Age subsample of spiral	16,418	6	Feb. 17 - May 2, 1986
1988				
Age 9	Age subsample-bridge to 1984	3,782	10	Jan. 4 - Mar. 11, 1988
	Bridge to 1986	3,711	3	Jan. 4 - Mar. 11, 1988
Age 13	Age subsample-bridge to 1984	4,005	10	Oct. 12 - Dec. 18, 1987
	Bridge to 1986	3,942	3	Oct. 12 - Dec. 18, 1987
Age 17	Age subsample-bridge to 1984	3,652	10	March 14 - May 13, 1988
	Age subsample-bridge to 1986	3,715	6	March 14 - May 13, 1988

*Age definitions for these samples are consistent with 1984 definitions.
**The number of blocks that include at least one reading scale item.

Some of the main differences between instruments and procedures for the 1984 and 1986 assessments were these:

- Reading was accompanied by writing in 1984 and by mathematics and science (and, at age 17, by computer science, history, and literature) in the 1986 trend samples.

- The composition of reading item blocks was not the same in 1984 and 1986. Therefore, items that appeared in both years did not necessarily appear in the same order or context, nor was the time allowed per item assured to be the same.

- In 1984, students responded to items by circling the letter of the correct response, whereas in 1986, students responded by filling in an oval.

Further detail on the differences between the two assessments appears in Chapters 2 and 3.

Like the 1984 assessment, the 1988 bridge to 1984 included reading and writing blocks. At each age, this bridge consisted of six of the 1984 booklets that contained at least one scaled reading block. (The 1984 balanced incomplete block [BIB] assessment included 57 such booklets at age 9 and 56 such booklets at ages 13 and 17.) The six bridge booklets included 10 of the 12 reading blocks scaled in 1984.

Like the 1986 assessment, the bridge to 1986 included reading, mathematics, and science blocks. At ages 9 and 13, this bridge contained all three booklets and all three reading blocks that were used for the estimation oi trend in 1986. At age 17, the bridge to 1986 included only six of the 35 booklets from 1986 that had at least one reading block, but these bridge booklets contained all six 1986 reading blocks.[2]

Tables and Figures Used in the Three Bridge Comparisons

The three bridge comparisons described in the following section are based on data displayed in Tables 5.2 to 5.9 and Figures 5.1 to 5.4.

Table 5.2 gives the mean percents correct for the two 1988 bridge samples and for the 1984 and 1986 assessments. These means are based on all multiple-choice items that were included in both bridge samples.[3] Standard errors obtained through jackknifing (see E. G. Johnson, Burke, Braden, Hansen, Lago, & Tepping, 1988) are given in parentheses.

[2] The bridge to 1984 included booklets 16, 17, 27, 34, 55, and 60 at age 9 and booklets 13, 16, 17, 21, 34, and 57 at ages 13 and 17 (see J. R. Johnson, 1987, pp. 120-121). The bridges to 1986 included booklets 1-3 at ages 9 and 13 and booklets 14, 36, 47, 62, 68, and 81 at age 17 (see Beaton, 1988a, pp. 421-423). At ages 9 and 13, each booklet in the bridge to 1986 included one block each of reading, math, and science. At age 17, two booklets contained two blocks of math and one block of reading, two booklets contained two blocks of science and one block of reading, one booklet contained one block each of reading, math, and science, and one booklet contained three reading blocks. Although some 1986 age 17 booklets combined reading with computer competence or with history and literature, these booklets were not included in the 1988 bridge to 1986.

[3] Percent correct is defined as R/(R + W + O + DK), where R, W, O, and DK represent the sum of the student weights for those who got the item right, those who got the item wrong, those who reached the item but omitted it, and those who indicated that they did not know the answer, respectively. Students who did not reach the item are not included in the computation. Note that the percents correct that appear in portions of *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988b) were computed using NAEP's earlier definition of the proportion correct, R/(R + W + DK). The change in definitions has very little impact on the reported results. Also, note that, in the 1988 report, a larger set of items was analyzed and that the 1984 results for ages 9 and 13 were based on sampling weights that have now been modified (Appendix C).

Table 5.2

Mean Percents Correct with Standard Errors for 1988 Bridges to 1984 and 1986 and for 1984 and 1986 Assessments*

	1984 Assessment	1988 Bridge to 1984	1986 Assessment	1988 Bridge to 1986
Age 9 (26 items)	60.0 (0.6)	62.2 (1.0)	59.3 (0.9)	62.1 (0.8)
Age 13 (19 items)	63.1 (0.4)	63.8 (0.7)	64.4 (0.7)	65.9 (0.4)
Age 17 (23 items)	75.9 (0.4)	76.6 (0.5)	73.5 (0.6)	73.8 (0.6)

*All multiple-choice items that were common to both bridges were used in this analysis.

Tables 5.3 - 5.8 give mean percents correct for NAEP's major reporting groups on these same items for the two bridge samples (Tables 5.3, 5.5, and 5.7) and for the 1984 and 1986 assessments (Tables 5.4, 5.6, and 5.8). The column labeled "N" gives the average number of students responding to each item. In addition, these tables include, for each sample, the average across items of the percent of st 'ents who did not reach the item. Differences between the two bridge samples and between 1986 and 1984 in percents correct and percents not reached are also given.

Table 5.9 gives means and standard deviations of reading scale values for these same samples of students, using the metric of the 1984 reading scale. Standard errors of means are giv.:n in parentheses. These results are based on NAEP plausible values technology (see Mislevy, 1988), which is a method of estimating proficiency distributions based oı students' item responses and background character'.stics (referred to in this context as conditioning variables). The analyses that produced the results in Table 5.9 included six conditioning variables: gender, ethnicity, size and type of community (STOC), region, parents' education, and TV watching.[4]

[4] The coding for these conditioning variables was the same as that given in Mislevy, 1988 (p. 198). The estimated coefficients of the conditioning variables for the two bridge samples appear in Table D.1 in Appendix D. Note that the results reported in Table 5.9 for the 1988 bridge to 1984 are not identical to those reported in *The Reading Report Card, 1971 to 1988* (Mullis & Jenkins, 1990) and in Tables 4.1, 6.2, 6.3, and 6.4 and Figures 1.2 and 6.1. For purposes of trend reporting, a more complete set of conditioning variables was used in order to maximize comparability with the 1984 assessment. The results reported here maximize the comparability between the two sets of 1988 bridge results. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

Table 5.3

NAEP 1988 Reading Bridges: Age 9
Weighted Mean Percents Correct and Percents Not Reached for 26 Multiple-choice Items Common Between Bridges*

	BRIDGE TO 1984			BRIDGE TO 1986			DIFFERENCE 1986 - 1984	
SUBGROUP	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	598	62.2 (1.0)	5.5	1135	62.1 (0.8)	8.4	0.0 (1.2)	2.8
SEX								
MALE	295	60.1 (1.3)	5.7	559	60.5 (0.9)	9.2	0.4 (1.6)	3.5
FEMALE	303	64.2 (0.9)	5.5	575	63.6 (1.1)	7.6	-0.6 (1.4)	2.1
ETHNICITY								
WHITE	361	65.5 (1.1)	4.2	689	66.1 (0.9)	7.4	0.6 (1.4)	3.2
BLACK	104	51.5 (2.2)	10.6	175	50.4 (1.4)	11.2	-1.1 (2.6)	0.6
HISPANIC	108	51.4 (2.3)	8.0	216	48.9 (1.9)	11.5	-2.5 (3.0)	3.5
OTHER	26	66.6 (2.4)	4.9	54	65.1 (2.7)	7.6	-1.6 (3.6)	2.7
REGION								
NORTHEAST	149	63.7 (1.4)	7.9	295	63.6 (1.5)	7.3	-0.1 (2.0)	-0.6
SOUTHEAST	159	59.8 (2.4)	6.0	316	59.9 (1.8)	9.4	0.1 (3.0)	3.4
CENTRAL	128	64.9 (1.6)	4.3	245	63.2 (1.5)	8.1	-1.[illegible] (2.2)	3.8
WEST	162	60.7 (2.0)	4.3	278	62.0 (1.5)	8.6	1.3 (2.5)	4.3
PARENTAL EDUCATION								
LESS THAN H.S.	25	53.4 (3.1)	7.7	43	49.5 (1.8)	6.6	-3.8 (3.6)	-1.2
GRADUATED H.S.	90	61.3 (1.5)	6.3	163	60.9 (1.5)	8.6	-0.3 (2.1)	2.3
POST H.S.	30	63.9 (3.5)	5.0	87	67.2 (2.1)	7.6	3.3 (4.1)	2.6
GRADUATED COLLEGE	245	67.9 (1.2)	4.5	490	68.0 (0.9)	6.3	0.0 (1.5)	1.9
UNKNOWN	207	56.8 (1.5)	5.9	342	54.6 (1.1)	11.3	-2.3 (1.8)	5.4

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding, the N's for subgroups may not sum to the N for the total group.

Table 5.4

NAEP 1984 and 1986 Reading Assessments: Age 9 Weighted Mean Percents Correct and Percents Not Reached for 26 Multiple-choice Items Common Between Bridges*

SUBGROUP	1984 ASSESSMENT			1986 ASSESSMENT			DIFFERENCE 1986-1984	
	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	1972	60.0 (0.6)	6.5	2102	59.3 (0.9)	9.3	-0.8 (1.1)	2.8
SEX								
MALE	998	57.3 (0.6)	6.7	1049	56.6 (0.9)	10.1	-0.7 (1.1)	3.4
FEMALE	974	62.8 (0.7)	6.2	1053	61.8 (1.0)	8.6	-1.0 (1.2)	2.3
ETHNICITY								
WHITE	1338	64.0 (0.7)	5.6	1386	63.4 (0.9)	8.6	-0.6 (1.2)	3.0
BLACK	281	46.9 (1.1)	9.8	249	46.5 (1.0)	12.0	-0.4 (1.5)	2.2
HISPANIC	264	49.3 (0.9)	8.0	307	46.5 (1.6)	10.6	-2.8 (1.8)	2.6
OTHER	89	61.5 (1.8)	5.9	160	58.5 (2.7)	10.2	-2.9 (3.2)	4.3
REGION								
NORTHEAST	446	62.2 (1.6)	6.1	524	61.5 (2.2)	8.6	-0.7 (2.7)	2.5
SOUTHEAST	489	57.7 (1.1)	7.2	476	55.7 (1.8)	9.5	-2.0 (2.1)	2.3
CENTRAL	567	62.5 (1.5)	6.3	517	61.5 (1.8)	9.2	-1.1 (2.3)	2.9
WEST	470	57.8 (0.7)	6.4	585	58.0 (1.8)	9.9	0.2 (2.0)	3.6
PARENTAL EDUCATION								
LESS THAN H.S.	115	49.6 (1.9)	8.1	88	48.0 (1.6)	11.1	-1.7 (2.5)	3.0
GRADUATED H.S.	372	59.2 (0.8)	6.3	321	55.2 (1.0)	9.9	-4.0 (1.3)	3.6
POST H.S.	99	60.2 (1.9)	5.4	146	66.1 (1.2)	6.4	6.0 (2.3)	1.0
GRADUATED COLLEGE	638	68.0 (0.8)	4.3	822	66.3 (0.9)	7.5	-1.7 (1.2)	3.3
UNKNOWN	727	55.9 (0.8)	7.6	721	53.1 (1.1)	11.3	-2.8 (1.4)	3.7

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding, the N's for subgroups may not sum to the N for the total group.

Table 5.5

NAEP 1988 Reading Bridges: Age 13
Weighted Mean Percents Correct and Percents Not Reached for 19 Multiple-choice Items Common Between Bridges*

	BRIDGE TO 1984			BRIDGE TO 1986			DIFFERENCE 1986 - 1984	
SUBGROUP	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	657	63.8 (0.7)	0.8	1280	65.9 (0.4)	4.8	2.1 (0.8)	4.0
SEX								
MALE	320	61.5 (0.8)	1.1	636	64.3 (0.6)	5.7	2.8 (1.0)	4.7
FEMALE	336	66.0 (0.8)	0.6	644	67.5 (0.6)	3.9	1.5 (1.0)	3.3
ETHNICITY								
WHITE	478	65.8 (0.8)	0.2	904	68.7 (0.4)	3.1	2.9 (0.9)	2.9
BLACK	92	58.9 (1.5)	2.9	196	59.2 (0.7)	10.3	0.3 (1.6)	7.4
HISPANIC	58	55.6 (2.1)	1.7	125	55.0 (1.8)	9.2	-0.6 (2.7)	7.4
OTHER	29	66.8 (2.9)	1.6	54	67.3 (2.0)	4.0	0.5 (3.6)	2.5
REGION								
NORTHEAST	144	64.9 (1.8)	0.1	287	68.0 (0.9)	5.3	3.1 (2.0)	5.2
SOUTHEAST	139	63.8 (1.3)	1.2	256	65.8 (1.3)	7.2	1.9 (1.8)	6.0
CENTRAL	196	62.6 (1.5)	1.4	368	65.3 (1.2)	3.2	2.7 (1.9)	1.9
WEST	177	63.9 (1.4)	0.6	369	64.9 (0.7)	3.7	1.0 (1.6)	3.1
PARENTAL EDUCATION								
LESS THAN H.S.	45	57.9 (1.8)	0.8	88	58.9 (2.2)	9.1	1.0 (2.9)	8.3
GRADUATED H.S.	213	61.9 (1.0)	0.7	321	61.8 (0.6)	5.5	-0.2 (1.1)	4.7
POST H.S.	68	66.3 (1.8)	0.4	208	67.8 (1.0)	2.5	1.5 (2.0)	2.2
GRADUATED COLLEGE	271	67.7 (0.9)	0.5	559	69.9 (0.5)	3.2	2.2 (1.0)	2.8
UNKNOWN	58	55.0 (1.9)	1.9	102	59.9 (1.8)	10.8	4.9 (2.6)	9.0

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding, the N's for subgroups may not sum to the N for the total group.

Table 5.6

NAEP 1984 and 1986 Reading Assessments: Age 13 Weighted Mean percents Correct and Percents Not Reached for 19 Multiple-choice Items Common Between Bridges*

	1984 ASSESSMENT			1986 ASSESSMENT			DIFFERENCE 1986-1984	
SUBGROUP	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	2208	63.1 (0.4)	2.4	1911	64.4 (0.7)	5.9	1.2 (0.8)	3.5
SEX								
MALE	1108	61.3 (0.6)	3.0	939	63.5 (0.7)	7.6	2.1 (0.9)	4.6
FEMALE	1100	65.1 (0.6)	1.8	972	65.3 (0.9)	4.2	0.2 (1.0)	2.4
ETHNICITY								
WHITE	1599	65.8 (0.5)	1.7	1175	66.9 (0.8)	4.0	1.1 (1.0)	2.3
BLACK	294	53.5 (1.2)	5.4	418	58.7 (1.3)	11.6	5.2 (1.7)	6.2
HISPANIC	241	54.8 (1.5)	4.5	254	52.8 (2.0)	12.9	-2.0 (2.5)	8.4
OTHER	74	62.7 (2.4)	2.2	63	64.4 (2.8)	2.5	1.7 (3.6)	0.3
REGION								
NORTHEAST	486	64.6 (0.8)	2.3	475	66.5 (1.4)	3.9	1.9 (1.6)	1.6
SOUTHEAST	548	63.0 (1.2)	2.8	427	64.1 (1.1)	8.0	1.0 (1.6)	5.2
CENTRAL	636	62.7 (1.1)	2.3	450	62.8 (2.2)	6.0	0.2 (2.5)	3.7
WEST	529	62.5 (0.6)	2.3	559	64.3 (1.3)	5.6	1.8 (1.4)	3.3
PARENTAL EDUCATION								
LESS THAN H.S.	193	54.5 (1.1)	3.8	152	57.8 (1.6)	9.9	3.2 (2.0)	6.1
GRADUATED H.S.	779	61.1 (0.7)	2.6	545	62.3 (0.8)	6.4	1.2 (1.1)	3.8
POST H.S.	220	67.8 (0.7)	2.0	296	67.2 (1.0)	3.6	-0.6 (1.3)	1.7
GRADUATED COLLEGE	792	68.4 (0.6)	1.4	724	68.5 (0.7)	4.4	0.1 (0.9)	2.9
UNKNOWN	203	52.9 (0.9)	3.9	159	52.1 (2.3)	12.4	-0.8 (2.5)	8.5

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding, the N's for subgroups may not sum to the N for the total group.

Table 5.7

NAEP 1988 Reading Bridges: Age 17
Weighted Mean Percents Correct and Percents Not Reached
for 23 Multiple-choice Items Common Between Bridges*

	BRIDGE TO 1984			BRIDGE TO 1986			DIFFERENCE 1986 - 1984	
SUBGROUP	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	604	76.6 (0.5)	0.5	867	73.8 (0.6)	2.8	-2.8 (0.8)	2.2
SEX								
MALE	277	74.3 (0.8)	0.6	412	71.1 (1.1)	3.5	-3.2 (1.4)	2.9
FEMALE	327	78.6 (0.7)	0.4	455	76.6 (0.7)	2.0	-2.0 (1.0)	1.6
ETHNICITY								
WHITE	425	78.8 (0.5)	0.3	639	76.2 (0.6)	1.8	-2.6 (0.8)	1.5
BLACK	106	71.4 (1.4)	0.3	133	65.5 (1.9)	4.7	-6.0 (2.4)	4.3
HISPANIC	48	66.0 (1.6)	1.4	66	65.5 (2.5)	4.8	-0.5 (2.9)	3.4
OTHER	26	78.2 (2.7)	2.3	29	76.6 (2.9)	7.7	-1.6 (3.9)	5.5
REGION								
NORTHEAST	131	79.1 (1.2)	0.5	203	75.5 (1.3)	2.6	-3.6 (1.8)	2.1
SOUTHEAST	155	74.6 (1.1)	0.3	221	73.8 (1.6)	2.0	-0.9 (1.9)	1.7
CENTRAL	106	77.3 (0.7)	0.1	156	73.7 (1.0)	2.7	-3.7 (1.2)	2.6
WEST	212	75.4 (1.1)	1.1	287	72.6 (1.1)	3.5	-2.7 (1.5)	2.4
PARENTAL EDUCATION								
LESS THAN H.S.	52	69.6 (1.7)	0.4	69	63.3 (1.9)	2.9	-6.3 (2.6)	2.5
GRADUATED H.S.	175	74.3 (0.8)	0.3	197	70.0 (1.2)	2.4	-4.3 (1.4)	2.1
POST H.S.	103	78.8 (1.3)	0.8	207	75.2 (1.1)	3.4	-3.7 (1.7)	2.6
GRADUATED COLLEGE	257	79.9 (0.8)	0.3	371	79.3 (0.8)	1.6	-0.7 (1.1)	1.3
UNKNOWN	16	61.4 (2.7)	2.6	21	58.0 (4.7)	4.3	-3.4 (5.4)	1.7

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding, the N's for subgroups may not sum to the N for the total group.

Table 5.8

NAEP 1984 and 1986 Reading Assessments: Age 17
Weighted Mean Percents Correct and Percents Not Reached
for 23 Multiple-choice Items Common Between Bridges*

	1984 ASSESSMENT			1986 ASSESSMENT			DIFFERENCE 1986 - 1984	
SUBGROUP	N	% CORRECT	% NOT RCH	N	% CORRECT	% NOT RCH	% CORRECT	% NOT RCH
-- TOTAL --	2390	75.9 (0.4)	1.8	1901	73.5 (0.6)	2.5	-2.4 (0.8)	0.7
SEX								
MALE	1191	73.8 (0.6)	2.4	954	70.9 (0.8)	2.8	-2.9 (1.0)	0.4
FEMALE	1198	78.2 (0.4)	1.2	947	76.2 (0.8)	2.2	-2.0 (0.9)	1.0
ETHNICITY								
WHITE	1745	78.2 (0.5)	1.2	1341	76.2 (0.7)	1.6	-2.0 (0.8)	0.4
BLACK	339	67.7 (0.9)	3.4	319	65.2 (0.9)	4.9	-2.5 (1.3)	1.5
HISPANIC	225	68.8 (1.7)	4.1	192	62.4 (1.5)	6.4	-6.4 (2.2)	2.4
OTHER	80	73.7 (2.1)	3.4	48	66.4 (3.1)	4.4	-7.4 (3.7)	1.0
REGION								
NORTHEAST	536	77.1 (1.7)	1.1	383	75.7 (1.1)	2.3	-1.5 (2.0)	1.2
SOUTHEAST	601	75.2 (0.7)	1.6	487	70.6 (0.9)	2.4	-4.5 (1.1)	0.8
CENTRAL	680	75.8 (0.9)	1.8	492	74.6 (1.5)	1.7	-1.2 (1.7)	-0.1
WEST	573	75.6 (0.6)	2.8	539	72.6 (1.2)	3.7	-3.0 (1.3)	0.8
PARENTAL EDUCATION								
LESS THAN H.S.	287	69.5 (0.9)	2.7	164	64.0 (1.5)	4.7	-5.4 (1 7)	1.9
GRADUATED H.S.	846	73.0 (0.6)	1.8	528	69.7 (0.5)	2.4	-3.4 (0.8)	0.6
POST H.S.	347	78.8 (0.5)	1.3	427	75.9 (0.9)	1.8	-3.0 (1.0)	0.5
GRADUATED COLLEGE	809	80.9 (0 6)	1.3	703	78.7 (0.8)	1.9	-2.1 (1.1)	0.5
UNKNOWN	77	60.4 (1.3)	6.1	69	56.1 (2.2)	5.7	-4.4 (2.6)	-0.4

*Standard errors are given in parentheses. The "N" column gives the average number of students responding to each item. Because of rounding. the N's for subgroups may not sum to the N for the total group.

Table 5.9

Reading Scale Means and Standard Deviations for 1988 Bridges to 1984 and 1986 and for 1984 and 1986 Assessments*

	1984 Assessment				1988 Bridge to 1984				1986 Assessment				1988 Bridge to 1986			
	Mean	SE	SD	N of Items	Mean	SE	SD	N of Items	Mean	SE	SD	N of Items	Mean	SE	SD	N of Items
Age 9	211.0	(1.0)	41.1	126	212.1	(1.1)	40.2	99	208.9	(1.2)	39.6	31	214.0	(1.0)	40.9	30
Age 13	257.1	(0.7)	35.5	124	257.5	(0.9)	33.9	99	259.4	(1.0)	35.7	25	263.7	(0.8)	37.1	24
Age 17	288.8	(0.9)	40.3	113	289.9	(1.3)	37.4	87	277.4	(1.1)**	49.4	62	281.9	(1.4)	46.9	58

* All results are based on NAEP plausible values technology. Standard errors of means are given in parentheses. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

** Standard error differs from column 4 of Table 4.1 because of a change in jackknife methodology.

All multiple-choice items that appeared in the bridges were included in these analyses; for the bridge to 1984, the number of items per cohort ranged from 87 to 99--substantially larger than in the percents correct analysis.[5] The item parameters used in these bridge sample analyses, which are the same as those used in the corresponding assessment years, are listed in the 1984 and 1986 technical reports (Beaton, 1987 and Beaton, 1988b) for the bridge to 1984 and the bridge to 1986, respectively. Chapter 6 of this report includes a description of the procedures used originally to estimate the 1984 and 1986 item parameters.

Because of improvements in estimation procedures noted in Chapter 4, some of the results in Table 5.9 for the 1984 and 1986 assessments differ from those that appeared in *The NAEP 1985-86 Reading Anomaly: A Technical Report* (Beaton, 1988a). The 1984 scale means for ages 9 and 13 are lower than the previously reported results by roughly two points and one point, respectively, because of the adjustments to the 1984 student weights (Appendix C). The 1986 results at ages 9 and 13 differ from those previously reported because of the correction of a specification error in the conditioning procedures in 1986.[6]

[5] Professionally scored items were excluded from the bridge scales so that investigation of the reading anomaly would not be complicated by changes in scoring patterns for these items. The exclusion of the professionally scored items from the bridge scaling accounts for the difference between 1986 and the 1988 bridge to 1986 in the number of items scaled. The number of items in the bridge to 1984 is substantially less than the number of items in the 1984 assessment because only a subset of the 1984 booklets was used in this bridge.

[6] At ages 9 and 13, the 1986 reading assessment included both a balanced incomplete block (BIB) spiral component and a bridge to 1984. For each of these two cohorts, the BIB and bridge samples were combined for purposes of generating plausible values. In estimating the conditional distributions, an indicator variable for sample membership (BIB or bridge) was included among the conditioning variables (see Mislevy, 1988). Also included was a variable that reflected whether students were above, at, or below modal grade. However, the modal grade was not the same for the BIB and bridge samples. The conditioning model was mis-specified in that it did not allow for an interaction between this

Neither the weight adjustment for 1984 results. nor the correction to the estimation procedure in 1986 affected the results for age 17, where the anomaly was the most pronounced.

<u>Figure 5.1</u> displays the NAEP reading scale results since 1971.

<u>Figures 5.2 - 5.4</u> show overlay graphs of the estimated distributions for the two 1988 bridge samples at each age level.[7]

1. <u>Comparison of the 1988 Bridge Samples</u>

The main findings of the comparison of the 1988 bridges to 1984 with the 1988 bridges to 1986 are these:

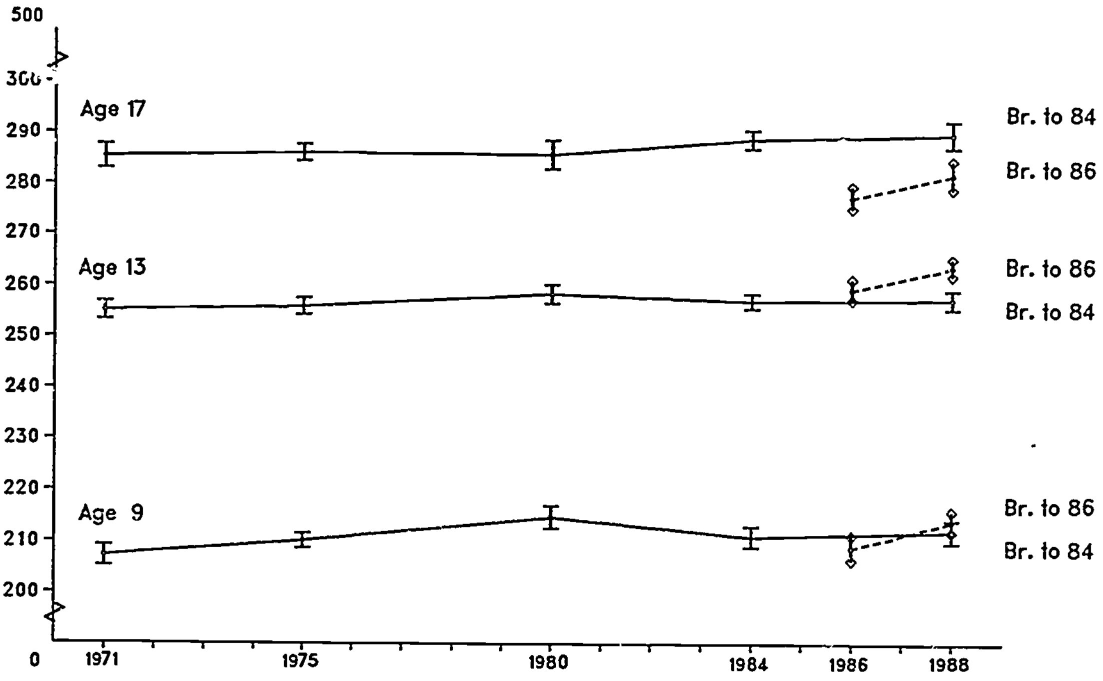
- At age 9, the mean percent correct for the two bridges was the same; the reading scale mean for the bridge to 1986 was slightly higher.
- At age 13, the performance of the bridge to 1986 was superior both in terms of mean percents correct and in terms of scale means.
- At age 17, the performance of the bridge to 1986 was inferior, both in terms of percents correct and in terms of scale means.

variable and the sample membership variable. (Since the age 17 assessment included only a BIB component, estimation of the age 17 results was not affected.) The problem was corrected by conditioning the BIB and bridge samples separately. The effect on the BIB results that are reported in *Who Reads Best?* (Applebee, Langer, & Mullis, 1988) was almost nil; the effect on the bridges, more substantial. The corrected mean for age 9 is 1.6 points higher than the previously reported value, whereas the corrected mean for age 13 is one point lower. At both ages, the standard deviations are about 5 points lower; unlike the earlier estimates, they do not appear large relative to 1984.

[7] Figures 5.1 to 5.4 are identical to Figures 4.1 to 4.4. They are reproduced here for convenience. These figures are based on all five sets of plausible values.

Figure 5.1

Reading Scale Results
1971 - 1988*



* Standard errors of means are approximately 1.0. Bands extend from two standard errors below to two standard errors above the mean. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

Weighted Reading Proficiency Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
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	211.0 (1.0)	257.1 (0.7)	288.8 (0.8)
1986	208.9 (1.2)	259.4 (1.0)	277.4 (1.1)**
1988 Br. to 84	212.1 (1.1)	257.5 (0.9)	289.9 (1.3)
1988 Br. to 86	214.0 (1.0)	263.7 (0.8)	281.9 (1.4)

** Standard error differs from column 4 of Table 4.1 because of a change in jackknife methodology.

Figure 5.2

Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 9

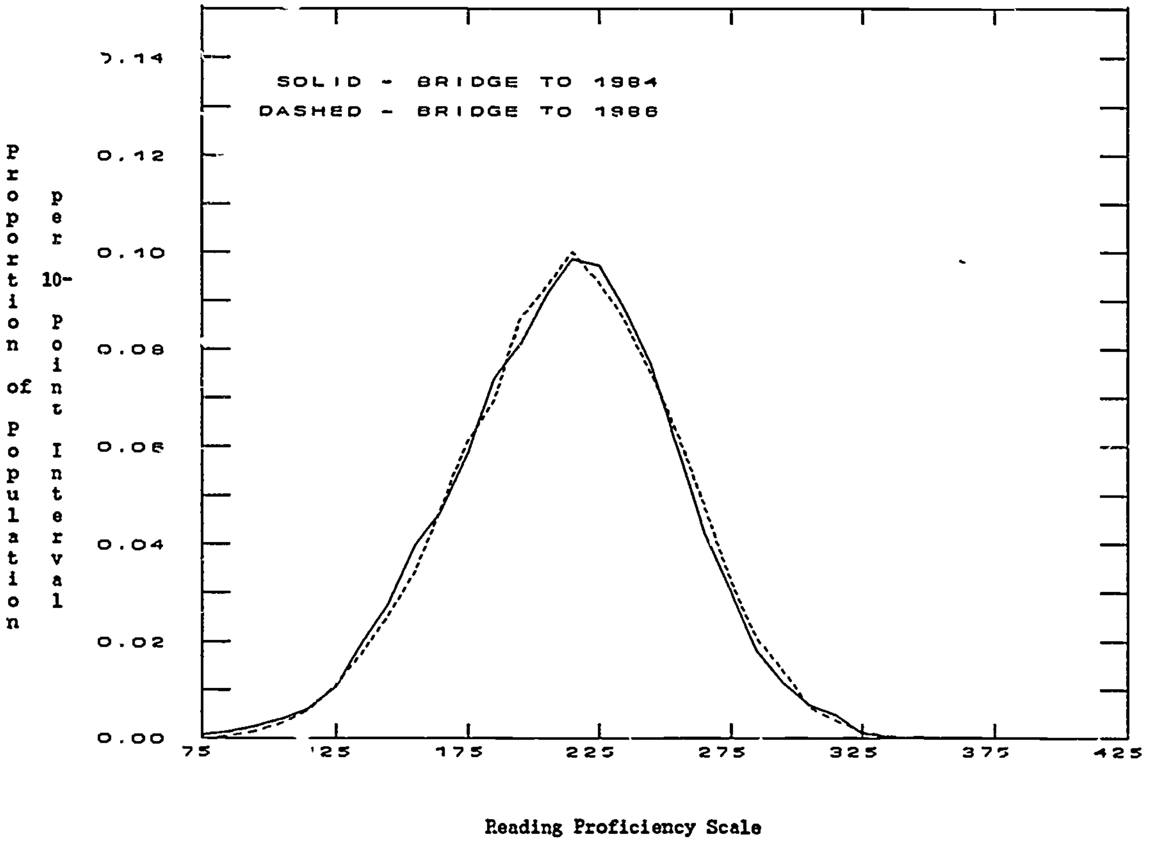
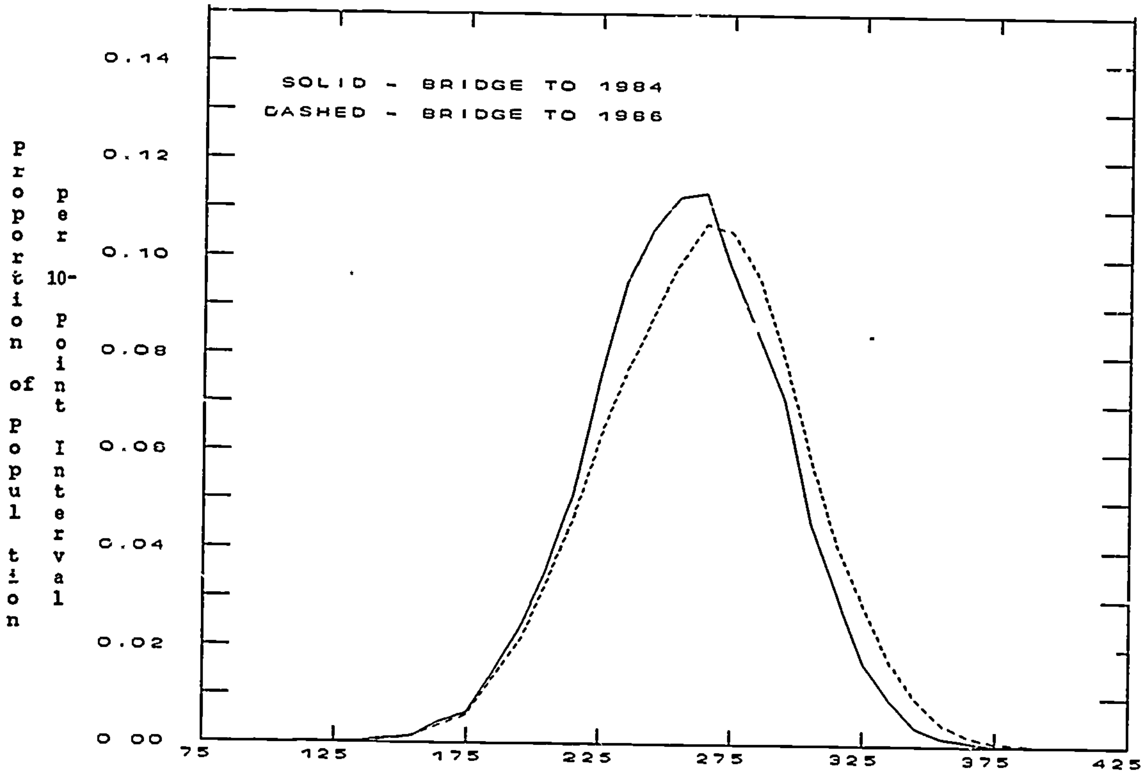


Figure 5.3

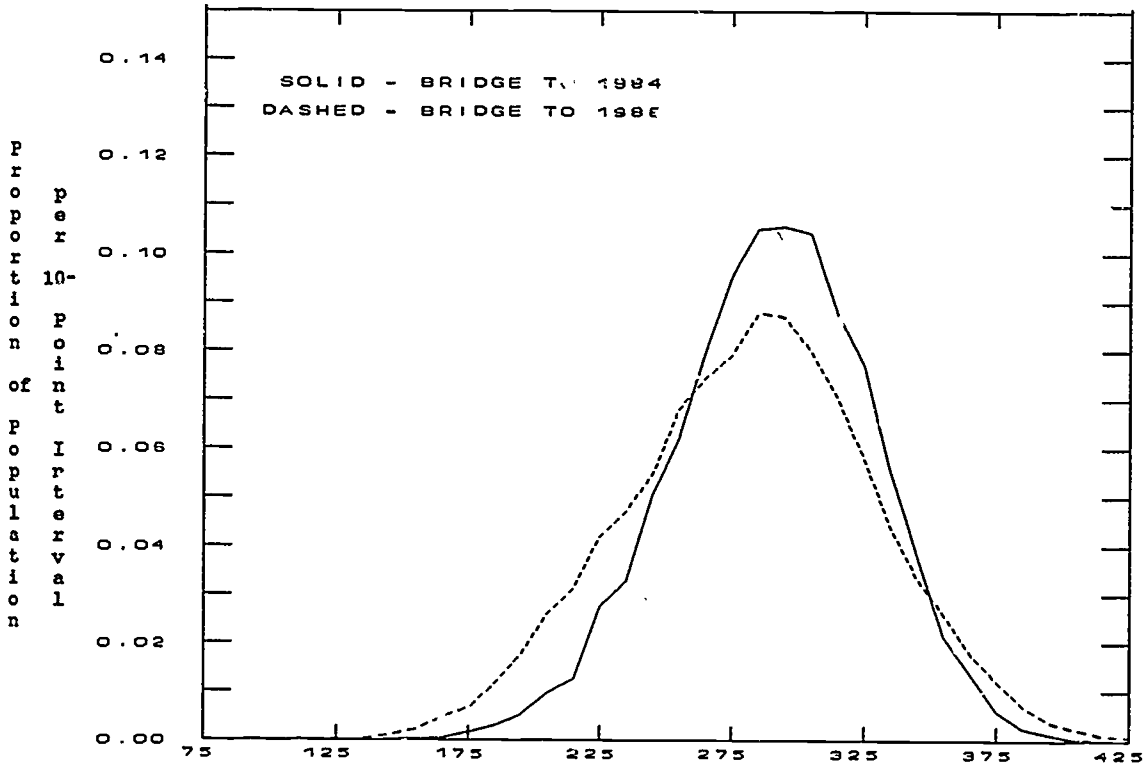
Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 13



Reading Proficiency Scale

Figure 5.4

Estimated Reading Proficiency Distributions for 1988 Bridge Samples, Age 17



Reading Proficiency Scale

At all three ages, the variance of the scale distributions was greater for the bridge to 1986 than for the bridge to 1984. The difference is particularly notable at age 17. The large difference in variances at age 17 is similar to that observed in comparing the actual 1984 and 1986 assessments. The graphs of distributions in Figures 5.2 through 5.4 show that the bridge to 1986 is characterized by a heavier upper tail than the bridge to 1984 at age 13 and heavier upper and lower tails at age 17. These findings parallel those obtained by comparing the 1984 and 1986 assessments. (Graphs of the 1984 and 1986 proficiency distributions are included in Chapter 6.)

There is some evidence of differential effects across subgroups. For example, there is a tendency for lower-scoring groups at age 17 to be more disadvantaged by the bridge to 1986 conditions than higher-scoring groups. This result parallels to some degree the findings in the actual 1984 and 1986 assessments.

An additional finding at age 9, which probably accounts for the slight inconsistency between the percents correct and the scale value results, is evidence of speededness of both the 1984 and 1986 instruments. On 18 of 26 items, the percentage of students who did not reach the item exceeded 10 in at least one of the bridge samples. These results parallel those obtained in the actual 1984 and 1986 assessments. Because the items were ordered differently in the two bridges (reflecting the corresponding prior assessments), the speededness tends to affect different items in each bridge. The median of the absolute differences between the bridges in not-reached percentages was 11.5; the corresponding value for the 1984 and 1986 assessments was 11.0.

Not-reached items are treated in different ways in the percents-correct analysis and the scaling analysis. In NAEP, the percent correct on an item is based only on students who reached the item. The implicit assumption is that students who do not reach the items have the same probability of answering correctly as students who do reach the items. In fact, the relation between percent not reached and percent correct is not a simple one: On about one-third of the age 9 items, the bridge sample with the greater not-reached percent had a higher percent correct on the item. In scaling, the treatment of not-reached items is consistent with the assumption that students who do not reach an item have the same probability of answering correctly as other students *who did reach the item and who gave the same responses to the preceding items and had the same values on the conditioning variables.*

In summary, there is evidence of differences between the 1984 and 1986 forms and administration procedures, but the effects are not the same at all three ages. The lack of consistency across ages is not surprising in light of the fact that changes in forms between 1984 and 1986, including, for example, the degree to which item positions were shifted, were not uniform across the age groups. The effects of these kinds of changes are discussed further in Chapters 6 and 8.

2. Comparison of the 1984 Assessment with the 1988 Bridge to 1984

The findings of the comparison of the reading results from the 1984 assessment to those from the corresponding 1988 bridge show slight evidence of an increase at age 9. There is little or no change, in terms of either mean percents correct or scale means at ages 13 and 17. The standard deviations of the scale values are also quite similar.

3. Comparison of the 1986 Assessment with the 1988 Bridge to 1986

Whereas the comparison of reading scale means for 1984 to those in the 1988 bridge to 1984 suggests that reading achievement has remained quite stable during the last four years, comparison of the 1986 assessment to its corresponding 1988 bridge reveals sizeable increases in scale means. The difference is largest at age 9, where the bridge mean exceeds the 1986 mean by 5.1 scale points. (The corresponding change of three points in mean percent correct is also very large.) In considering these differences, it is important to note that, for reading assessments that occurred between 1971 and 1984, the largest change between successive assessments was 4.6 scale points. This change took place in a five-year interval (1975 to 1980) at age 9 (see Beaton, 1988a, p. 7). This makes a five-point change in two years appear unlikely.

Tr^re ar(two ways in which the bridge to 1986 is known to differ from the 1986 assessment. First, because of the relatively small size of the bridge samples, it was not possible to re-create the large assessment sessions that occurred in some instances in the 1986 assessment of 17-year-olds. However, this explanation seems unlikely to account for the relatively steep rise between 1986 and 1988, particularly since a small investigation of effects of session size in the 1986 assessment showed a slight tendency for medium sessions (26-50 students) to have lower results than small (1-25 students) or large (more than 50 students) sessions.

The other known difference between 1986 and 1988 is that, as shown in Table 5.1, the times of testing for the bridge to 1986 were slightly different from the time of testing for the 1986 assessment. This occurred because it was desirable to match the times of testing for the two sets of 1988 bridge samples. The 1988 main NAEP assessment affords an opportunity to assess the effects of time of testing, since two random half-samples were assessed at each age, one in the winter and one in the spring. On the average, the spring samples were tested about two months later than the winter samples. At age 9, which displayed the largest difference between the half-samples, the mean reading proficiency increased 2 scale points between winter and spring, with a standard error of 2.1. The difference at age 13 was 0, with a standard error of 2.1, and at age 17, there was a drop of 0.8 points between winter and spring, with a standard error of 1.8. Clearly, within-year changes were not large. Interpolations based on these results suggest that the changes in time of testing between 1986 and 1988 would have had little impact.

An additional hypothesis that was considered was that teaching to the test may have occurred in schools that were included in both the 1986 and the 1988 bridge to 1986 assessments. However, an investigation showed that only two schools were included in both assessments. It is therefore implausible that teaching to the test could have contributed to the large gains between 1986 and 1988.

The large differences between the 1986 assessment and the 1988 bridge to 1986, which are paralleled in the mathematics and science results (see Chapter 7) suggest that there were aspects of the 1986 assessment that were not duplicated in the 1988 bridge to 1986.

Summary

The comparison of the 1988 bridge samples yielded the following conclusions:

- The 1986 instruments and conditions appear to have been advantageous to 13-year-olds and disadvantageous to 17-year-olds, relative to the 1984 assessment.
- At age 9, the percents of students who failed to reach certain items were substantially different in the two assessments.
- Based on the 1988 bridge to 1984, there appears to have been little change in reading proficiency between 1984 and 1988.

Despite the somewhat puzzling gain between 1986 and the 1988 bridge to 1986 at age 9, the findings of these bridge comparisons suggested that it would be useful to pursue the idea of using the bridge data to equate the 1984 and 1986 results. The ensuing analyses are described in Chapter 6.

Chapter 6

ADJUSTMENT OF THE 1986 READING RESULTS TO ALLOW FOR CHANGES IN ITEM ORDER AND CONTEXT

Rebecca Zwick[1]

Overview

Although the potential effect of item context on proficiency estimation has been discussed in the measurement literature (see Leary & Dorans, 1985, for a review and Wise, Chia, & Park, 1989, for a recent example), the prevailing view has been that item-parameter estimates derived through item response theory (IRT) methods are relatively robust to changes in item context. Current testing practices, such as item banking and adaptive testing, as well as IRT common-item equating methods, such as the one applied to NAEP reading data in 1986, rest on the assumption of invariance of item parameters across different test forms.

The analyses described in Chapters 5, 6, and 8 of this report show that in the case of the 1984 and 1986 NAEP assessments, the effects of changes in item context, position, and administration conditions were large enough to produce significant differences in item functioning, which, in turn, led to a violation of the item-parameter invariance assumption. One manifestation of the impact of changes in item position on item functioning was the large difference between the two assessments in the percents of students reaching

[1] Analysis plans for the adjustment method described in this chapter were developed in collaboration with Albert Beaton and Robert Mislevy. David Freund provided statistical programming, with assistance from Minhwei Wang and Kate Pashley. David Freund and Jo-Ling Liang produced the figures in this chapter.

certain items. Neither IRT nor any other method in existence can yield invariant parameters under these circumstances. In theory, more complex models could be developed that would take into account explicitly any changes in item position or context, but such models are unlikely to be available in the near future. Our findings have led us to appreciate the need to avoid changes in instruments and procedures when assessing trend.

The value of the 1988 bridge data is that it made possible the equating of the 1984 and 1986 instruments without reliance on common-item assumptions. By using common-population equating, rather than common-item equating, we could allow for the possibility that items common to 1984 and 1986 were, in fact, functioning as different items in each assessment and that form and administration changes may have had a different impact at each age level.

Common-population equating is possible when two random samples from the same population are available. The equating is achieved by matching certain properties of one sample proficiency distribution (in this case, the first two moments) to those of the other sample distribution, as described in detail below. The transformation of the proficiency scale that achieves this match implies a set of transformations for the item parameters. In contrast, our attempt in 1986 to link the 1986 results to the 1984 reading scale through common-item equating was based on the assumption of item-parameter invariance. The analyses of the bridge data reported in this chapter not only permitted us to investigate the impact of relaxing this invariance assumption but yielded new item parameters for the 1986 instrument which were used to adjust the 1986 results. To explain how this was done, it is necessary to first describe the estimation of reading item parameters in 1984 and 1986.

Estimation of Reading Item Parameters in 1984 and 1986

In 1984, item-parameter estimates were obtained simultaneously for reading items administered to all three age cohorts. The initial estimates of item parameters and reading proficiency, which were on an arbitrary scale, were linearly transformed to produce a reading proficiency scale with a mean of 250.5 and a standard deviation of 50 across all three cohorts (see Mislevy & Sheehan, 1987).

In 1986, reading items were administered to balanced incomplete block (BIB) spiral samples for all three cohorts and to bridge samples to 1984 at ages 9 and 13. Students from all five of these samples (along with students in a special language minority study [Baratz-Snowden, Rock, Pollack, & Wilder, 1988]) were included in a single item calibration (see Zwick, 1988b). The steps for obtaining the 1986 item parameters were as follows:

1. Assume the three-parameter logistic model,

   $P(X_i = 1|\theta) = c_i + (1 - c_i) \{1 + \exp [-1.7a_i(\theta - b_i)]\}^{-1}$,

   where X_i is the response to item i, which is scored "1" if correct, θ represents ability, a_i is a discrimination parameter for item i, b_i is a difficulty parameter, and the lower asymptote, c_i, represents the probability that a very low-ability examinee answers the item correctly. Use the BILOG program (Mislevy & Bock, 1982) to obtain provisional parameter estimates (designated 86-P) for all items for all three cohorts, ignoring the distinction between old and new items.

2. Apply the Stocking-Lord procedure (see Sheehan, 1988) to the items common to 1984 and 1986 in order to find the best-fitting linear transformation for mapping the 86-P parameters to the 1984 parameters.

3. Use these transformed parameters (designated L(86-P) to indicate a linear transformation of the 86-P parameters) for the items that were new to 1986.

4. Substitute the original 1984 parameters for the items that dated back to 1984[2].

The 1986 bridges-to-1984 at ages 9 and 13 included only items that dated back to 1984. Therefore, the item-parameter estimates used to obtain results for these two cohorts were a subset of the estimates obtained in 1984. However, at age 17, trend results were to be based on the age 17 subsample of the BIB-spiraled assessment, which received 43 items from 1984 and 19 new items. Therefore, the age 17 results were based on 1984 parameters for the 43 old items and L(86-P) parameters for the 19 new items. The reading results at age 17 may have appeared particularly anomalous because the 19 new items were linked into the scale using a common-item equating procedure that, we have since learned, was inappropriate for linking the 1984 and 1986 measures, despite its frequent use in similar applications outside NAEP.

[2] An alternative would have been to [illegible] the L(86-P) parameters for these items as well, or to use some weighted combination of the original 1984 and L(86-P) parameters. In fact, we did estimate the reading mean for age 17 using the L(86-P) parameters for *all* items and found the mean to be even lower than the reported anomalous mean.

1988 Adjustment Procedure

To derive an adjustment of the 1986 reading results, we took advantage of the fact that, for each cohort, the 1988 bridges provided reading data for the 1984 and 1986 instruments and procedures based on random samples from the identically defined populations of 9-, 13-, and 17-year-olds. We were, therefore, able to abandon common-item equating procedures and make use of common-population equating. Then, under the assumption that the relation between the 1984 and 1986 measurement systems was stable over time, it was possible to use the equating functions to adjust the 1986 reading results. The steps in the adjustment procedure were as follows:

1. Use the RESOLVE program (which implements the procedures described in Mislevy, 1984) to obtain a nonparametric estimate of the proficiency distribution for each cohort of the 1988 bridge to 1986, using the 86-P parameters. (The results in Table 5.9 for the 1988 bridge to 1986 rely on common-item assumptions and could not, therefore, serve as the basis of the equating.)

2. For each cohort, obtain the linear transformation that matches the mean and standard deviation for the 1986 bridge, obtained from Step 1, to those of the 1984 bridge. (Note that the linear transformations were not constrained to be the same from one age to another.)

3. Use these transformations to adjust the 86-P item parameters, yielding three new sets of item parameters, $L'_9(86)$, $L'_{13}(86)$, and $L'_{17}(86)$ (different linear transformations at each age). These adjusted parameters appear in Tables D.2 through D.4 in Appendix D.

4. Apply these $L'_9(86)$, $L'_{13}(86)$, and $L'_{17}(86)$ parameters to the reading data collected in 1986 to reestimate the 1986 reading scale values, using plausible values methodology (Mislevy, 1988), a method of estimating proficiency distributions based on students' item responses and background characteristics (referred to in this context as conditioning variables). The same conditioning variables and coding scheme were used as in the original 1986 analysis (Mislevy, 1988, p. 198). The estimated coefficients for the conditioning variables appear in Tables D.5 through D.7 in Appendix D.

Results of the Adjustment

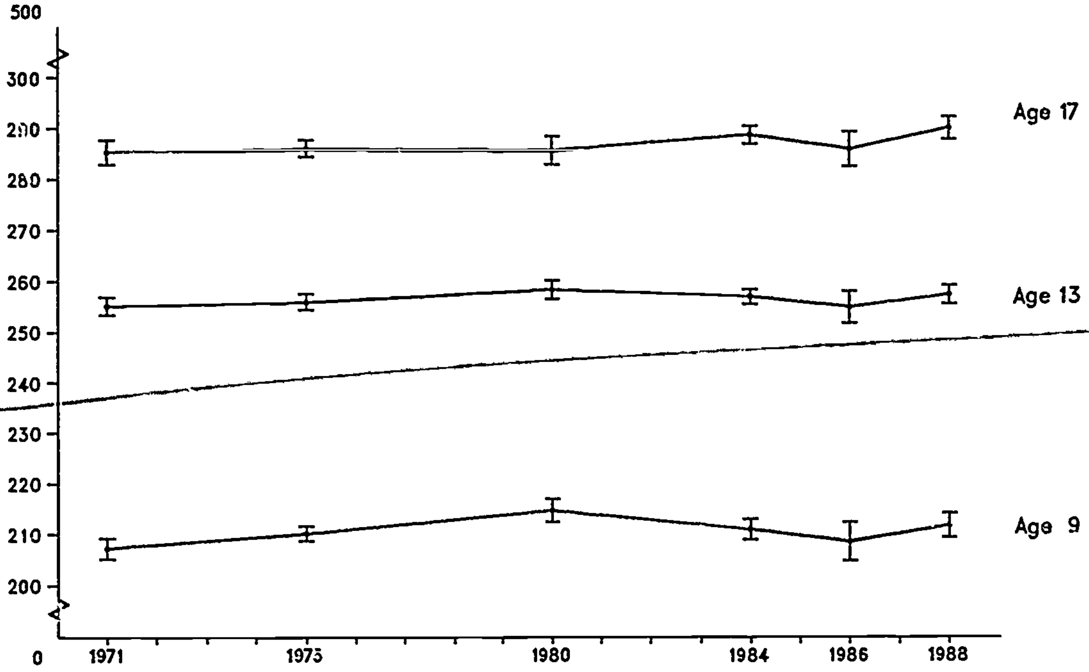
The reading scale means from the adjustment procedure are shown in Figure 6.1[3], along with earlier reading results and 1988 results based on the bridge to 1984. The means and standard deviations for 1984, 1986, and the adjusted 1986 results are also given in Table 6.1 for each cohort.[4] (The

[3] This figure is identical to Figure 1.2. It is repeated here for convenience.

[4] For reasons described in Chapters 4 and 5, the 1984 and 1986 results for ages 9 and 13 differ from previously published results. The 1984 results in Table 6.1 incorporate the adjustment in sample weights; the original 1986 results incorporate the modification in the conditioning model. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

Figure 6.1

Modified Results
Reestimated Average Reading Performance, 1971 - 1988
(and standard errors*)



*Bands extend from two estimated standard errors below to two estimated standard errors above the mean. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

Modified Results
Weighted Reading Proficiency Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
1971	207.3 (1.0)	255.2 (0.9)	285.4 (1.2)
1975	210.2 (0.7)	256.0 (0.6)	286.1 (0.8)
1980	214.8 (1.1)	258.5 (0.9)	285.8 (1.4)
1984	211.0 (1.0)	257.1 (0.7)	288.8 (0.9)
1986 Adjusted	208.6 (1.9)	255.0 (1.6)	286.0 (1.7)
1988	211.8 (1.2)	257.5 (0.9)	290.1 (1.1)

Table 6.1

Results of Adjustment of the 1986 Reading Scale[a]

	1984[b]			Unadjusted 1986[c]			Adjusted 1986		
	Mean	S.E.	S.D.	Mean	S.E.	S.D.	Mean	S.E.	S.D.
Age 9	211.0	(1.0)	41.1	208.9	(1.2)	39.6	208.6	(1.9)	38.6
Age 13	257.1	(0.7)	35.5	259.4	(1.0)	35.7	255.0	(1.6)	34.7
Age 17	288.8	(0.9)	40.3	277.4	(1.1)[d]	49.4	286.0	(1.7)	39.5

[a] Standard errors are given in parentheses. See Appendix E for computation of standard errors for adjusted means. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

[b] Results incorporate the adjustment to the sampling weights.

[c] Results incorporate the modification in the conditioning model.

[d] Standard error differs from column 4 of Table 4.1 because of a change in jackknife methodology.

standard errors for the adjusted means take into account the error associated with estimating the equating functions. Details of the standard error computations are given in Appendix E.) Overlays of the 1984 and adjusted 1986 distributions for each cohort appear in Figures 6.2 through 6.4. These can be contrasted with the overlays of the 1984 and original 1986 distributions for each cohort that appear in Figures 6.5 through 6.7.[5]

The result of the adjustment is most dramatic at age 17, where the mean increased by 8.6 points and the standard deviation was reduced by 9.9 points. At age 13, the mean decreased by 4.4 points. The adjustment reduced the heavy upper and lower tails at age 17, as well as the heavy upper tail at age 13. The mean for age 9 stayed virtually the same, as did the standard deviations at ages 9 and 13. The changes from 1984 to the adjusted 1986 results are quite consistent across ages. The decreases in the means are 2.4, 2.1, and 2.8 for ages 9, 13, and 17, respectively.

Figures 6.8 through 6.10 give another perspective on the results of the adjustment for each of the three cohorts. To construct these graphs, 2,000 students were selected within each cohort. A proficiency estimate for each student was computed[6] based on both the original and the adjusted item parameters. The original values were plotted along the x-axis; the adjusted values, along the y-axis. The changes in the tails of the distribution at age 17 are very evident in Figure 6.10.

[5] Figures 6.2 to 6.7 are based on all five plausible values.

[6] The proficiency estimates used in this graph are the mean plausible values. Note that the same graph could have been obtained by estimating each point separately for each of the five sets of plausible values and obtaining a final estimate by averaging the five sets of x-coordinates and the five sets of y-coordinates of these points. The method used here is, therefore, consistent with NAEP's analysis recommendations (see Mislevy, 1988).

Figure 6.2

Estimated Reading Proficiency Distributions for 1984 and 1986 (Adjusted)
Age 9

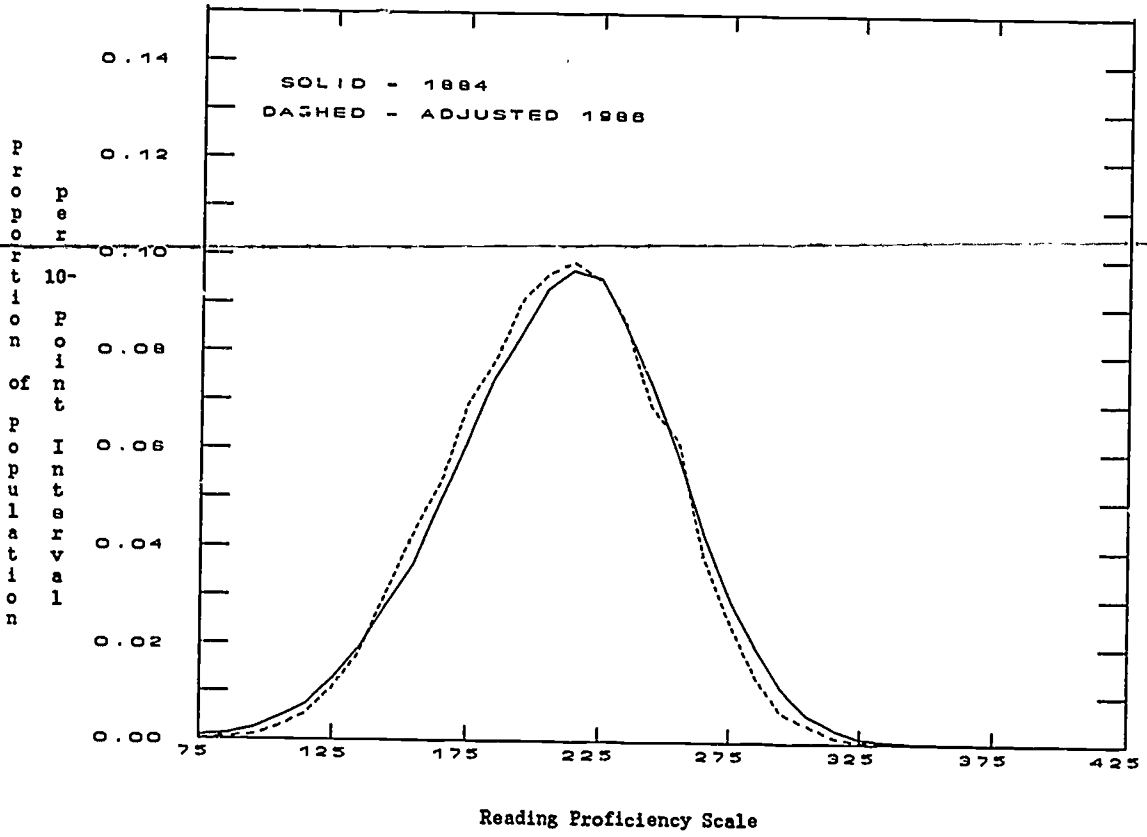
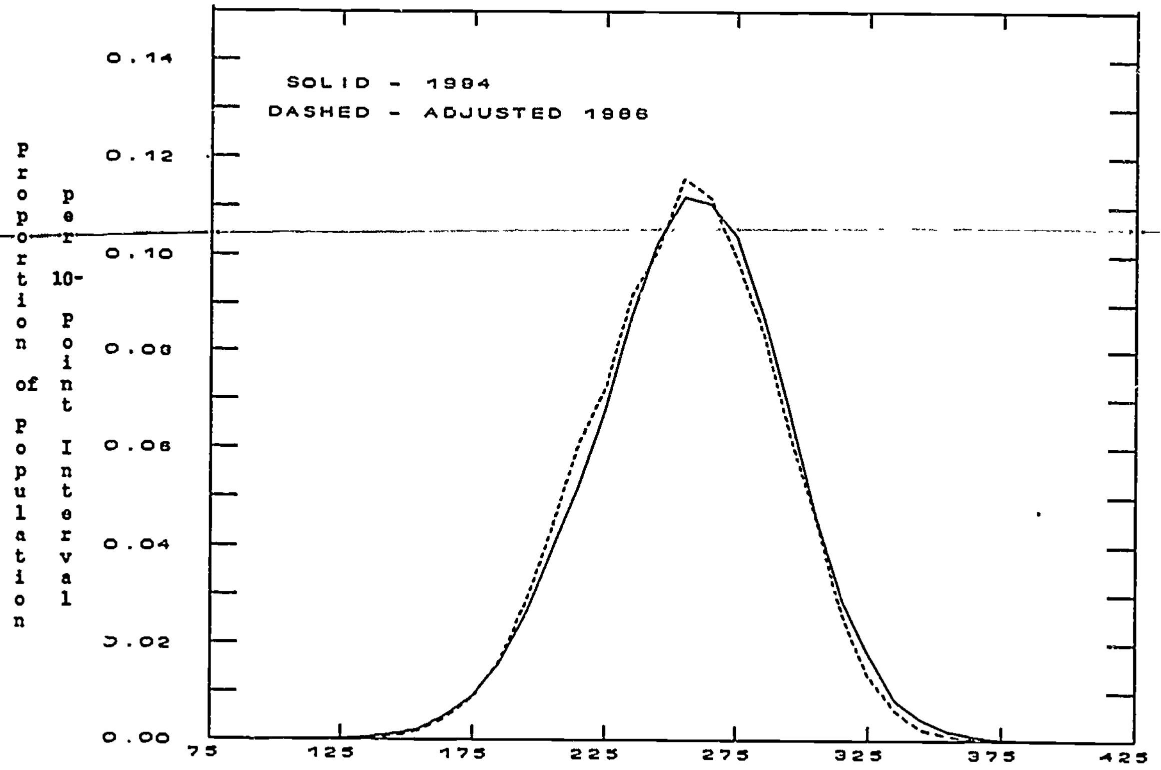


Figure 6.3

Estimated Reading Proficiency Distributions for 1984 and 1986 (Adjusted)
Age 13



Reading Proficiency Scale

Figure 6.4

Estimated Reading Proficiency Distributions for 1984 and 1986 (Adjusted)
Age 17

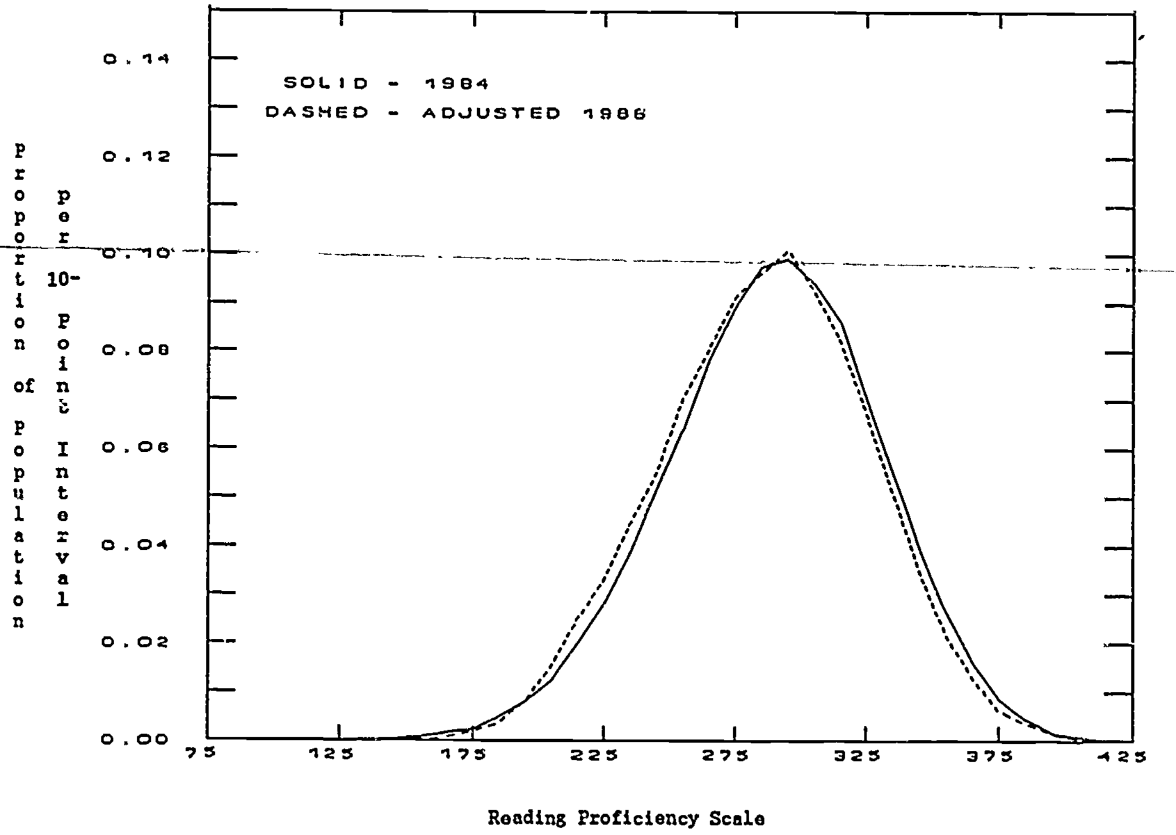


Figure 6.5

Estimated Reading Proficiency Distributions for 1984 and 1986 (Original)
Age 9

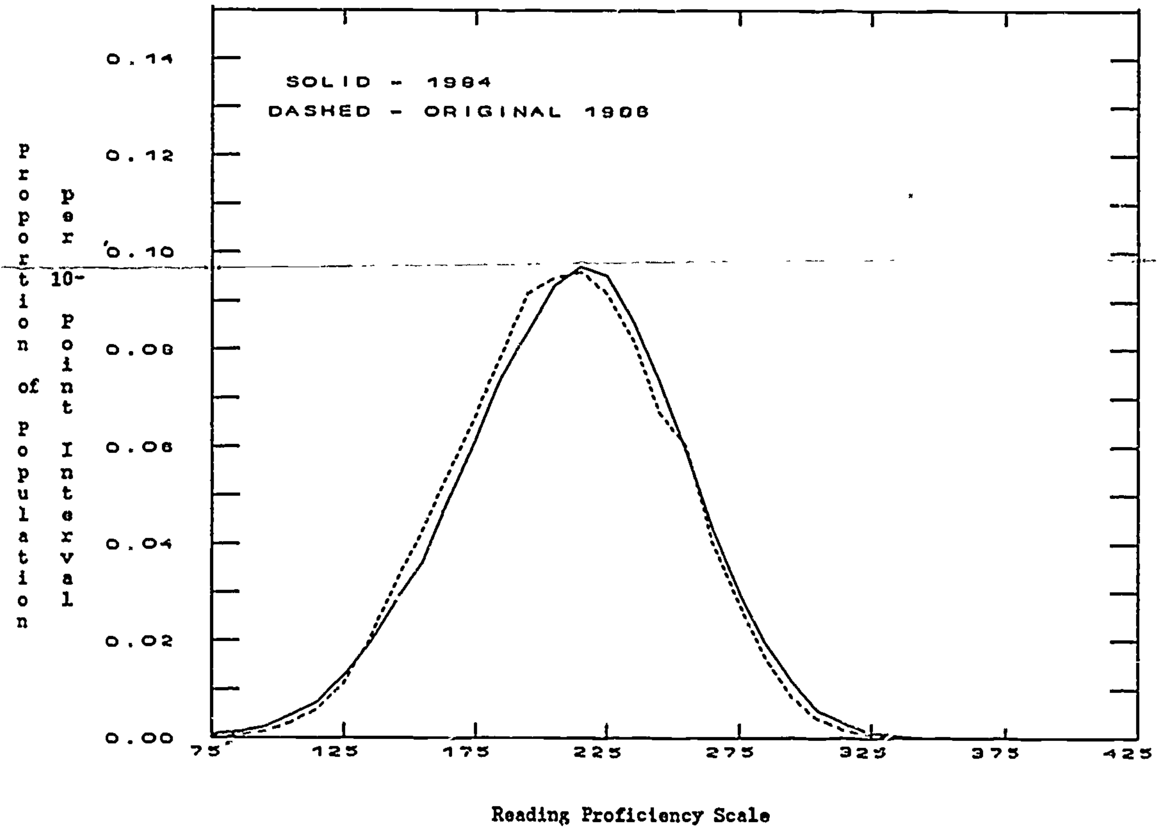


Figure 6.6

Estimated Reading Proficiency Distributions for 1984 and 1986 (Original)
Age 13

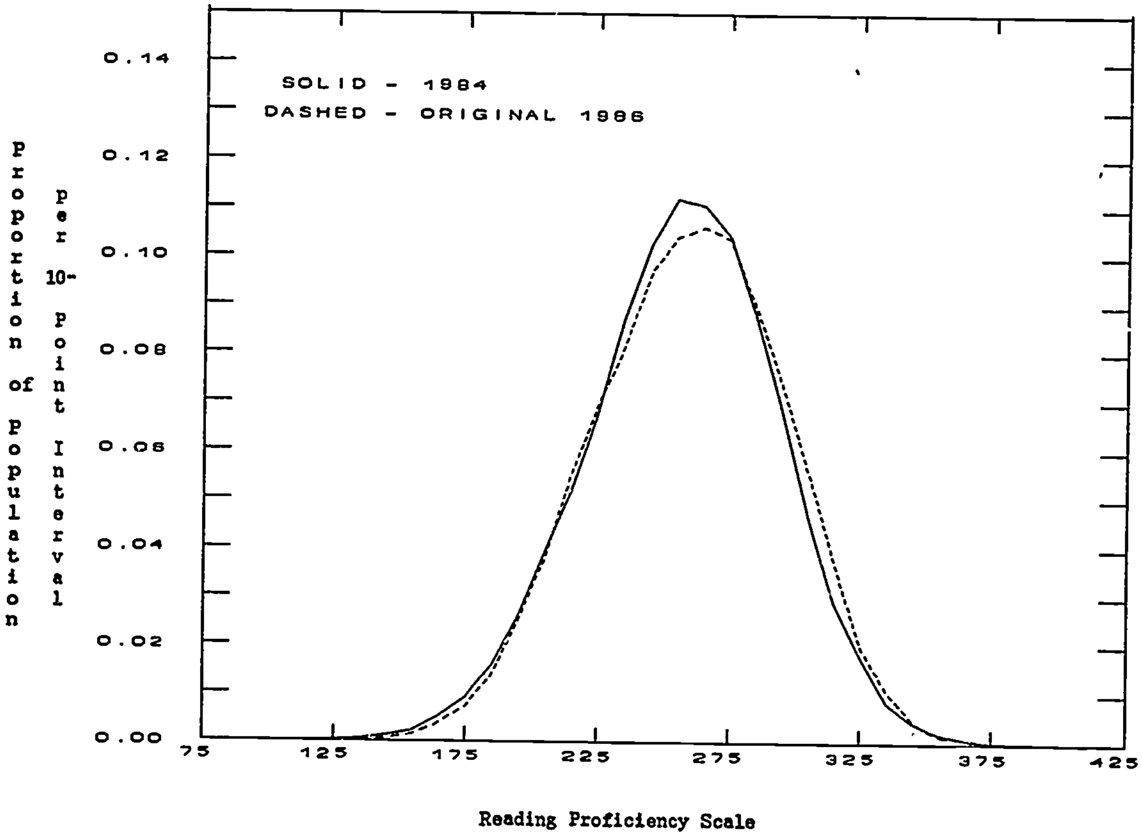


Figure 6.7

Estimated Reading Proficiency Distributions for 1984 and 1986 (Original)
Age 17

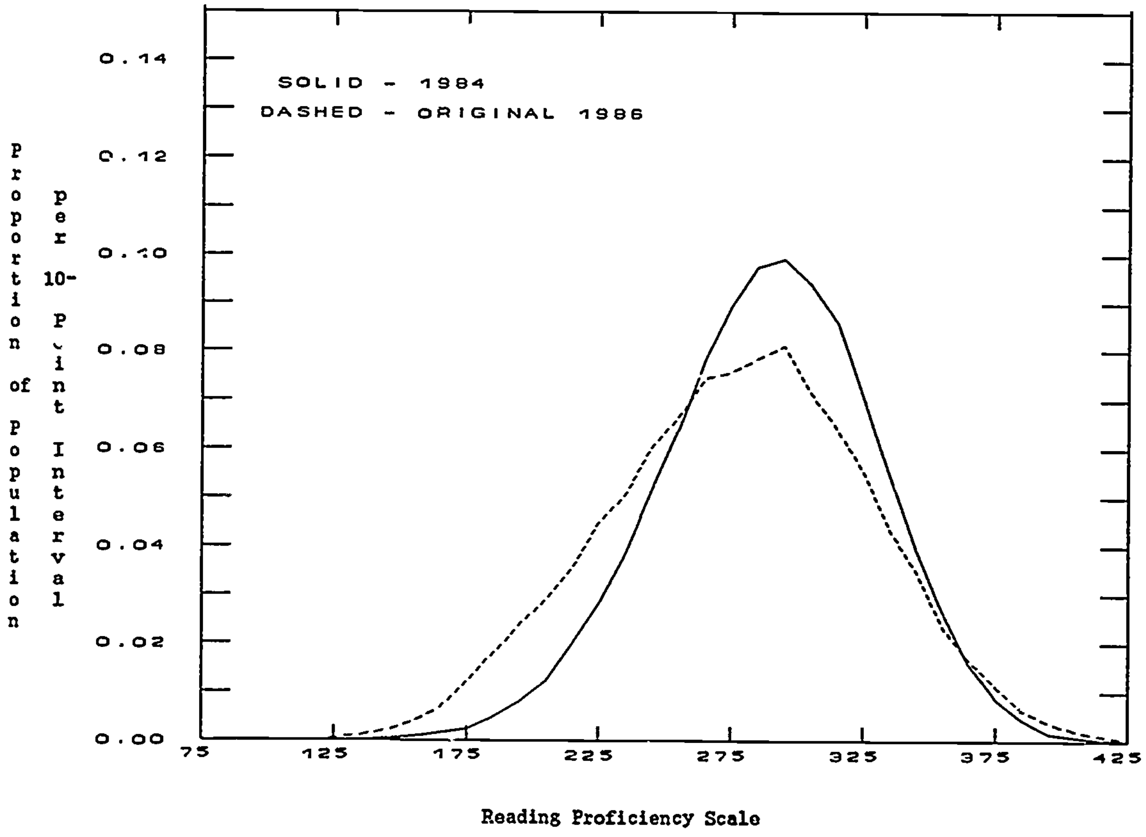


Figure 6.8

Relation Between Reading Scale Values Constructed with Original and Adjusted Item Parameters, Age 9

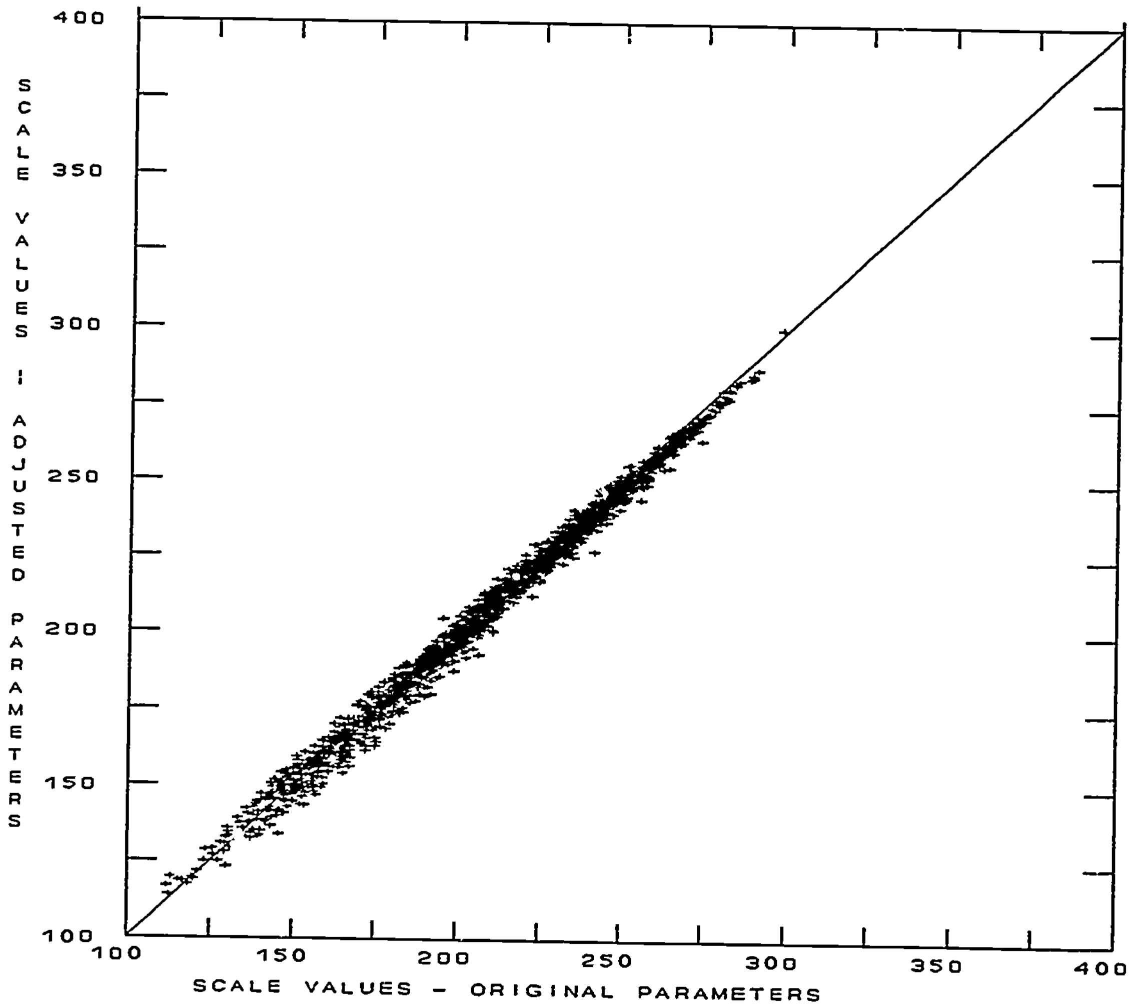


Figure 6.9

Relation Between Reading Scale Values Constructed with Original and Adjusted Item Parameters, Age 13

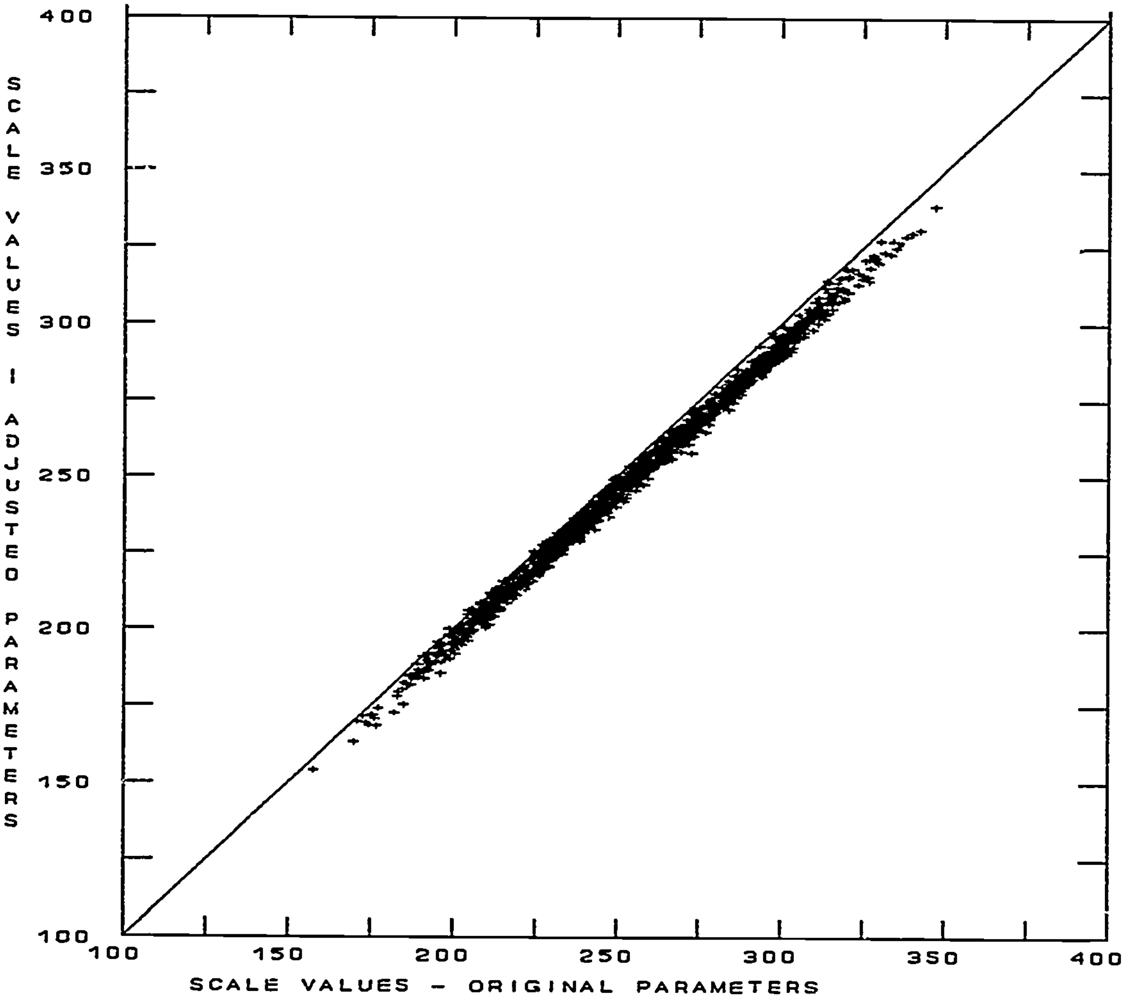
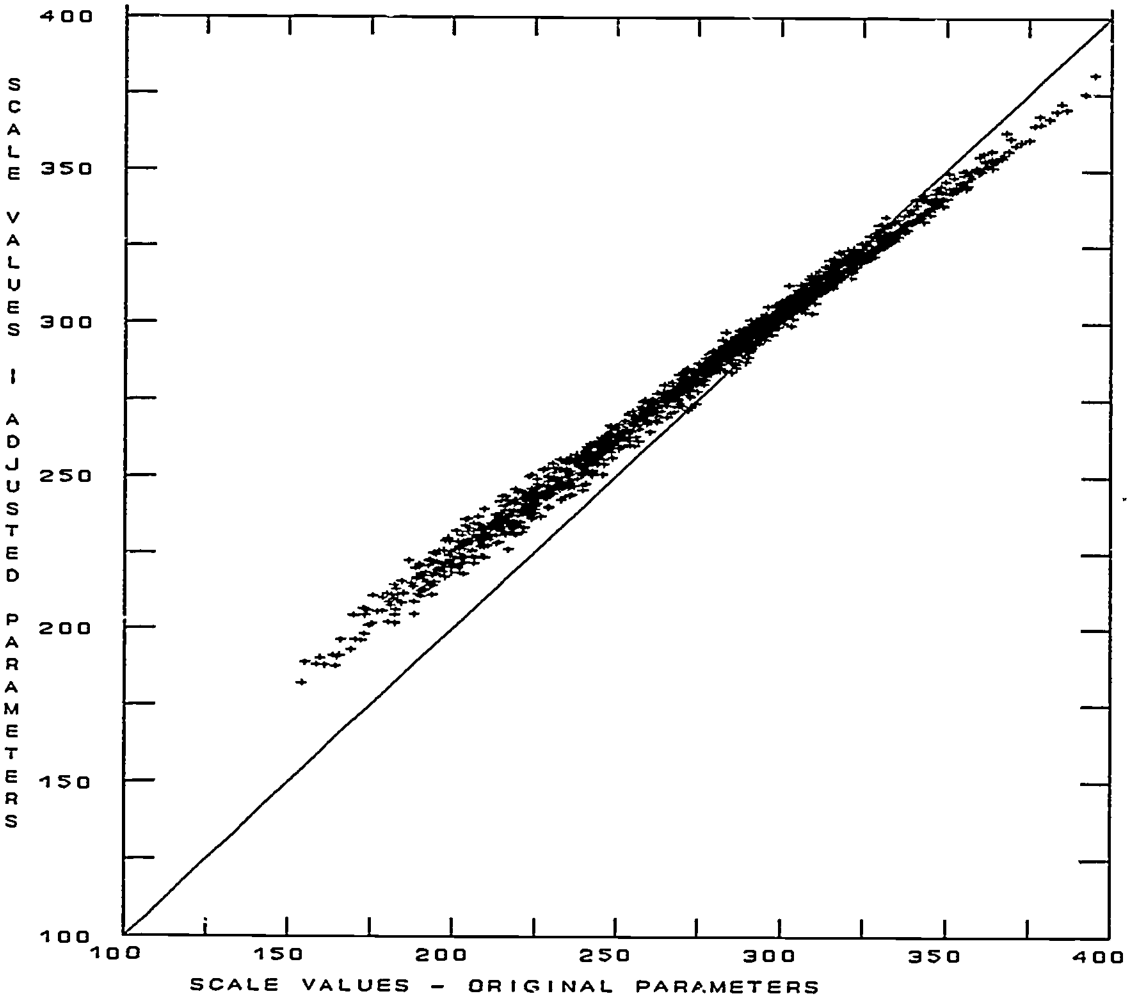


Figure 6.10

Relation Between Reading Scale Values Constructed with Original and Adjusted Item Parameters, Age 17



Adjusted means for subgroups are given in Tables 6.2 through 6.4. At age 17, the adjustment led to some substantial changes in the patterns of subgroup differences. For example, as shown below, the adjustment increased the mean for Black students by 13.2 points and the mean for Hispanic students by 12.3, resulting in values close to those observed in 1984. Because the adjustment produced an increase of only 7.5 points for White students, the differences between White students and minority students in the adjusted results are smaller than in the unadjusted results; in fact, they are smaller than those observed in 1984. Also, the adjustment led to an increase of about 15 points for the students below modal grade, whereas the mean for students at modal grade increased seven points and the mean for students above modal grade increased five points. The adjusted 1986 results resemble closely the 1984 results.

	1984		1986		Adjusted 1986	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
White	295.6	(0.7)	283.9	(1.2)	291.4	(1.8)
Black	264.2	(1.2)	251.8	(1.5)	265.0	(2.0)
Hispanic	268.1	(1.9)	255.3	(2.6)	267.6	(2.6)
Below Modal Grade	258.8	(1.2)	242.8	(1.3)	257.7	(1.9)
At Modal Grade	295.7	(0.7)	286.0	(1.0)	293.1	(1.7)
Above Modal Grade	303.8	(1.3)	296.0	(2.7)	301.0	(2.6)

Summary

Our analysis of the 1988 bridge data helped us to understand the reading anomaly and also provided a means of adjusting the 1986 results. The existence of the bridge data allowed us to equate the 1984 and 1986 instruments through common-population equating, rather than common-item equating. Using the transformed item parameters that resulted from the

Table 6.2

Reading Trend Results Including 1986 Adjusted Values: Age 9
(standard errors in parentheses)

	1971	1975	1980	1984	Unadjusted 1986	Adjusted 1986 [a]	1988 [b]
-- TOTAL --	207.3(1.0)	210.2(0.7)	214.8(1.1)	211.0(1.0)	208.9(1.2)	208.6(1.9)	211.8(1.2)
SEX							
MALE	200.9(1.1)	204.4(0.8)	209.7(1.3)	207.7(1.1)	204.7(1.3)	204.5(1.9)	207.5(1.5)
FEMALE	213.7(1.1)	215.9(0.8)	220.0(1.1)	214.2(1.0)	213.1(1.4)	212.8(2.0)	216.3(1.4)
OBSERVED ETHNICITY/RACE							
WHITE [c]	213.8(1.0)	216.6(0.7)	221.3(0.9)	218.3(0.8)	215.3(1.3)	214.9(1.9)	217.7(1.5)
BLACK	170.0(1.6)	181.3(1.1)	189.2(1.6)	185.7(1.2)	184.9(1.6)	185.0(2.3)	188.5(2.6)
HISPANIC	*****(0.0)	182.8(2.3)	189.5(3.3)	187.2(1.6)!	189.4(3.4)	189.8(3.7)	193.7(3.9)
OTHER	193.4(4.7)!	207.9(5.1)!	218.5(4.1)!	222.6(2.7)	204.0(6.8)!	203.7(6.9)!	228.4(5.0)
REGION							
NORTHEAST	213.0(1.7)	214.8(1.4)	220.9(2.5)	215.9(2.0)	212.8(2.9)	212.3(3.1)	215.2(2.8)
SOUTHEAST	194.3(2.8)	201.2(1.1)	210.2(2.3)	204.3(2.2)	202.4(2.8)!	202.5(3.1)!	207.2(2.3)
CENTRAL	214.4(1.4)	215.5(1.1)	216.5(1.2)	215.6(1.6)	213.3(2.7)	212.9(3.1)	218.2(2.5)
WEST	204.6(1.8)	207.1(2.0)	212.4(2.2)	209.1(2.0)	206.6(3.0)	206.5(3.3)	207.9(2.8)
PARENTAL EDUCATION							
NOT GRAD. H.S.	188.4(1.3)	190.0(1.2)	193.9(1.6)	195.1(1.5)	189.5(2.8)	189.5(3.2)	192.5(5.3)
GRADUATED H.S.	207.7(1.1)	211.3(0.9)	212.7(1.3)	208.9(1.2)	201.9(2.0)	202.2(2.4)	210.8(2.0)
POST H.S.	223.7(1.3)	221.6(0.9)	225.9(1.2)	222.9(1.1)	219.7(1.3)	219.0(2.0)	220.0(1.6)
DO NOT KNOW [d]	197.4(1.0)	203.2(0.8)	206.0(1.0)	204.4(1.0)	200.8(1.6)	200.9(2.2)	204.4(1.9)
MISSING	150 0(19.2)!	*****(0.0)	*****(0.0)	160.8(4.7)!	189.3(10.9)!	189.2(11.0)!	162.3(14.3)
GRADE							
< MODAL GRADE	177.5(1.2)	183.5(1.2)	188.8(1.3)	186.9(1.2)	188.8(1.4)	189.4(2.1)	192.6(1.8)
AT MODAL GRADE	216.8(1.1)	218.3(0.7)	225.0(0.9)	222.8(0.9)	219.1(1.2)	218.5(1.9)	222.8(1.5)
> MODAL GRADE	231.8(3.7)	225.8(3.6)	243.3(6.3)!	253.6(5.4)!	244.4(11.9)!	241.9(11.5)!	262.4(10.0)
MISSING	174.5(10.4)!	199.6(8.6)!	190.1(2.2)	*****(0.0)	*****(0.0)	*****(0.0)	*****(0.0)
ITEMS IN THE HOME							
0 - 2 ITEMS	186.2(1.0)	193.9(0.9)	197.7(1.4)	196.4(0.9)	194.3(1.5)	194.6(2.1)	198.5(2.1)
3 ITEMS	207.9(1.0)	212.2(0.7)	216.6(1.0)	216.6(0.9)	210.9(1.4)	210.6(2.0)	214.8(1.5)
4 ITEMS	222.8(0.9)	225.0(0.8)	227.9(1.0)	227.1(1.0)	221.2(1.2)	220.6(1.9)	223.0(1.7)
DO NOT KNOW	161.7(6.5)	177.8(14.8)!	167.0(10.2)!	155.7(6.7)!	169.0(17.6)!	171.2(18.0)!	181.4(25.0)
MISSING	180.1(13.1)!	183.6(5.0)!	204.9(3.2)!	158.5(3.5)!	243.1(25.2)!	235.3(24.4)!	160.2(13.7)
TELEVISION WATCHED PER DAY [e]							
0 - 2 HOURS	*****(0.0)	*****(0.0)	219.9(1.1)	219.3(1.3)	212.3(1.9)	212.1(2.4)	217.0(1.7)
3 - 5 HOURS	*****(0.0)	*****(0.0)	222.3(0.7)	218.3(0.9)	217.0(1.2)	216.4(1.9)	218.2(1.6)
6 HOURS OR MORE	*****(0.0)	*****(0.0)	211.0(0.8)	198.9(1.0)	195.0(1.6)	195.2(2.1)	198.1(1.6)
MISSING	*****(0.0)	*****(0.0)	153.8(2.4)	192.3(2.1)!	211.5(14.1)!	208.1(14.3)!	194.0(7.8)

[a] Using adjustment data and adjusted standard errors. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

[b] Based on the 1988 bridge to 1984

[c] Includes Hispanic students in 1970-71

[d] Includes "MISSING" in 1970-71, 1974-75, and 1979-80

[e] Unavailable in 1970-71 and 1974-75

! Interpret with caution--the sampling error cannot be accurately estimated, since the coefficient of variation of the estimated total number of students in the subpopulation exceeds 20 percent.

Table 6.3

Reading Trend Results Including 1986 Adjusted Values: Age 13
(standard errors in parentheses)

					Unadjusted	Adjusted	
	1971	1975	1980	1984	1986	1986 [a]	1988 [b]
-- TOTAL --	255.2(0.9)	256.0(0.8)	258.5(0.9)	257.1(0.7)	259.4(1.0)	255.0(1.6)	257.5(0.9)
SEX							
MALE	248.5(1.0)	249.6(0.8)	254.3(1.1)	252.7(0.8)	255.1(1.0)	250.9(1.6)	251.8(1.2)
FEMALE	260.9(0.9)	262.4(0.9)	262.7(0.9)	261.7(0.8)	263.6(1.4)	259.1(1.9)	263.0(1.0)
OBSERVED ETHNICITY/RACE							
WHITE [c]	260.9(0.8)	262.1(0.7)	264.4(0.6)	262.6(0.5)	263.4(1.3)	258.8(1.8)	261.3(1.0)
BLACK	222.4(1.1)	225.7(1.2)	232.4(1.5)	236.0(1.2)	242.9(1.6)	239.3(2.1)	242.9(2.3)
HISPANIC	*****(0.0)	232.5(3.4)	236.8(2.1)	239.6(1.6)!	246.3(3.2)	242.2(3.4)	240.1(3.5)
OTHER	251.4(3.5)!	255.4(4.9)!	252.8(4.8)!	260.1(2.9)	268.8(4.3)	263.9(4.4)	269.3(4.3)
REGION							
NORTHEAST	261.2(2.0)	258.8(1.8)	260.1(1.8)	260.4(0.7)	263.6(2.2)	258.7(2.5)	258.6(2.0)
SOUTHEAST	245.0(1.7)	249.3(1.5)	252.7(1.7)	256.4(1.8)	259.0(1.6)!	254.8(2.0)!	257.6(1.9)
CENTRAL	260.0(1.9)	261.6(1.4)	264.6(1.5)	258.7(1.2)	254.5(3.6)	250.8(3.9)	255.9(2.0)
WEST	253.5(1.2)	253.1(1.6)	256.3(2.1)	253.9(1.4)	260.8(1.8)	256.0(2.2)	257.9(2.1)
PARENTAL EDUCATION							
NOT GRAD. H.S.	238.5(1.1)	238.6(1.2)	238.5(1.3)	240.1(1.2)	248.9(2.9)	244.2(3.2)	246.5(2.2)
GRADUATED H.S.	255.5(0.8)	254.6(0.7)	253.6(0.8)	253.2(0.8)	253.5(1.1)	249.3(1.7)	252.7(1.2)
POST H.S.	270.2(0.8)	269.9(0.8)	270.9(0.8)	267.7(0.7)	267.3(0.9)	262.7(1.5)	265.3(1.4)
DO NOT KNOW [d]	233.1(1.1)	234.9(1.0)	233.3(1.7)	236.5(1.4)	239.6(3.2)	236.4(3.4)	240.4(2.8)
MISSING	*****(0.0)	*****(0.0)	*****(0.0)	255.0(4.4)!	264.5(15.8)!	259.6(15.8)!	224.7(19.2)
GRADE							
< MODAL GRADE	229.5(1.0)	232.3(1.0)	239.6(1.5)	239.1(0.9)	242.2(1.5)	238.4(1.9)	242.8(1.3)
AT MODAL GRADE	264.8(0.8)	264.9(0.7)	266.1(0.9)	266.7(0.6)	267.7(1.0)	263.0(1.5)	266.7(1.1)
> MODAL GRADE	278.1(2.6)	278.1(4.0)	274.5(4.9)!	295.3(8.6)!	280.4(6.3)!	275.8(6.3)!	271.8(11.4)
MISSING	225.2(9.8)!	204.9(15.8)!	249.7(10.7)	*****(0.0)	*****(0.0)	*****(0.0)	*****(0.0)
ITEMS IN THE HOME							
0 - 2 ITEMS	226.6(1.2)	231.5(1.2)	235.8(1.4)	238.4(1.0)	242.7(1.6)	238.4(2.0)	242.9(1.8)
3 ITEMS	248.9(0.9)	249.7(0.8)	253.1(1.1)	254.3(0.7)	256.6(1.3)	252.9(1.7)	255.6(1.0)
4 ITEMS	266.5(0.7)	267.4(0.7)	268.5(0.7)	266.1(0.7)	266.0(1.0)	263.1(1.6)	264.2(1.3)
DO NOT KNOW	218.6(7.3)!	218.5(14.5)!	222.9(6.1)!	204.7(10.5)!	278.5(81.3)!	275.0(81.0)!	212.7(17.1)
MISSING	224.4(9.9)!	227.3(14.2)!	247.3(6.7)!	248.8(5.5)!	264.1(18.1)!	259.0(18.2)!	223.4(21.6)
TELEVISION WATCHED PER DAY [e]							
0 - 2 HOURS	*****(0.0)	*****(0.0)	263.3(0.9)	268.1(0.8)	265.1(1.9)	260.3(2.2)	264.3(1.4)
3 - 5 HOURS	*****(0.0)	*****(0.0)	257.1(0.9)	261.6(0.6)	262.0(1.0)	257.5(1.6)	258.7(1.0)
6 HOURS OR MORE	*****(0.0)	*****(0.0)	243.2(1.3)	244.2(0.9)	245.5(1.6)	241.9(2.0)	243.5(2.0)
MISSING	*****(0.0)	*****(0.0)	233.2(4.1)	238.0(1.3)	227.6(17.6)!	220.9(18.6)!	227.9(3.0)

[a] Using adjustment data and adjusted standard errors. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

[b] Based on the 1988 bridge to 1984

[c] Includes Hispanic students in 1970-71

[d] Includes "MISSING" in 1970-71, 1974-75, and 1979-80

[e] Unavailable in 1970-71 and 1974-75

! Interpret with caution--the sampling error cannot be accurately estimated, since the coefficient of variation of the estimated total number of students in the subpopulation exceeds 20 percent.

Table 6.4

Reading Trend Results Including 1986 Adjusted Values: Age 17 (standard errors in parentheses)

	1971	1975	1980	1984	Unadjusted 1986	Adjusted 1986 [b]	1988 [c]
-- TOTAL --	285.4(1.2)	286.1(0.8)	285.8(1.4)	288.8(0.9)	277.4(1.1)[a]	286.0(1.7)	290.1(1.1)
SEX							
MALE	279.0(1.2)	280.1(0.9)	282.1(1.4)	283.8(0.9)	269.8(1.4)	279.6(1.9)	286.0(1.5)
FEMALE	291.5(1.3)	291.8(0.9)	289.5(1.4)	293.9(1.1)	285.2(1.2)	292.6(1.8)	293.8(1.6)
OBSERVED ETHNICITY/RACE							
WHITE [d]	291.4(1.0)	293.0(0.6)	293.1(1.2)	295.6(0.7)	283.9(1.2)	291.4(1.8)	294.7(1.3)
BLACK	238.6(1.7)	240.4(1.9)	242.5(2.0)	264.2(1.2)	251.8(1.5)	265.0(2.0)	274.4(2.6)
HISPANIC	*****(0.0)	252.2(3.6)	260.7(3.3)	268.1(1.9)!	255.3(2.6)	267.5(2.6)	270.8(4.0)
OTHER	276.3(7.1)!	275.3(4.3)!	280.7(4.0)	284.5(3.1)	266.4(5.5)	276.0(4.7)	290.0(5.7)
REGION							
NORTHEAST	292.2(2.5)	289.5(1.7)	285.4(2.4)	292.0(2.1)	286.0(2.5)	293.1(2.5)	294.8(2.5)
SOUTHEAST	270.8(2.5)	277.3(1.4)	281.0(2.6)	284.6(2.3)	269.4(1.3)	279.4(1.8)	285.5(2.1)
CENTRAL	290.8(2.1)	291.9(1.5)	288.6(3.2)	290.1(1.5)	279.7(2.0)	288.1(2.5)	291.2(1.8)
WEST	283.7(1.7)	282.3(1.8)	286.6(1.7)	289.1(1.6)	273.5(1.9)	282.7(2.2)	289.0(2.2)
PARENTAL EDUCATION							
NOT GRAD. H.S.	261.6(1.5)	263.3(1.4)	261.9(1.7)	269.3(1.4)	253.4(1.7)	266.3(2.1)	267.4(2.4)
GRADUATED H.S.	283.3(1.2)	281.7(1.0)	277.4(1.1)	281.1(1.0)	264.9(1.0)	275.9(1.7)	282.0(1.5)
POST H.S.	302.3(1.0)	300.9(0.7)	299.3(1.2)	301.2(0.8)	289.3(1.1)	295.8(1.7)	299.5(1.3)
DO NOT KNOW [e]	261.8(6.5)	240.2(2.8)	249.5(3.9)	256.5(2.1)	237.9(2.7)	253.4(2.7)	254.7(6.1)
MISSING	*****(0.0)	*****(0.0)	*****(0.0)	280.8(7.9)!	249.3(7.7)!	261.8(6.4)!	230.5(27.8)
GRADE							
< MODAL GRADE	238.6(1.5)	242.8(1.8)	243.8(2.3)	258.8(1.2)	242.8(1.3)	257.7(1.9)	265.4(2.2)
AT MODAL GRADE	291.3(1.0)	292.5(0.7)	291.5(1.2)	295 7(0.7)	286.0(1.1)	293.1(1.7)	296.5(1.1)
> MODAL GRADE	302.9(1.6)	301.8(1.0)	301.2(2.2)	303.8(1.3)	296.0(2.7)	301.0(2.6)	304.6(2.6)
MISSING	257.6(17.6)!	259.5(13.2)!	241.3(22.3)!	*****(0.0)	*****(0.0)	*****(0.0)	*****(0.0)
ITEMS IN THE HOME							
0 - 2 ITEMS	246.2(1.8)	251.7(2.1)	257.6(2.2)	264.1(1.4)	250.7(1.6)	264.1(2.0)	268.8(2.4)
3 ITEMS	273.9(1.4)	275.8(1.1)	278.5(1.8)	283.0(1.1)	270.9(1.4)	280.6(1.9)	287.1(1.7)
4 ITEMS	295.6(1.0)	296.1(0.6)	295.6(1.1)	296.3(0.8)	286.5(1.0)	293.5(1.7)	295.8(1.2)
DO NOT KNOW	238.3(18.0)!	205.3(21.6)!	*****(0.0)	220.6(8.3)!	222.9(14.1)!	242.1(11.1)!	227.4(14.0)
MISSING	284.6(57.6)!	239.4(11.7)	244.2(5.8)!	221.4(10.4)!	240.7(9.9)!	254.8(8,1)!	199.0(17.9)
TELEVISION WATCHED PER DAY [f]							
0 - 2 HOURS	*****(0.0)	*****(0.0)	291.0(1.3)	297.4(0.9)	287.5(1.3)	294.3(1.8)	295.6(1.2)
3 - 5 HOURS	*****(0.0)	*****(0.0)	277.1(1.3)	284.5(0.9)	273.5(1.0)	282.8(1.7)	285.4(1.8)
6 HOURS OR MORE	*****(0.0)	*****(0.0)	257.7(2.9)	267.8(1.4)	249.1(1.6)	263.0(2.0)	268.6(4.1)
MISSING	*****(0.0)	*****(0.0)	256.1(7.2)	245.9(2.2)	241.2(9.8)!	255.6(8.0)!	232.9(10.3)

[a] Standard error differs from column 4 of Table 4.1 because of a change in jackknife methodology

[b] Using adjustment data and adjusted standard errors. Appendix A (p. 171) gives a summary of which modifications and adjustments of reading scale results are used in the tables and figures in this report.

[c] Based on the 1988 bridge to 1984

[d] Includes Hispanic students in 1970-71

[e] Includes "MISSING" in 1970-71, 1974-75, and 1979-80

[f] Unavailable in 1970-71 and 1974-75

! Interpret with caution--the sampling error cannot be accurately estimated, since the coefficient of variation of the estimated total number of students in the subpopulation exceeds 20 percent.

procedure, adjusted 1986 results were produced. At age 17, the adjustment resulted in an increase of 8.6 points in the mean and a decrease of 9.9 points in the standard deviation. Some changes occurred in the patterns of subgroup differences. At age 13, the mean decreased by 4.4 points and, at age 9, the results were essentially unchanged. The adjusted 1986 results are 2 to 3 points lower than the 1984 results at all three ages.

Our findings on the impact of item position and context lead us to the conclusion that common-item equating procedures should not be assumed to be appropriate when form changes have taken place. This reinforces the importance of using *intact blocks of items* for purposes of scale equating in NAEP. In all cases in which the estimation of trend is planned, our designs for 1990 and 1992 include intact blocks that retain the same position within the assessment booklet. This more conservative approach should greatly diminish the likelihood of anomalous results such as those that occurred in 1986.

Chapter 7

MATHEMATICS AND SCIENCE TREND DATA ANALYSIS

Kentaro Yamamoto[1]

Because mathematics and science items were part of the bridge samples used in 1988 to illuminate the anomalous 1986 reading results, data for these subject areas were also available for trend analyses. This chapter describes the technical details of the item-parameter estimation ard scaling performed for trend analyses of responses to mathematics and science cognitive items in the 1988 assessment.

To maintain the comparability of measurement instruments, booklets for the 1988 reading bridge to 1986 were identical to those used in 1986 and therefore included science and mathematics blocks. The 1988 mathematics and science trend analyses are limited to data from blocks that appeared in the same booklets as reading blocks in the 1986 assessment. For age 17, the number of mathematics and science blocks available for trend analysis was fewer in 1988 than in 1986. However, since every 1986 trend booklet for ages 9 and 13 contained a block from each of the three subject areas, the complete sets of trend blocks for those ages were available for analysis in 1988.

[1] Maxine Kingston, Edward Kulick, Michael Narcowich, and Minhwei Wang performed the data analyses for this chapter; Jo-Ling Liang and Edward Kulick produced the figures. Robert Mislevy provided consultation on scaling and Rebecca Zwick provided valuable editorial assistance.

The combination of blocks within booklets, the composition of item blocks, the mode of administration, the sample definition, and the time of testing were identical for the age 9 and age 13 samples in the 1986 assessment and the 1988 bridge to 1986. Consequently, trend analyses for these two ages were straightforward, but analyses of trends for age 17 were not.

In 1986, the reading trend for age 17 was assessed as part of the BIB spiral portion of the assessment, while the science and mathematics trends were assessed apart from reading under a paced-tape mode of administration. Since the overarching aim of the 1988 bridge study was to replicate the booklets and administration procedures for the 1986 assessment of trends in reading, booklets from the BIB spiral portion of the 1986 assessment were again administered in 1988 under the same administration conditions as in 1986. In particular, the administration of mathematics and science items in the spiral portion was by paper and pencil, rather than by paced tape. This means that the data from the 1988 age 17 trend assessments of mathematics and science are comparable to the 1986 BIB assessment and not directly to the 1986 trend assessment. This made the equating design to align the 1988 trend point for age 17 student to the past trend more complicated than before. For age 17 in 1988, two types of equating were necessary--one based on common populations across different modes of administration for the 1986 BIB and trend, and one based on common items (similarly placed) for the 1986 BIB and the 1988 trend.

The main objective of the 1988 trend assessments of mathematics and science was to evaluate the differences between the 1986 and 1988 assessments. The 1988 trend point was to be added to the existing trend line. Since these analyses closely follow those conducted in 1986, readers desiring more detailed descriptions are referred to various chapters in the 1986 technical report (Beaton, 1988b), such as Chapter 8 by Mislevy for the underlying theory of measurement and the imputation of plausible values. Chapter 3 by Hansen, Rust, and Burke for sampling design, Chapter 14 by Johnson, Burke, Braden, Hansen, Lago, and Tepping for sample weights, Chapter 10 by Johnson for mathematics data analysis, and Chapter 11 by Yamamoto for science data analysis. This chapter will consider details specific to the 1988 analysis.

7.1 Sampling of Students and Items for Mathematics and Science

For ages 9 and 13, the combination of blocks, composition of item blocks, mode of administration, sample definitions, and time of testing were identical in 1986 and the 1988 bridge to 1986. Three booklets, identical to those used in 1986, were used to measure trend for these ages. Each booklet contained one reading, one mathematics, and one science block. Each student in the sample was administered one of these booklets. The mathematics and science portions were presented aurally using a tape recorder as in past assessments. The tape recorder was turned off for the reading block.

For age 17, the mathematics and science booklets of the 1986 trend assessment were not used in 1988, since the 1986 mathematics and science trend booklets for age 17 did not include reading tasks. Instead, the booklets used in 1988 were identical to a subset of booklets used for the 1986 BIB assessment and consisted of six booklets, five of which contained at least one reading block and either a mathematics or a science trend block from the 1986 assessment. The sixth booklet, which did not contain mathematics or science blocks, was included for the reading assessment in 1988. Only one of the two trend blocks of either mathematics or science was included in four of the booklets; the fifth booklet contained both a mathematics and a science block.

The 1988 age 17 sample was defined using the same age definition as the 1986 BIB assessment and received a print-administered assessment instead of the paced administration of the pre-1988 trend assessments. Unlike the samples at ages 9 and 13, in which every student received both a mathematics and a science block, about one-fifth of the age 17 sample received both; the rest received a block of either mathematics or science items.

The proficiencies of the three ages cannot be placed on a single scale without a cross-sectional study or a vertical equating across ages, neither of which were possible in the 1988 mathematics and science trend assessment. The mathematics and science scales were derived from the 1986 cross-sectional assessment (see E. G. Johnson, 1988d, and Yamamoto, 1988). The 1988 trend analysis added a new trend point to the existing trend line up to 1986.

The specific mathematics and science samples for 1988 and 1986 follow.

Sample (yr:age)	Type	Time	Mode	Age Definition	Sample	Size	Modal Grade
Mathematics and Science							
86:9a	Bridge	Winter	Tape	Calendar yr.	Age	6932	4
88:9c	Bridge	Winter	Tape	Calendar yr.	Age	3711	4
86:13a	Bridge	Fall	Tape	Calendar yr.	Age	6200	8
88:13c	Bridge	Fall	Tape	Calendar yr.	Age	3942	8
Mathematics							
86:17	Main	Spring	Print	Not calendar yr.	Age/grade	6151*	11
86:17b	Bridge	Spring	Tape	Not calendar yr.	Age	3868	11
88:17c	Bridge	Spring	Print	Not calendar yr.	Age/grade	1852	11
Science							
86:17	Main	Spring	Print	Not calendar yr.	Age/grade	5611*	11
86:17b	Bridge	Spring	Tape	Not calendar yr.	Age	3868	11
88:17c	Bridge	Spring	Print	Not calendar yr.	Age/grade	1862	11

* Number of age-only BIB sample students who answered any one of the trend blocks.

Note: 1) For all three ages, mathematics 1988 trend blocks are identical to those administered in 1986; 2) Only the subset of the 86:17 and 88:17c samples that were age-eligible and received trend blocks were used, and numbers on the table reflect such samples.

The items used for the analysis of the 1988 data set are the same as those used for the 1986 trend analyses; that is, the same items were excluded as in 1986 for reasons of lack of fit of the estimated item response function to the empirical regression curve. Three mathematics items, one from each age group, were excluded from scaling. Three science items were dropped from the scaling for age 9 and three from the scaling for age 17; one science item was dropped for age 13.

Using current methods, it is possible to assess the change over time in either item characteristics or proficiencies of populations, but not both at the same time. This is true for any analysis, whether based on classical test theory, item response theory, or proportions correct. To assess change in item characteristics, we are forced to assume that the ability distribution of the population remains stable; to assess change in the ability distribution of the population, we must assume the stability of item characteristics (see the discussion of common-item equating in Chapter 6). However, we know that this is not strictly true. Societal and instructional changes may produce gradual alterations in item functioning over time. If there is evidence that this is occurring, it may be desirable to allow for changes in the parameters of these common items. Permitting item characteristics to vary in this way is feasible only if common-population equating methods are available to link the newly obtained results to past trend lines. This is the approach that was used in analyzing the 1988 mathematics data at age 17 and science data at all three ages.

7.2 Scaling of the Mathematics Trend Data

Ages 9 and 13

From the item analysis, it was found that the 1988 response distributions of all response choices, including "omits," were quite similar to the 1986 data. The mean weighted proportion correct at the block level was computed; these values were compared with the 1986 results, as shown in Table 7.1. At each block level for all age groups, the 1988 sample showed higher weighted proportion correct values than the 1986 sample.

In estimating item parameters in 1986, combined data from the three most recent trend assessments (1977, 1982, and 1986) were used. Thus, the 1986 trend analysis assumed the characteristics of all items were stable across the three assessments. Item parameters estimated in 1986 were kept unchanged for the 1988 assessment for ages 9 and 13. Consequently, the same constants were used to transform provisional imputed values to the mathematics proficiency scale.

To justify the use of the parameter estimates from the 1986 assessment, the fit of the IRT item parameters to the 1988 bridge data was examined by means of the chi-square test. At ages 9 and 13, the use of previously estimated item parameters appeared to be justified, but this was not the case at age 17. Hence, the item parameters applicable to age 9 and age 13 were kept unchanged for the mathematics trend analysis; they are presented in Tables D.8 and D.9 in Appendix D.

The coexistence of item parameters that fit in various degrees to the data from a particular year comes from the need to place several samples from different years on a scale based upon common-item equating. When common-item parameters are estimated on multiple data sets, the fit of the estimated item

Table 7.1

Mathematics Weighted Mean Proportion Correct

	Block	1986	(N)	1988	(N)	Item[c]	
Age 17	1	59.1	(2211)[a]	61.3	(619)	35	
(paper)	2	63.4	(2233)[a]	65.7	(624)	35	
	3	65.3	(2263)[a]	67.0	(609)	24	(19)
	Total	52.3	(6151)[a]	64.4	(1852)	94	
	Noncalculator	61.0		62.7		75	
Age 17	1	60.3	(1934)[b]			35	
(taped)	2	62.1	(1934)[b]			35	
	3	64.5	(1934)[b]			24	(19)
	Total	62.0	(3868)[b]			94	
	Noncalculator	60.8				75	
Age 13	1	63.9	(2075}	65.3	(1405)	37	
(taped)	2	58.5	(2054)	60.5	(1281)	37	
	3	57.4	(2071)	60.0	(1256)	24	(16)
	Total	60.3	(6200)	62.2	(3942)	98	
	Noncalculator	61.4		63.2		82	
Age 9	1	55.2	(2315)	58.2	(1274)	26	
(taped)	2	57.3	(2361)	62.4	(1240)	26	
	3	73.0	(2256)	76.7	(1197)	16	(11)
	Total	60.2	(6932)	64.2	(3711)	68	
	Noncalculator	57.1		62.1		57	

[a] Age-only BIB sample with at least one mathematics trend block.

[b] 1986 Age 17 trend sample blocks 1 and 2 were paired.

[c] Includes some items that were excluded from IRT scaling; parentheses in this column indicate the number of calculator items excluded from IRT scaling.

regression curve to the weighted means of proportions correct, given an ability level, is maximized. Because of this averaging, it is possible that the estimated item parameters fit very well to the combined data sets as a whole, but less well to each data set separately.

For ages 9 and 13, the same common-item equating procedure that was employed in the 1986 trend analysis was used to align the 1988 point to the trend up to 1986. A brief description of the procedure follows. From the item parameters estimated in 1986 and background variables of 1988, the proficiency scores were imputed for the 1988 bridge data for each age using the M-GROUP computer program based on the plausible values methodology (Sheehan, 1985; see Mislevy, 1988, for a detailed discussion). The conditioning variables and the estimated conditioning effects for ages 9 and 13 are given in Tables D.10 and D.11 in Appendix D. The same linear constants of 1986 were used to transform provisional imputed scores to the final proficiency scores of the mathematics trend. The transformation constants for all three ages are listed in Table 7.2.

Table 7.2

Coefficients of the Linear Transformation of the Trend Scale from Original Units to the Mathematics Proficiency Scale

Age	Intercept	Slope
9	218.42	35.84
13	266.58	34.57
17	303.25	31.84

Age 17

For age 17, new item parameters were estimated using the subsample from the 1986 BIB assessment equivalent to the 1988 trend sample. Use of the estimated item parameters from 1986 is not appropriate for the 1988 assessment for age 17, because of the different mode of administration for the 1986 and the 1988 trend assessments for that age. For example, on all five items of a type referred to as "estimate" items, use of paper and pencil instead of a tape recorder had a dramatic effect. "Estimate" items ask the student to select an answer among several options, all of which are rounded so that none of them is exactly correct. The property of the response options is indicated by the word "about" being positioned before "how much" or "how many" in a question. When an "estimate" item was presented under taped administration, enough time was allowed for rough estimation of the (typically) large number, but not enough time was allowed for the numerical calculation of the answer. However, because under paper-and-pencil administration it is possible to spend more time to answer, the examinee may opt to perform the calculation rather than the estimation. In such a case, it is more appropriate to treat an "estimate" item as two different items under different modes of administration. The observed item regression curves of the 1986 BIB data and 1986 bridge data of one of the "estimate" items are presented in Figure 7.1.

Therefore, for age 17, both equating methods, common-item (between the 1986 BIB and 1988 bridge samples) and common-population (between the 1986 BIB and 1986 bridge samples), were used to place the 1988 trend sample on a scale comparable to the 1986 reported scale. The procedure took place as follows. The item parameters for the total set of 73 items were estimated based on the two data sets: the 1986 BIB assessment and the 1988 bridge to 1986. Both samples included grade- and age-eligible students in order to maintain an adequate sample size for the estimation accuracy. This resulted in a second set of item parameters for age 17. The new item parameters are listed in

Figure 7.1

A Plot of Observed Proportion Correct of the 1986 BIB Spiral and Trend Assessments with the Estimated Item Regression Curve for an "Estimate" Item

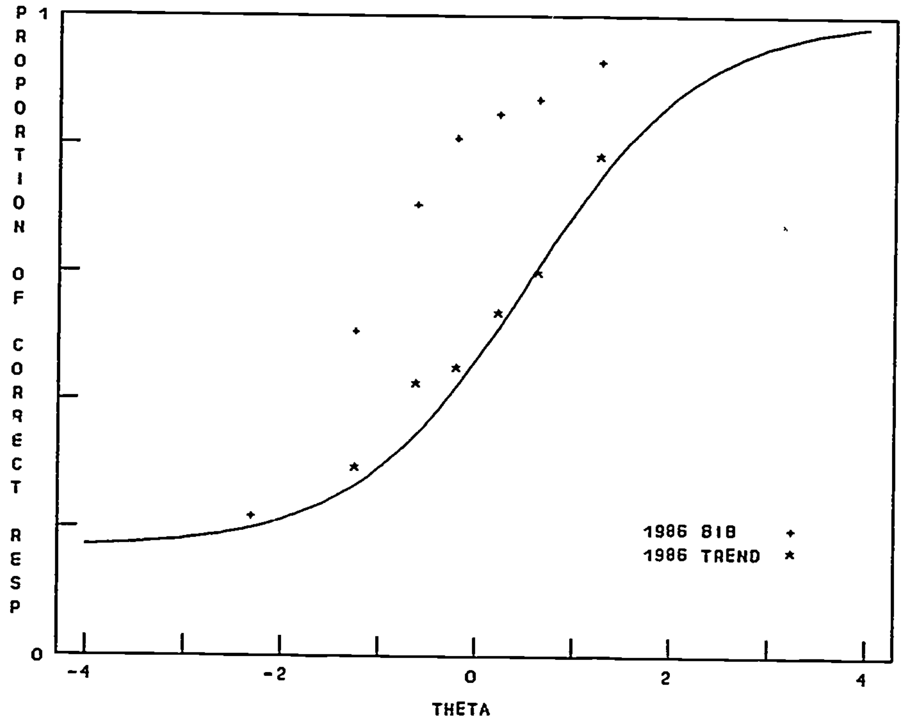


Table D.12 in Appendix D; the old parameters appear in Beaton (1988b). The rationale for estimating parameters for all items instead of only "estimate" items comes from the main objective of the 1988 bridge to 1986, namely to examine the possibility of effects due to changes in assessment procedures.

From the above estimated item parameters and background information for the appropriate sample, proficiency scores were imputed for each student in the 1986 BIB and 1988 bridge-to-1986 samples. The conditioning variables and the estimated conditioning effects for age 17 are given in Table D.13 in Appendix D. Then the mean and standard deviation of the imputed scores of the age-only subsample of the 1986 BIB were calculated. Constants were found to match the means and standard deviations of the proficiency scores of the 1986 trend sample and the age-only subsample of the 1986 BIB sample. Subsequently, by applying the same linear transformation to the provisional imputed values of the 1988 trend age-only sample, the 1988 trend point was aligned with the trend line up to 1986. The transformation constants for age 17 data are listed in Table 7.2.

The trends in mean proficiency with jackknifed standard errors for subpopulations of the three age samples are listed in Tables 7.3, 7.4, and 7.5. The 1986 and 1988 posterior distributions of mathematics proficiency were calculated for each cohort separately at 40 quadrature points. Overlays of distributions from the two assessment years appear in Figures 7.2 through 7.4. For age 17, the 1986 distribution is calculated on the 1986 bridge sample as well as on the age-only subsample of the 1986 BIB sample, which is comparable to the 1988 bridge sample of age 17. The shape of the distributions of the two assessments is quite similar for ages 9 and 13.

Table 7.3

Weighted Mathematics Proficiency Means and Standard Errors for Age 9

	1978		1982		1986		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	218.6	(0.8)*	219.0	(1.1)*	221.7	(1.0)*	229.0	(1.1)
Sex								
Male	217.4	(0.7)*	217.1	(1.2)*	221.7	(1.1)*	229.1	(1.6)
Female	219.9	(1.0)*	220.8	(1.2)*	221.7	(1.2)*	229.0	(1.1)
Ethnicity								
White	224.1	(0.9)*	224.0	(1.1)*	226.9	(1.1)*	234.5	(1.2)
Black	192.4	(1.1)*	194.9	(1.6)*	201.6	(1.6)	206.3	(2.6)
Hispanic	202.9	(2.3)*	204.0	(1.3)*	205.4	(2.1)*	215.9	(3.4)
Other	227.2	(3.2)*	238.5	(4.2)	221.8	(7.5)*	242.9	(4.2)
Grade								
< modal	190.9	(1.1)*	193.1	(1.4)*	198.1	(1.0)*	208.8	(1.8)
= modal	228.5	(0.9)*	230.1	(1.0)*	233.8	(1.0)*	239.0	(1.2)
> modal	240.5	(5.7)	258.3	(11.0)	248.8	(10.8)	260.1	(9.7)
Region								
Northeast	226.9	(1.9)	225.7	(1.7)	226.0	(2.7)	233.5	(3.1)
Southeast	208.9	(1.2)*	210.4	(2.9)*	217.8	(2.5)	222.4	(2.9)
Central	224.0	(1.5)*	221.1	(2.4)*	226.0	(2.3)*	233.9	(1.7)
West	213.5	(1.4)*	219.3	(1.7)*	217.2	(2.4)*	226.9	(2.0)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of three comparisons (each year compared to 1988).

Table 7.4

Weighted Mathematics Proficiency Means
and Standard Errors for Age 13

	1978		1982		1986		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	264.1	(1.1)*	268.6	(1.1)*	269.0	(1.2)*	273.3	(0.8)
Sex								
Male	263.6	(1.3)*	269.2	(1.4)*	270.0	(1.1)*	275.3	(1.1)
Female	264.7	(1.1)*	268.0	(1.1)	267.9	(1.5)	271.2	(1.0)
Ethnicity								
White	271.6	(0.9)*	274.4	(1.0)*	273.6	(1.3)*	279.1	(0.9)
Black	229.6	(1.9)*	240.4	(1.6)*	249.2	(2.3)	250.3	(1.2)
Hispanic	238.0	(2.2)*	252.4	(1.6)	254.3	(2.9)	254.7	(3.9)
Other	272.5	(3.5)	274.5	(3.8)	282.7	(3.4)	288.6	(5.9)
Grade								
< modal	239.6	(1.4)*	247.2	(1.4)*	251.1	(1.1)*	255.8	(1.1)
= modal	273.8	(1.1)*	276.6	(0.9)*	277.6	(1.0)*	282.5	(0.8)
> modal	297.6	(7.7)	303.9	(7.6)	296.9	(7.7)	320.5	(7.2)
Region								
Northeast	272.7	(2.4)	276.9	(2.2)	276.6	(2.2)	278.7	(2.2)
Southeast	252.7	(3.2)*	258.1	(2.4)*	263.5	(1.4)	268.2	(2.9)
Central	269.4	(1.8)	272.8	(1.9)	266.1	(4.5)	271.3	(1.5)
West	260.0	(1.9)*	266.0	(2.3)*	270.4	(2.1)	274.6	(1.7)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of three comparisons (each year compared to 1988).

Table 7.5

Weighted Mathematics Proficiency Means
and Standard Errors for Age 17

	1978		1982		1986		1986(BIB)		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	300.4	(0.9)*	298.5	(0.9)*	302.0	(0.9)	302.0	(0.8)	305.4	(1.2)
Sex										
Male	303.8	(1.0)	301.5	(1.0)*	304.7	(1.2)	303.2	(1.0)	306.7	(1.8)
Female	297.1	(1.0)*	295.6	(1.0)*	299.4	(1.0)*	300.8	(0.8)*	304.2	(1.4)
Ethnicity										
White	305.9	(0.9)	303.7	(0.9)*	307.5	(1.0)	306.4	(0.9)	309.5	(1.4)
Black	268.4	(1.3)*	271.8	(1.3)*	278.6	(2.1)*	280.0	(1.0)*	289.2	(2.1)
Hispanic	276.3	(2.2)*	276.7	(2.0)*	283.1	(2.9)*	286.0	(1.8)*	294.3	(3.5)
Other	312.9	(3.4)	309.4	(8.8)	304.7	(7.2)	314.6	(6.0)	314.2	(7.0)
Grade										
< modal	272.7	(1.1)*	274.1	(1.5)*	277.3	(1.6)*	278.9	(1.4)*	283.4	(1.8)
= modal	304.7	(1.0)*	302.5	(0.9)*	306.7	(0.9)*	307.7	(0.7)*	312.5	(1.1)
> modal	309.3	(1.0)	306.5	(1.4)	309.1	(3.0)	312.3	(2.0)	314.2	(3.0)
Region										
Northeast	306.7	(1.7)	304.0	(2.1)	307.4	(1.9)	308.4	(1.4)	309.5	(2.6)
Southeast	292.3	(1.7)*	292.3	(2.1)*	297.3	(1.4)	295.0	(1.2)	300.2	(2.3)
Central	305.2	(1.8)	302.0	(1.1)	303.6	(1.9)	303.7	(1.6)	305.5	(2.8)
West	295.5	(1.8)*	294.1	(2.0)*	299.3	(2.7)	299.9	(1.7)	306.4	(2.2)

* shows statistically significant difference from 1988, where $\alpha = .05$ per set of four comparisons (each year compared to 1988).

Figure 7.2

Estimated Mathematics Proficiency Distributions
for the 1986 and 1988 Bridge Samples, Age 9

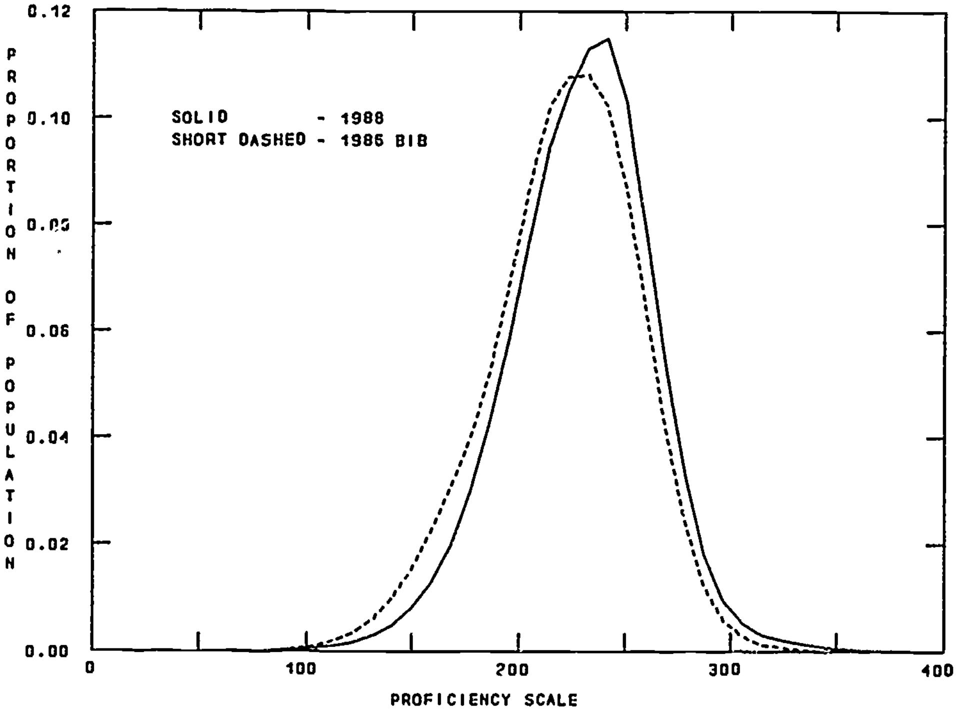


Figure 7.3

Estimated Mathematics Proficiency Distributions
for the 1986 and 1988 Bridge Samples, Age 13

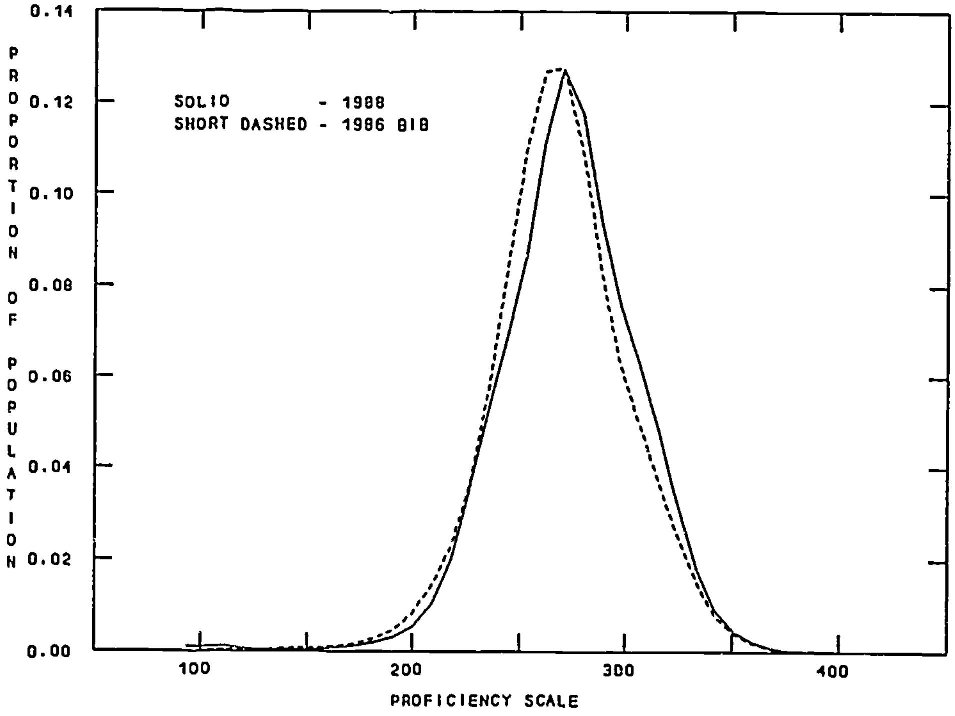
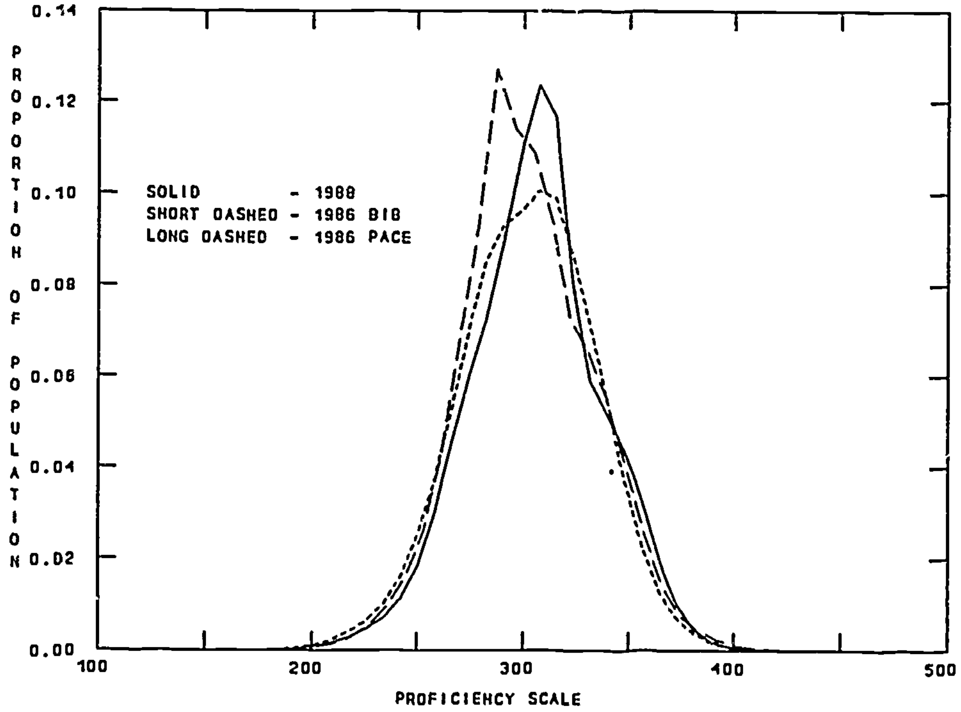


Figure 7.4

Estimated Mathematics Proficiency Distributions
for the 1986 Age-only BIB Sample and the 1988 Age-only Bridge Sample
Age 17



In 1986, using the range of student performance on the NAEP mathematics scale, five levels of mathematics proficiency were established and described in detail in the *Mathematics Report Card* (Dossey, Mullis, Lindquist, & Chambers, 1988): Level 150--Simple Arithmetic Facts, Level 200--Beginning Skills and Understanding, Level 250--Basic Operations and Beginning Problem Solving, Level 300--Moderately Complex Procedures and Reasoning, and Level 350--Multi-step Problem Solving and Algebra. Table 7.6 shows the percentage of students at ages 9, 13, and 17 who attained each level of proficiency in the 1978, 1982, 1986, and 1988 assessments.

7.3 Scaling of the Science Trend Data

The 1988 science trend analysis followed procedures and methods similar to those for the mathematics analysis. From the item analysis, it was found that the 1988 response distributions of all response choices, including "omits," were quite similar to the 1986 data. The mean weighted proportion correct at a block level was computed; these values were compared with the 1986 results, and are presented in Table 7.7.

In 1986, item parameters were estimated for the age 9, 13, and 17 samples. The trend items for age 13 and age 17 were estimated together because the majority of the items were common to both ages. For the 1988 data, because of the change in the mode of administration for age 17, those ms had to be estimated separately from the age 13 items. To obtain the best estimates of proficiencies for the two years, items for age 13 were reestimated using BILOG (Mislevy & Bock, 1983) on the 1986 and 1988 bridge data sets. For age 9, it was found that the 1986 score key for one of 63 items did not distinguish "I don't know," hence the responses to that item

Table 7.6

Mathematics Trends for 9-, 13-, and 17-Year-Old Students:
Percentage of Students at or Above
the Five Proficiency Levels, 1978-1988

		Assessment Year							
		1978		1982		1986		1988	
Proficiency Levels	Age	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Level 150	9	96.5	(0.2)	97.2	(0.3)	97.8	(0.2)	99.0	(0.2)
Simple Arithmetic	13	99.8	(0.0)	99.9	(0.0)	100.0	(0.0)	100.0	(0.0)
Facts	17	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)	100.0	(0.0)
Level 200	9	70.3	(0.9)*	71.5	(1.1)*	73.9	(1.1)*	81.5	(1.1)
Beginning Skills	13	94.5	(0.4)	97.8	(0.4)	98.5	(0.2)	98.6	(0.3)
and Understandings	17	99.8	(0.0)	99.9	(0.1)	99.9	(0.1)	100.0	(0.0)
Level 250	9	19.4	(0.6)*	18.7	(0.8)*	20.8	(0.9)*	27.0	(1.3)
Basic Operations and	13	64.9	(1.2)*	71.6	(1.2)*	73.1	(1.5)	76.8	(0.9)
Beginning Problem Solving	17	92.1	(0.5)	92.9	(0.5)	96.0	(0.4)	97.5	(0.4)
Level 300	9	0.8	(0.1)	0.6	(0.1)	0.6	(0.2)	1.4	(0.3)
Moderately Complex	13	17.9	(0.7)	17.8	(0.9)	15.9	(1.0)*	20.5	(0.9)
Procedures and Reasoning	17	51.4	(1.1)*	48.3	(1.2)*	51.1	(1.2)*	58.7	(1.5)
Level 350	9	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)
Multi-step Problem	13	0.9	(0.2)	0.5	(0.1)	0.4	(0.1)	0.7	(0.2)
Solving and Algebra	17	7.4	(0.4)	5.4	(0.4)	6.4	(0.4)	6.5	(1.0)

* Shows statistically significant difference from 1988, where α = .05 per set of three comparisons (each year compared to 1988). (No significance test is reported when the proportion of students is either >95.0 or <5.0.) Standard errors are presented in parentheses.

were treated as wrong when they should have been treated as "omit." This error was found only in the 1986 bridge data set for age 9. The consequence of this error on the proficiency score is very small for two reasons: It involved only 8 percent of the responses for a particular item, and the subjects who selected the "I don't know" option had the lowest mean proportion correct among all options. In fact, using the trend item parameters from 1986 estimated on the incorrect data sets, we compared the means of the ability distributions of two data sets with and without correction of the 1986 age 9 trend and found that they differed by about .07 in the proficiency scale. In order to assess administration effect as accurately as possible, however, the item parameters for all items were estimated for age 9 based on the 1986 and 1988 corrected bridge data sets. The estimated item parameters for three ages are listed in Tables D.14, D.15, and D.16 in Appendix D.

The imputed proficiency values of the 1988 sample were calculated from the responses on cognitive items and background questions based on the item parameters estimated on the trend samples of 1986 and 1988. At this point, the imputed values of the 1988 sample were not comparable to the trend scale of 1986. Note that the 1986 sample was used to obtain two separate sets of trend item parameters, the one for the data up to and including 1986 and the other for the data from 1986 and 1988. This design enabled us to use common-population equating based on the same sample, and also to express the difference in the distribution of proficiency between 1986 and 1988 in terms of the trend scale established in 1986. The linear transformations were derived separately for ages 9 and 13 to match, within each age cohort, the two means and standard deviations of proficiencies of the 1986 bridge sample, one based on the item parameters estimated on the data until 1986 and the other

Table 7.7

Science Weighted Mean Proportion Correct

	Block	1986	(N)	1988	(N)	Item[c]
Age 17	1	60.5	(2223)[a]	60.6	(634)	27
(paper)	2	59.0	(1935)[a]	60.7	(619)	32
	3	53.7	(2282)[a]	56.3	(609)	23
	Total	58.0	(5611)[a]	59.5	(1862)	82
Age 17	1	63.3	(1934)[b]			27
(taped)	2	63.4	(1934)[b]			32
	3	58.9	(1934)[b]			23
	Total	62.1	(3868)[b]			82
Age 13	1	52.5	(2075)	53.8	(1405)	25
(taped)	2	54.2	(2054)	54.7	(1281)	31
	3	56.2	(2071)	57.8	(1256)	27
	Total	54.3	(6200)	55.5	(3942)	83
Age 9	1	59.4	(2315)	62.6	(1274)	18
(taped)	2	52.5	(2361)	53.5	(1240)	25
	3	68.5	(2256)	69.0	(1197)	20
	Total	59.5	(6932)	61.0	(3711)	63

[a] Age-only BIB sample with at least one science trend block.

[b] 1986 age 17 trend sample blocks 1 and 2 were paired.

[c] Includes some items that were excluded from IRT scaling.

based on the item parameters estimated on the 1986 and 1988 data. The linear constants derived from those transformations were applied to the 1988 data set to obtain trend points for 1988. For age 17, we applied an equating method identical to that used for age 17 mathematics data. The conditioning variables and the estimated conditioning effects are given in Tables D.17 to D.19 (Appendix D) for all three ages. The linear coefficients used for the three ages are presented in Table 7.8.

Table 7.8

Coefficients of the Linear Transformation of the Trend Scale from Original Units to the Science Proficiency Scale

Age	Intercept	Slope
9	225.59	41.15
13	254.19	36.92
17	289.34	43.05

The trends in mean proficiency with jackknifed standard errors for subpopulations of the three age samples are listed in Tables 7.9 - 7.11. The 1986 and 1988 posterior distributions of science proficiency were calculated for each cohort separately at 40 quadrature points. Overlays of distributions from the two assessment years appear in Figures 7.5 through 7.7. For age 17, the 1986 distribution is calculated on the 1986 bridge sample as well as on the age-only subsample of the 1986 BIB sample, which is comparable to the 1988 bridge sample of age 17. The shape of the distributions of the two assessment years is quite similar for each cohort.

In 1986, using the range of student performance on the NAEP science scale, five levels of science proficiency were established and described in detail in the *Science Report Card* (Mullis & Jenkins, 1988): Level 150--Knows Everyday Science Facts, Level 200--Understands Simple Scientific Principles,

Table 7.9

Weighted Science Proficiency Means
and Standard Errors, Age 9

	1977		1982		1986		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	219.9	(1.2)*	220.9	(1.8)*	224.3	(1.2)*	228.9	(1.3)
Sex								
Male	222.1	(1.3)*	221.0	(2.3)*	227.3	(1.4)	232.1	(1.6)
Female	217.7	(1.2)*	220.7	(2.0)	221.3	(1.4)	225.7	(1.6)
Ethnicity								
White	229.6	(0.9)*	229.1	(1.9)*	231.9	(1.2)*	237.4	(1.3)
Black	174.9	(1.9)*	187.1	(3.0)*	196.2	(1.9)	200.1	(2.5)
Hispanic	191.9	(2.9)	189.0	(4.1)	199.4	(3.1)	201.0	(6.1)
Other	214.4	(7.5)	222.8	(5.4)	220.6	(4.6)	229.7	(3.6)
Grade								
<modal	197.6	(1.6)*	197.5	(2.9)*	204.9	(1.6)*	213.8	(2.3)
=modal	227.0	(1.2)*	230.7	(2.2)	234.3	(1.2)	236.2	(1.4)
>modal	243.9	(6.1)	265.9	(15.1)	235.0	(10.7)*	278.7	(13.3)
Region								
Northeast	224.5	(1.6)	221.8	(2.7)	228.2	(3.5)	228.9	(3.8)
Southeast	205 1	(3.0)*	214.0	(3.9)	218.8	(3.1)	223.7	(2.5)
Central	225.3	(2.2)*	226.3	(3.4)	227.9	(2.2)*	236.7	(2.9)
West	220.9	(2.3)	219.9	(4.1)	222.1	(3.2)	226.8	(2.0)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of three comparisons (each year compared to 1988).

Table 7.10

Weighted Science Proficiency Means and Standard Errors, Age 13

	1977		1982		1986		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	247.4	(1.1)*	250.2	(1.3)*	251.4	(1.4)*	257.3	(0.9)
Sex								
Male	251.1	(1.3)*	255.7	(1.5)*	256.1	(1.6)*	262.2	(1.2)
Female	243 8	(1.2)*	245.0	(1.3)*	246.9	(1.5)*	252.4	(1.0)
Ethnicity								
White	256.1	(0.8)*	257.3	(1.1)*	259.2	(1.4)*	265.2	(0.9)
Black	208.1	(2.4)*	217.2	(1.3)*	221.6	(2.5)*	229.4	(1.2)
Hispanic	213.4	(2.2)*	225.5	(3.9)	226.1	(3.1)	229.3	(4.2)
Other	235.1	(3.4)*	262.4	(11.5)	253.0	(4.0)	265.7	(5.2)
Grade								
< modal	223.4	(1.6)*	228.6	(1.6)*	234.2	(1.9)*	242.2	(1.8)
= modal	256.0	(1.0)*	258.5	(1.3)*	259.8	(1.3)*	265.3	(0.7)
> modal	284.7	(4.9)	287.4	(8.0)	266.4	(6.3)*	304.2	(9.6)
Region								
Northeast	255.3	(2.4)	254.1	(2.4)	257.6	(3.1)	259.9	(2.8)
Southeast	235.1	(1.8)*	238.7	(2.4)*	247.1	(2.2)	253.1	(3.0)
Central	253.8	(1.8)	253.9	(2.4)	249.4	(5.3)	259.2	(1.6)
West	243.0	(2.3)*	252.4	(3.0)	252.3	(2.7)	256.9	(1.7)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of three comparisons (each year compared to 1988).

Table 7.11

Weighted Science Proficiency Means
and Standard Errors, Age 17

	1977		1982		1986		1986(BIB)		1988	
Subgroup	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Total	289.6	(1.0)*	283.3	(1.1)*	288.5	(1.4)*	288.5	(1.1)*	294.2	(1.5)
Sex										
Male	297.1	(1.2)	291.9	(1.4)*	294.9	(1.9)*	294.6	(1.4)*	302.5	(2.3)
Female	282.3	(1.1)	275.2	(1.3)*	282.3	(1.5)	282.3	(1.2)	285.6	(1.9)
Ethnicity										
White	297.7	(0.7)	293.2	(1.0)*	297.5	(1.7)	296.1	(1.2)*	301.9	(1.7)
Black	240.3	(1.5)*	234.8	(1.7)*	252.8	(2.9)	254.9	(1.9)	260.0	(3.4)
Hispanic	262.3	(2.5)*	248.7	(2.4)*	259.3	(3.8)*	258.6	(2.0)*	281.8	(5.2)
Other	284.4	(4.1)	269.1	(4.9)	276.8	(11.2)	288.9	(9.9)	295.9	(11.6)
Grade										
< modal	253.2	(1.4)*	250.8	(2.2)*	259.2	(2.7)	256.7	(2.2)*	266.3	(2.9)
= modal	295.0	(0.9)*	288.9	(1.1)*	294.0	(1.6)*	296.7	(1.2)	300.6	(1.5)
> modal	300.8	(1.5)*	292.6	(2.6)*	298.6	(4.3)*	297.0	(2.8)*	317.0	(4.2)
Region										
Northeast	296.4	(2.3)	284.4	(1.9)*	292.2	(4.3)	296.6	(2.0)	303.3	(4.5)
Southeast	276.4	(1.9)*	276.3	(2.8)*	283.5	(2.0)	279.6	(2.1)*	288.0	(3.1)
Central	294.1	(1.6)	289.3	(2.4)	294.4	(2.3)	291.1	(2.3)	292.0	(3.1)
West	286.6	(1.6)	280.9	(2.7)*	283.2	(3.8)	285.5	(2.3)*	294.2	(3.2)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of four comparisons (each year compared to 1988).

Figure 7.5

Estimated Science Proficiency Distributions for the 1986 and 1988 Bridge Samples, Age 9

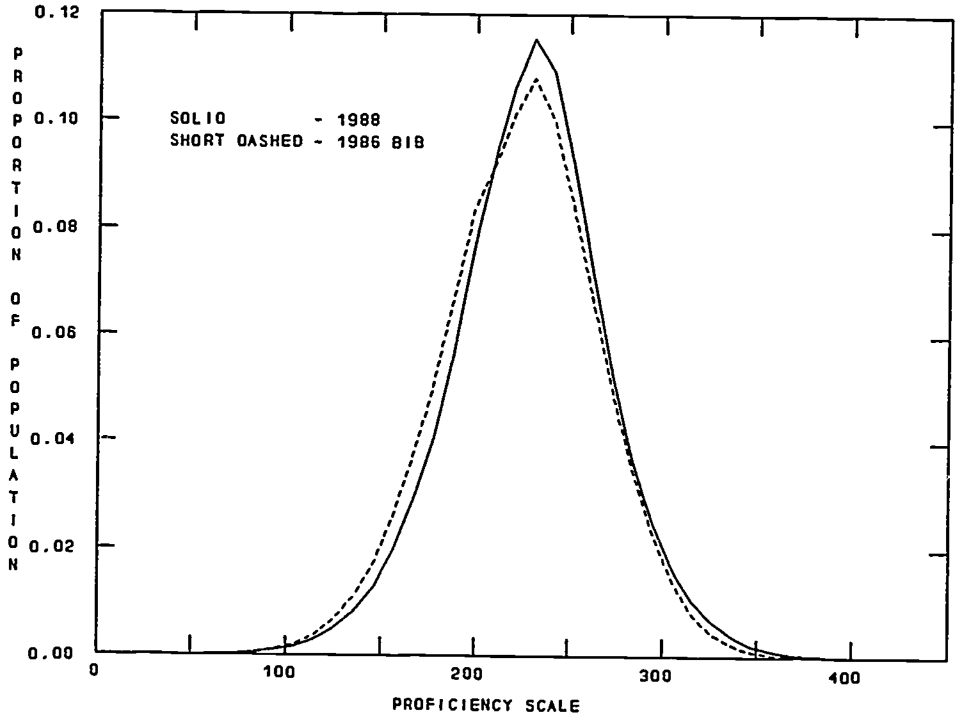


Figure 7.6

Estimated Science Proficiency Distributions
for the 1986 and 1988 Bridge Samples, Age 13

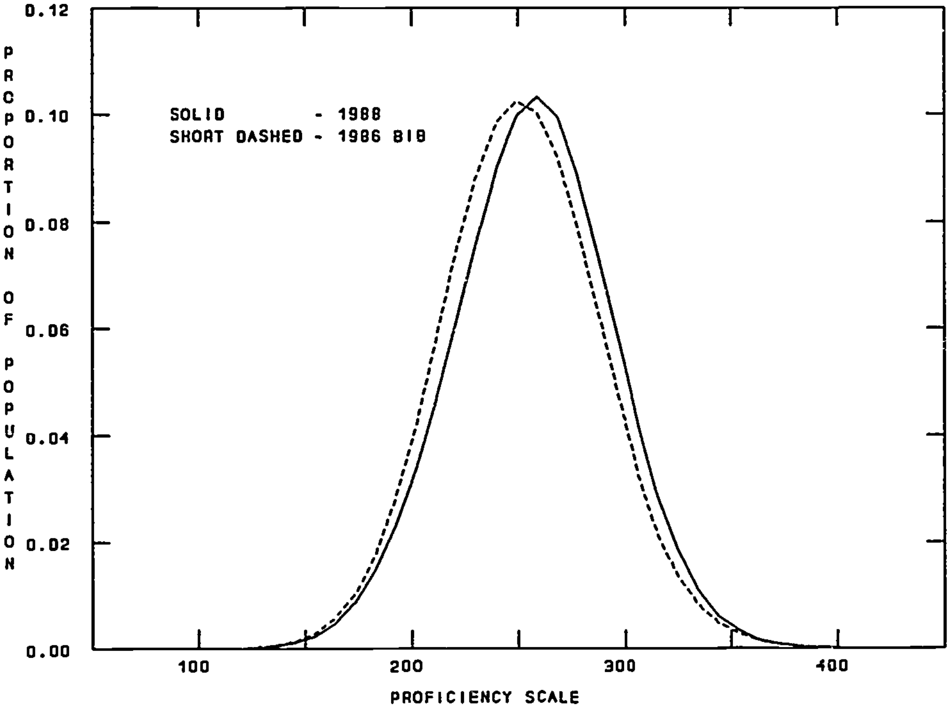
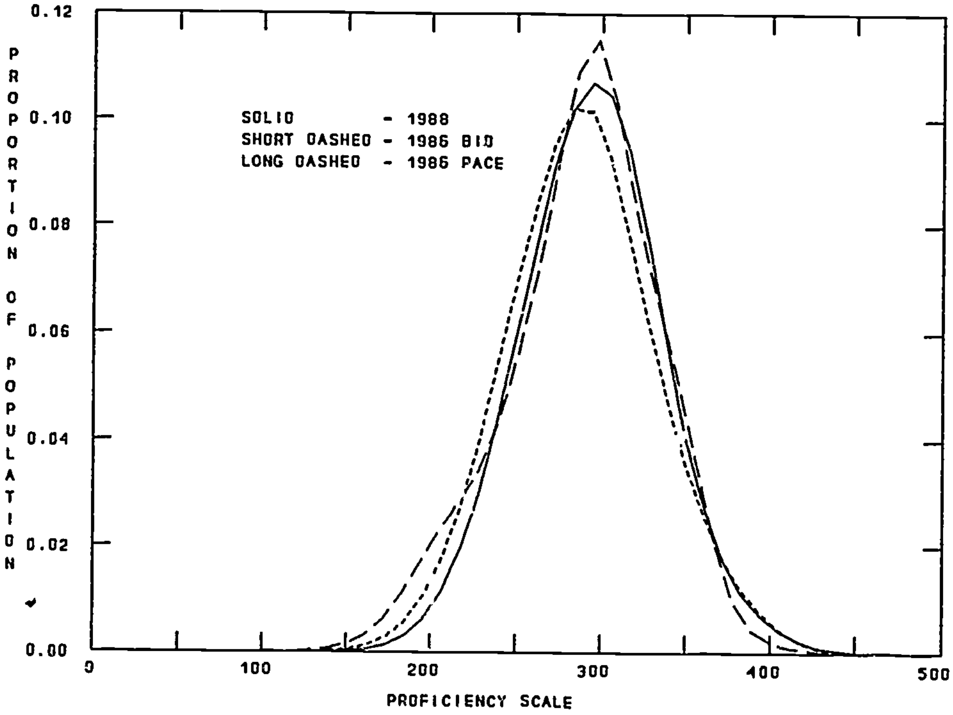


Figure 7.7

Estimated Science Proficiency Distributions
for the 1986 Age-only BIB Sample and the 1988 Age-only Bridge Sample
Age 17



Level 250--Applies Basic Scientific Information, Level 300--Analyzes Scientific Procedures and Data, and Level 350--Integrates Specialized Scientific Information. Table 7.12 shows the percentage of students at ages 9, 13, and 17 who attained each level of proficiency in the 1978, 1982, 1986, and 1988 assessments.

7.4 Major Findings for Mathematics and Science Trend Data

The four main findings of the comparison of the 1988 and the 1986 trend samples for mathematics and science are as follows:

1) For all three ages, the 1988 trend sample showed higher weighted mean proportions correct than the corresponding 1986 trend sample. This was true at the block level as well as for overall performance in mathematics and science.

2) In terms of proficiency scale means of mathematics and science for the entire sample, the 1988 sample's performance was superior to the comparable 1986 sample's performance. The improvements were statistically significant for all samples except the age 17 mathematics sample (see Tables 7.3 - 7.5 and 7.9 - 7.11), note that the desired Type I error rate was divided by the number of contrasts using a Bonferroni approach. This was true for most of the subpopulation levels as well. The means and standard deviations for mathematics and science for all three ages since 1969 are presented in Table 7.13. They are plotted in Figures 7.8 and 7.9.

3) The differences between paced administration and paper-and-pencil administration for age 17 in 1986 were not statistically significant for mathematics and science at any reporting subpopulation levels (e.g., gender, race/ethnicity, grade, or region).

4) Trends of mean proficiencies in mathematics and science closely parallel each other. Strictly speaking, any direct comparison of the value of the proficiency means in different subject areas and the changes in proficiency over time across subject areas has limited meaning. However, the shape and relative magnitudes can be compared across subject areas. The apparent large increases in mathematics and science proficiencies from 1986 to 1988 exist even though there was no context change in regard to the item order, composition, and mode and timing of presentation.

Table 7.12

Science Trends for 9-, 13-, and 17-Year-Old Students:
Percentage of Students at or Above
the Five Proficiency Levels, 1978-1988

		Assessment Year							
		1977		1982		1986		1988	
Proficiency Levels	Age	Mean	S.E.	Mean	S.E.	Mean	S.E.	Mean	S.E.
Level 150	9	93.6	(0.5)	95.0	(0.5)	96.3	(0.3)	97.3	(0.3)
Knows Everyday	13	98.6	(0.1)	99.6	(0.1)	99.8	(0.1)	99.7	(0.1)
Science Facts	17	99.8	(0.0)	99.7	(0.1)	99.9	(0.1)	100.0	(0.0)
Level 200	9	67.9	(1.1)*	70.4	(1.6)*	71.4	(1.0)*	76.4	(0.9)
Understands Simple	13	85.9	(0.7)*	89.6	(0.7)*	91.8	(0.9)	93.4	(0.6)
Scientific Principles	17	97.2	(0.2)	95.7	(0.4)	96.7	(0.4)	98.9	(0.4)
Level 250	9	26.2	(0.7)*	24.8	(1.7)*	27.6	(1.0)	31.2	(1.4)
Applies Basic	13	49.2	(1.1)*	51.5	(1.4)*	53.4	(1.4)*	59.0	(0.8)
Scientific Information	17	81.1	(0.7)*	76.8	(1.0)*	80.8	(1.2)*	86.4	(0.9)
Level 300	9	3.5	(0.2)	2.2	(0.6)	3.4	(0.4)	3.9	(0.5)
Analyzes Scientific	13	10.9	(0.4)	9.4	(0.6)*	9.4	(0.7)*	12.4	(0.7)
Procedures and Data	17	41.7	(0.8)	37.5	(0.8)*	41.4	(1.4)	44.6	(1.9)
Level 350	9	0.0	(0.0)	0.1	(0.1)	0.1	(0.1)	0.2	(0.1)
Integrates Specialized	13	0.7	(0.1)	0.4	(0.1)	0.2	(0.1)	0.5	(0.1)
Scientific Information	17	8.5	(0.4)	7.2	(0.4)	7.5	(0.6)	8.2	(1.0)

* Shows statistically significant difference from 1988, where $\alpha = .05$ per set of three comparisons (each year compared to 1988). (No significance test is reported when the proportion of students is either >95.0 or <5.0.) Standard errors are presented in parentheses.

Table 7.13

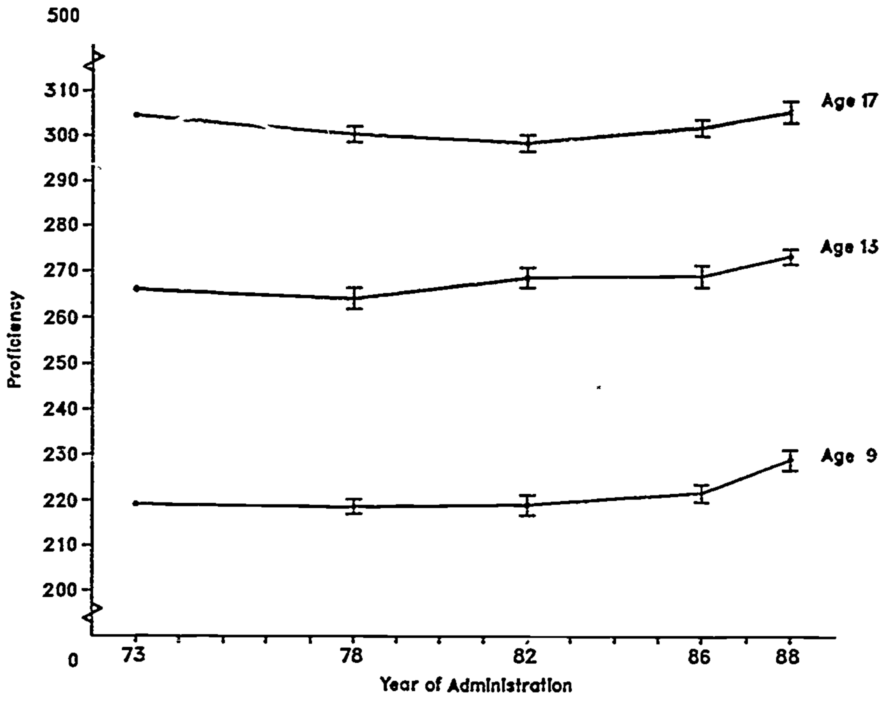
Trend of Proficiency Scale Means and Standard Deviations for Mathematics and Science

		1970	1973	1977		1978		1982		1986		1988	
		Mean	Mean	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Age 17	Math		304.4			300.4	(34.86)	298.5	(32.39)	302.0	(31.09)	305.4	(29.74)
Age 17	Sci	304.8	295.8	289.6	(44.58)			283.3	(46.67)	288.5	(44.48)	294.2	(41.37)
Age 13	Math		266.0			264.1	(38.99)	268.6	(33.36)	269.0	(30.84)	273.3	(31.74)
Age 13	Sci	254.9	249.5	247.4	(43.11)			250.2	(38.65)	251.4	(36.63)	257.3	(37.20)
Age 9	Math		219.1			218.6	(36.02)	218.0	(34.80)	221.7	(33.98)	229.0	(33.09)
Age 9	Sci	224.9	220.3	219.9	(44.88)			220.9	(40.93)	224.3	(41.48)	228.9	(40.96)

Note: 1970-1973 results are interpolated backward; standard deviations of proficiency are in parentheses.

Figure 7.8

Trend of Proficiency Scale Means for Mathematics
1973 - 1988*



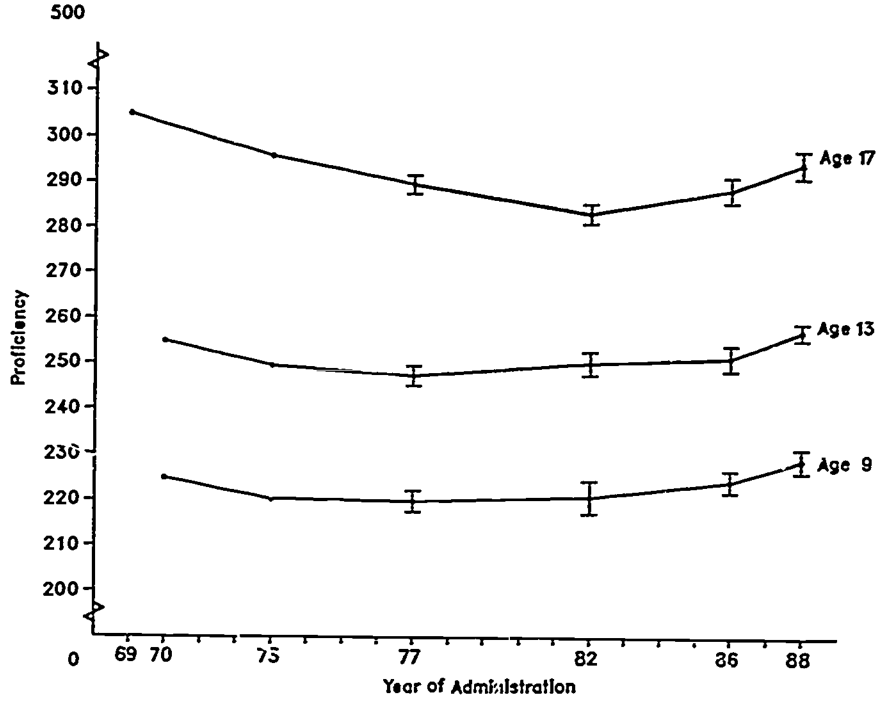
* 1973 results were interpolated for this plot. Bands extend from two standard errors below to two standard errors above the mean.

Mathematics Scale Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
1973	219.1*	266.0*	304.4*
1978	218.6 (0.8)	264.1 (1.1)	300.4 (0.9)
1982	219.0 (1.1)	268.6 (1.1)	298.5 (0.9)
1986	221.7 (1.0)	269.0 (1.2)	302.0 (0.9)
1C88	229.0 (1.1)	273.3 (0.8)	305.4 (1.2)

Figure 7.9

Trend of Proficiency Scale Means for Science 1969 - 1988*



* 1969, 1970, and 1973 results were interpolated for this plot. Bands extend from two standard errors below to two standard errors above the mean.

Science Scale Means and Standard Errors

Year	Age 9 (S.E.)	Age 13 (S.E.)	Age 17 (S.E.)
1969	--	--	304.8*
1970	224.9*	254.9*	--
1973	220.3*	249.5*	295.8*
1977	219.9 (1.2)	247.4 (1.1)	289.6 (1.0)
1982	220.9 (1.8)	250.2 (1.3)	283.3 (1.1)
1986	224.3 (1.2)	251.4 (1.4)	288.5 (1.4)
1988	228.9 (1.3)	257.3 (0.9)	294.2 (1.5)

Chapter 8

ITEM-BY-FORM VARIATION IN 1984 AND 1986 NAEP READING SURVEYS[1]

Robert J. Mislevy

8.1 Introduction

The 1984 and 1986 NAEP reading surveys employed overlapping sets of test items, but administered those items in forms that differed in length, composition, timing, and administration conditions. As discussed elsewhere in this report, it has been hypothesized that the main effects of such changes were responsible to some degree for the anomalous results observed in 1986; that is, the cumulative effect of such changes caused the assessment in a particular age/grade to become easier or harder, leading to the large, and frankly, unbelievable, differences initially observed between the 1984 and 1986 percent-correct results. This chapter investigates the magnitudes of item-by-form variation, above and beyond main effects.

While the primary investigations focus upon main effects, this ancillary study capitalizes upon the bridge data to highlight a key issue in instrument design. To anticipate, we find that modifying assessment forms can cause the accuracy of measures of change to plummet, and that similar magnitudes of this extraneous variation were found in the 1984-86 period and historical NAEP

[1] I am grateful to Nancy Allen, Al Beaton, Gene Johnson, John Tukey, and Rebecca Zwick for discussions and comments on the analyses described here.

reading assessments. The results support maintaining absolutely identical portions of instruments between successive time points to measure change, while introducing innovations in other portions.

What follows is an analysis of the components of variance of item percents-correct, which were the basis of NAEP trend reports prepared under the aegis of the Education Commission of the States. These procedures illuminate, without the complexities of the scale-score methodology, the same sources of variation that affect NAEP scale-score reports prepared under the aegis of Educational Testing Service. In the spirit of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), estimates of variance components for items-by-test-forms, items-by-time, and student-sampling shed light upon the sources of uncertainty in NAEP estimates of change, their relative magnitudes, and their likely effects under alternative assessment designs.

8.2 Background

The item percents-correct from initial analyses of the 1986 NAEP reading survey indicated sudden declines in average proficiency for 9- and 17-year-olds in the two-year period since 1984 that exceeded the largest changes ever seen in NAEP's history of comparisons over four-, five-, and six-year periods. A number of analyses were carried out with the 1986 data to check hypotheses about mechanisms that could have led to spurious declines of this magnitude (Beaton, 1988a). No proximate cause was identified in those analyses.

Other hypotheses could not be checked with those data, however, including the possibilities of effects due to the rearrangements of items on test forms, changes in administration procedures, changes in time allocations, and response modes. Changing the response mode from circling a correct answer to filling in a bubble, for example, or tightening time limits, could tend to make *all items* in the assessment more difficult. Changing an item's position from the beginning to the end of a block, changing the print size so that hyphenations appeared in different words, or moving the item from the context of easier to that of barder items, could tend to make it appear more difficult *relative to other items.*

To investigate the possibility of cumulative effects of these types, an experiment was embedded in the 1988 NAEP reading survey. At each age/grade, randomly equivalent samples of students were administered representative booklets from the 1984 survey and the 1986 survey. Each "bridge sample" survey was carried out with timings and administration conditions that matched the actual 1984 or 1986 conditions as closely as practical. The average difference over items between the two bridge samples estimates the main effect of changes in forms and administration procedures, as discussed in Chapter 5. The item-by-form variation between the bridge samples quantifies the magnitude of changes in relative item difficulties, above and beyond the main effect. These latter variations are the subject of the present chapter.

8.3 The Data

Data were obtained at each age, in the form of percents of correct responses to all the multiple-choice items that appeared in both the 1988 bridges; that is, bridges to the 1984 and 1986 assessments. There were 26 such items at age 9, 19 at age 13, and 23 at age 17. This collection is a subset of all the items that appeared in both the full 1984 and full 1986 instruments. The single professionally scored, open-ended item common to the bridges was not included due to possible scoring inconsistencies. The average number of students responding to the items in each Age, Form, Time-Point combination are shown in Table 8.1. The key features of the four samples at each age are as follows:

Sample Name	Test Form	Student Population
1984	1984 form	1984 students
1986	1986 form	1986 students
1984 b	1984 form	1988 students
1986 b	1986 form	1988 students

The item percents-correct, or "item p's," are shown in Tables 8.2 through 8.4. Because the data were obtained under complex sample designs, student responses were weighted in accordance with the students' selection probabilities. (The revised poststratification weights described in Appendix C have been used here with the age 9 and age 13 data, thereby eliminating one factor that contributed to the anomalies originally seen in 1986.)

8.4 A Model for Item-Ps

The variance-components analysis is based on a linear model for item-p's:

$$p_{ift} = \mu_i + \alpha_{if} + \beta_{it} + \varepsilon_{ift} ,$$

where

p_{ift} is the item-p for Item i on Form f from the sample of students at Time-point t.

μ_i is the population average for Item i over forms and time points.

Table 8.1

Average Sample Sizes for Item Percents-Correct

		----------- Assessment Year -----------		
Test Form	Age	1984	1986	1988
		[1984 sample]		[1984b sample]
	9	1972	-	598
1984	13	2208	-	657
	17	2381	-	604
			[1986 sample]	[1986b sample]
	9	-	2102	1135
1986	13	-	1911	1280
	17	-	1901	867

Table 8.2

NAEP Bridge Study Weighted Percents-correct, Age 9

Item	1984	1986	1984b	1986b	86-84	84b-84	86b-86	86b-84b
N001501	0.792	0.765	0.833	0.825	-0.027	0.041	0.060	-0.008
N001502	0.517	0.518	0.544	0.555	0.001	0.027	0.037	0.011
N001503	0.671	0.659	0.704	0.691	-0.012	0.033	0.032	-0.013
N001504	0.587	0.586	0.609	0.615	-0.001	0.022	0.029	0.006
N002001	0.293	0.342	0.299	0.362	0.049	0.006	0.020	0.063
N002002	0.362	0.377	0.391	0.397	0.015	0.029	0.020	0.006
N002003	0.405	0.423	0.453	0.433	0.018	0.048	0.010	-0.020
N002801	0.536	0.505	0.580	0.549	-0.031	0.044	0.044	-0.031
N002802	0.549	0.572	0.600	0.609	0.023	0.051	0.037	0.009
N003101	0.534	0.528	0.611	0.598	-0.006	0.077	0.070	-0.013
N003102	0.347	0.376	0.330	0.429	0.029	-0.017	0.053	0.099
N004101	0.664	0.597	0.691	0.619	-0.067	0.027	0.022	-0.072
N008601	0.654	0.618	0.675	0.653	-0.036	0.021	0.035	-0.022
N008602	0.562	0.540	0.574	0.560	-0.022	0.012	0.020	-0.014
N008603	0.625	0.589	0.667	0.631	-0.036	0.042	0.042	-0.036
N008901	0.678	0.709	0.695	0.754	0.031	0.017	0.045	0.059
N008902	0.697	0.713	0.684	0.767	0.016	-0.013	0.054	0.083
N009401	0.731	0.732	0.759	0.760	0.001	0.028	0.028	0.001
N009801	0.918	0.913	0.895	0.899	-0.005	-0.023	-0.014	0.004
N010201	0.860	0.813	0.876	0.837	-0.047	0.016	0.024	-0.039
N010301	0.838	0.827	0.814	0.803	-0.011	-0.024	-0.024	-0.011
N010401	0.700	0.646	0.701	0.653	-0.054	0.001	0.007	-0.048
N010402	0.361	0.376	0.370	0.371	0.015	0.009	-0.005	0.001
N010403	0.250	0.222	0.280	0.259	-0.028	0.030	0.037	-0.021
N013301	0.811	0.853	0.844	0.868	0.042	0.033	0.015	0.024
N014201	0.667	0.610	0.681	0.657	-0.057	0.014	0.047	-0.024
Average	0.600	0.593	0.622	0.621	-0.008	0.021	0.029	-0.000
Variance	0.031	0.029	0.030	0.028	0.00095	0.00055	0.00047	0.00150
Std.Dev.	0.177	0.170	0.174	0.167	0.03084	0.02336	0.02168	0.03868

Table 8.3

NAEP Bridge Study Weighted Percents-correct, Age 13

Item	1984	1986	1984b	1986b	86-84	84b-84	86b-86	86b-84b
N001501	0.934	0.958	0.975	0.940	0.024	0.041	-0.018	-0.035
N001502	0.785	0.785	0.791	0.805	0.000	0.006	0.020	0.014
N001503	0.841	0.887	0.900	0.885	0.046	0.059	-0.002	-0.015
N001504	0.805	0.819	0.847	0.821	0.014	0.042	0.002	-0.026
N002001	0.621	0.668	0.647	0.682	0.047	0.026	0.014	0.035
N002002	0.668	0.671	0.644	0.722	0.003	-0.024	0.051	0.078
N002003	0.733	0.759	0.717	0.784	0.026	-0.016	0.025	0.067
N002801	0.854	0.892	0.881	0.913	0.038	0.027	0.021	0.032
N002802	0.895	0.911	0.908	0.922	0.016	0.013	0.011	0.014
N003001	0.301	0.251	0.338	0.269	-0.050	0.037	0.018	-0.069
N003003	0.095	0.098	0.060	0.100	0.003	-0.035	0.002	0.040
N003101	0.845	0.834	0.841	0.844	-0.011	-0.004	0.010	0.003
N003102	0.746	0.782	0.744	0.786	0.036	-0.002	0.004	0.042
N004601	0.580	0.592	0.571	0.613	0.012	-0.009	0.021	0.042
N004602	0.687	0.659	0.644	0.704	-0.028	-0.043	0.045	0.060
N004603	0.818	0.764	0.785	0.838	-0.054	-0.033	0.074	0.053
N005001	0.221	0.240	0.238	0.243	0.019	0.017	0.003	0.005
N005002	0.352	0.446	0.367	0.439	0.094	0.015	-0.007	0.072
N005003	0.215	0.212	0.222	0.214	-0.003	0.007	0.002	-0.008
Average	0.631	0.644	0.638	0.659	0.012	0.007	0.016	0.021
Variance	0.065	0.067	0.068	0.069	0.00114	0.00078	0.00045	0.00149
Std.Dev.	0.255	0.258	0.261	0.262	0.03374	0.02786	0.02122	0.03861

Table 8.4

NAEP Bridge Study Weighted Percents-correct, Age 17

Item	1984	1986	1984b	1986b	86-84	84b-84	86b-86	86b-84b
N001501	0.964	0.941	0.974	0.951	-0.023	0.010	0.010	-0.023
N001502	0.888	0.846	0.916	0.841	-0.042	0.028	-0.005	-0.075
N001503	0.919	0.893	0.939	0.865	-0.026	0.020	-0.028	-0.074
N001504	0.900	0.869	0.911	0.830	-0.031	0.011	-0.039	-0.081
N002001	0.777	0.721	0.775	0.729	-0.056	-0.002	0.008	-0.046
N002002	0.811	0.755	0.808	0.799	-0.056	-0.003	0.044	-0.009
N002003	0.858	0.831	0.861	0.823	-0.027	0.003	-0.008	-0.038
N002801	0.947	0.927	0.936	0.923	-0.020	-0.011	-0.004	-0.013
N002802	0.962	0.938	0.949	0.932	-0.024	-0.013	-0.006	-0.017
N003001	0.468	0.387	0.462	0.397	-0.081	-0.006	0.010	-0.065
N003003	0.223	0.302	0.202	0.270	0.079	-0.021	-0.032	0.068
N003101	0.936	0.872	0.917	0.896	-0.064	-0.019	0.024	-0.021
N003102	0.885	0.856	0.886	0.865	-0.029	0.001	0.009	-0.021
N004601	0.704	0.681	0.689	0.699	-0.023	-0.015	0.018	0.010
N004602	0.795	0.791	0.826	0.814	-0.004	0.031	0.023	-0.012
N004603	0.877	0.856	0.890	0.868	-0.021	0.013	0.012	-0.022
N003201	0.912	0.881	0.935	0.874	-0.031	0.023	-0.007	-0.061
N003202	0.851	0.810	0.829	0.823	-0.041	-0.022	0.013	-0.006
N003203	0.736	0.701	0.778	0.693	-0.035	0.042	-0.008	-0.085
N003204	0.836	0.795	0.869	0.807	-0.041	0.033	0.012	-0.062
N005001	0.393	0.417	0.452	0.401	0.024	0.059	-0.016	-0.051
N005002	0.508	0.547	0.519	0.546	0.039	0.011	-0.001	0.027
N005003	0.314	0.288	0.284	0.327	-0.026	-0.030	0.039	0.043
Average	0.759	0.735	0.766	0.738	-0.024	0.006	0.003	-0.028
Variance	0.046	0.040	0.048	0.040	0.00110	0.00050	0.00041	0.00153
Std.Dev.	0.215	0.200	0.218	0.200	0.03321	0.02226	0.02014	0.03916

α_{if} is an effect for Item i specific to Form f. If, for a fixed f, the αs have a mean of zero, there is no main effect such as the ones suspected in the anomaly. A nonzero mean is an (undesired) form effect.

β_{it} is an effect for Item i specific to the Time t population. Nonzero β means for different values of t are the mean differences over time that the assessment is intended to measure.

ε_{ift} is an error term specific to Item i on Form f for Time t. We assume that these terms have means of zero, and are independent over items, time-points, and forms. (This independence is the only assumption in this model.)

Associated with each term is a variance component. The components relevant to present purposes are those for α, β, and ε:

σ_f^2 is item-by-form variance. Insofar as measuring change is concerned, this is noise. Its deleterious effects are not reduced by increasing the number of students in the sample, and, as we shall see, it can come to dominate the variance of estimates of change. These effects can be reduced to zero by maintaining identical forms and administration conditions.

σ_t^2 is item-by-time variance. It arises because the items in a content area become easier or harder by different amounts. Its impact on the uncertainty of change can be reduced in two ways: by increasing the number of items in a subject area, and by reporting in subject areas in which items are likely to exhibit similar changes over time.

$\sigma^2_{\epsilon ft}$ is sampling variance within Form f and Time-point t. It arises from the fact that only a sample of the population is surveyed, and can be reduced by increasing the student sample size. In this report we shall denote the sampling variance in the four samples involved in the study as σ^2_{84}, σ^2_{86}, σ^2_{84b}, and σ^2_{86b}.

We suppose that item-by-time and item-by-form variances are similar in magnitude at all time points and over forms, respectively, but allow sampling variances to differ in different assessments because NAEP examinee sample sizes often vary considerably.

An estimate of change from time point A to time point B, if the same form F is used at both time points, is the average over items of n item-p differences, where n is the number of items. Its expectation (over items i) is

$$E(\beta_{iA} - \beta_{iB}) + E(\varepsilon_{iFA} - \varepsilon_{iFB}) = E(\beta_{iA} - \beta_{iB}) ,$$

an unbiased estimate of the true average change if the items have been selected at random, and its variance is

$$[\ 2\sigma^2_t + \sigma^2_{\varepsilon_{FA}} + \sigma^2_{\varepsilon_{FB}}] \ / \ n \ .$$

If change from Time A to Time B is estimated using two different forms, F and G, the expected difference over items is

$$E(\beta_{iA} - \beta_{iB}) + E(\alpha_{iF} - \alpha_{iG}) + E(\varepsilon_{iFA} - \varepsilon_{iGB})$$

$$= E(\beta_{iA} - \beta_{iB}) + E(\alpha_{iF} - \alpha_{iG}) ,$$

which confounds change over time with difference in form. It is an unbiased estimate of the true average change only if the average item-by-form effect is zero. Its variance is

$$[\ 2\sigma_t^2 + 2\sigma_f^2 + \sigma_{\varepsilon_{FA}}^2 + \sigma_{\varepsilon_{FB}}^2\ ] \ / \ n \ .$$

Note that even if item-by-form interactions are zero on the average, the sensitivity of differences in average percent-correct as a measure of change is degraded by the additional term $2\sigma_f^2$.

Although our focus is on variance components, we should mention the relevance of the preceding paragraph to the anomaly. The items in this study are only a little more than half of those that appeared in common between 1984 and 1986, but the anomaly is reflected clearly in age 17 data (Table 8.4), where the 1986 mean item-p lies .024 below the 1984 mean. This difference may seem small from the perspective of measuring individuals, but it is larger than the change in means between any two previous assessments over time spans two to three times as long. If one ignores item-by-form and item-by-time variance when gauging the statistical significance of this difference (as was traditionally done in NAEP), the resulting t-statistic for change is about -8; comparable values for the longer time spans in the past rarely exceeded 2 in absolute value.[2] That the mean of the 1986 bridge sample lies .028 below the mean of the 1984 bridge sample--two randomly equivalent samples from the 1988 population--suggests, however, that the 1984 to 1986 drop could be due in part to a difference in test forms.

[2] A more appropriate error term, also based on Table 8.4, is the standard deviation of the item-by-item differences between 1986 and 1984, divided by the square root of the number of items, or $.033/\sqrt{23} = .007$; this gives a t-statistic of $-.024/.007 = -3.5$.

8.5 Estimating Variance Components

The variance components introduced above can be estimated from the data in Tables 8.1 through 8.4 in the following way:

Step 1. Approximate σ^2_{84}, σ^2_{86}, σ^2_{84b}, and σ^2_{86b} using item-p's, sample sizes, and design effects. The sampling variance for a particular p_{ift} is approximately

$$\frac{p_{ift}(1-p_{ift})}{N_{ift}/deff},$$

where N_{ift} is the sample size upon which p_{ift} is based and deff is a design effect that acts to increase the variance estimate due to the complex sample design. Since N_{ift} values varied little across items for fixed f and t, the average value was used in this report for all N_{ift}s in a given sample. Based on the studies summarized by E. G. Johnson (1987a), a design effect of 1.5 was employed for all items and all samples. In each sample, the average sampling variance over items was used as if it applied to all items.[3] The resulting values are as follows:

Age	σ^2_{84}	σ^2_{86}	σ^2_{84b}	σ^2_{86b}
9	16	15	51	27
13	11	13	37	18
17	9	11	33	27

Note: all entries multiplied by 10^5; e.g., 9 means .00009.

[3] While it might be preferable to transform item-p's to arcsins to better justify the use of a single sampling variance value, the average of varying sampling variance values for untransformed item-p's was employed for ease of computation and, we hope, comprehensibility.

Step 2. Compute the variances among differences between the item-p's for given items in selected pairs of samples. The expectations of the variances of these differences can be expressed as functions of the variance components of interest:

$$\mathrm{Var}(p_{86b} - p_{84b}) = 2\sigma_f^2 + \sigma_{86b}^2 + \sigma_{84b}^2 \qquad (1)$$

$$\mathrm{Var}(p_{86b} - p_{86}) = 2\sigma_t^2 + \sigma_{86b}^2 + \sigma_{86}^2 \qquad (2)$$

$$\mathrm{Var}(p_{84b} - p_{84}) = 2\sigma_t^2 + \sigma_{84b}^2 + \sigma_{84}^2 \qquad (3)$$

$$\mathrm{Var}(p_{86} - p_{84}) = 2\sigma_f^2 + 2\sigma_t^2 + \sigma_{86}^2 + \sigma_{84}^2 \qquad (4)$$

Step 3. Replacing estimated error variances from Step 1 and observed variances among item-p differences into the formulas in Step 2, solve for σ_f^2 and σ_t^2. This is done separately for each age. Equations 2 and 3 both yield approximations of σ_t^2; the average is also reported. Equation 1 yields an approximation of σ_f^2, which is reported. Substituting the estimate of σ_t^2 into Equation 4 yields a second approximation of σ_f^2, which is also reported.

8.6 Results

The estimates of variance components are shown below.

	σ_t^2			σ_f^2		
Age	Eq 2	Eq 3	Ave	Eq 1	Eq 2-4	Ave
9	3	-6*	-1	36	34	35
13	7	15	11	47	34	41
17	2	4	3	47	42	44
Average			4			40

Note: all entries multiplied by 10^5; e.g., 3 means .00003.

*The estimated value of -6×10^{-5} has been carried through for the purpose of averaging, although variances must, of course, be nonnegative.

Note that item-by-form variances, which are avoidable, dwarf item-by-time variances, which are not, roughly by a factor of ten. Also, recall that the variance of change in average item-percents correct are sums of components for sampling variance, item-by-time variance, and, if different forms are used, item-by-form variance.

Using the sampling variance figures from the 1984 assessment, we can compare the total variance that might be expected when comparing percents-correct from 1984 to 1986 had the same form been used with the same sample size at both occasions, with the total variance that might be expected with the different forms that were actually employed. The total variance using *different* forms, each comprising the same n items, is modelled as

$$2\,(\sigma_{84}^2 + \sigma_t^2 + \sigma_f^2)/n \;; \qquad (5)$$

The total variance using the *same* form, comprising, say, n* items, is

$$2\ (\sigma_{84}^2 + \sigma_t^2)/n^* . \tag{6}$$

Using averages over ages of variance component estimates, we obtain (10 + 4 + 40)/n and (10 + 4)/n* respectively for (5) and (6). These values are equal when n/n* = (10 + 4 + 40)/(10 + 4) = 3.86. Thus, even in the absence of main effects for test forms (i.e., no "anomalies"), it takes about *four times as many items* to get the same accuracy for measuring change using methodologies that differ as little as the 1984 and 1986 reading surveys, compared to using the same methodology at both occasions.

From another perspective, we can ask how many fewer *respondents* would be needed to achieve the same precision when forms are kept the same, compared to when they are different. To answer this question, we again work with the 1984 BIB error variance, and denote the student sample size by N. The modelled variance when *different* forms are used is as follows:

$$2\ (\sigma_{84}^2 + \sigma_t^2 + \sigma_f^2)/n . \tag{7}$$

A comparison based on the same forms with an examinee sample size of N* and the same design effect would have the following modelled variance:

$$2\ [(N/N^*)\ \sigma_{84}^2 + \sigma_t^2\ ]/n . \tag{8}$$

With the average values 10, 4, and 40 for σ^2_{84}, σ^2_t, and σ^2_f respectively, (7) is equal to (8) when $N/N^*=5$. This can be interpreted as saying ***the respondent sample must be five times as large*** to achieve the same precision for measuring change with *different* forms, compared to what is required when using the *same* forms at both occasions.

8.7 A Quick Comparison with Paced Presentations

Prior to 1984, NAEP reading assessments were conducted with tape recordings that paced students through their survey forms with controlled allocations of time for each item. The order of items and the length of the surveys was allowed to vary from one assessment year to the next. Time allocation under the present BIB-spiraling conditions is controlled only at the level of blocks of items approximately 15 minutes in length. In order to get a feel for the combined extent of item-by-time and item-by-form interactions in paced-administration data, item-p's were examined for 20 items at each age that appeared in NAEP in the 1975 and 1980 assessments.

As in Equation 4, the variance among item-p's across two assessment years with different paced forms confounds item-by-form and item-by-time interactions. There is a five-year difference between the 1975 and 1980 NAEP assessments, so we compared these item-p difference variances with the (1986b-1984) differences from the bridge study discussed above, which had a four-year time span; this ensures that the item-by-time components of the BIB and the paced total variances will be similar. The total variances in item-p's at ages 9, 13, and 17 were .00130, .00072, and .00092 for the (1986b-1984) BIB data, and .00067, .00094, and .00247 for the (1975 1980) paced data.

The comparable magnitudes of the BIB and paced total variances suggests that controlling the certain key aspects of the local environment of items (e.g., time allotted for a given item) in the paced format, but not others (e.g., location in assessment booklet, preceding exercises) did not produce significantly lower item-by-form variances. That is, the item-by-form variance noted above in BIB is undesirable and largely avoidable, but it does not represent a great increase over variances of the same kind that appeared to have existed under paced administration in earlier NAEP reading surveys. It may be the case, however, that controlling item-level timing and administration conditions succeeded to a larger degree in avoiding form main effects, so that anomalies like the one seen in 1986 did not arise.[4]

8.8 Assessment Versus Individual Measurement

In view of the major impact of item-by-form variation upon the sensitivity of an assessment instrument, one wonders why such effects were not anticipated and avoided, or at least incorporated into standard errors for estimates of change, since NAEP's inception. One reason may be that effects of exactly the same size often truly are negligible in the setting of individual measurement, the arena in which the "common wisdom" about educational measurement has for the most part accumulated.

Consider measuring an individual student when alternative test forms exhibit the magnitude of item-by-form variance detected between the 1984 and 1986 NAEP forms, about .00040 at the level of the individual item As in the

[4] Because different factors can contribute to form main effects and item-by-form interaction, if this could have been measured separately, finding similar magnitudes of item-by-form interaction in BIB and paced assessments would not address the question of whether anomalies qua form main effects occurred in the past.

assessment setting, the measure is imperfect for two reasons: item-by-form variation and sampling variation. In individual measurement, sampling variance at the item level is driven by observing but a single binary response; on an item with an item-p of .7, this value is .7 x.3 = .21 for a typical student. Adding item-by-form variance of .0004 increases total variance beyond sampling variance by less than *two-tenths of one percent.* In contrast, the item-level sampling variance component of the 1984 and 1986 NAEP assessments was driven by securing responses from some 2,000 students, producing a value of about .00010. Adding to this the item-by-form variance of .00040 increased total variance beyond sampling variance by *four hundred percent.* In this example, a researcher interested in individual measurement could safely ignore an item-by-form variance that would devastate precision in an assessment. (Sheehan & Mislevy, 1988, demonstrate similar effects in the setting of item response theory.)

8.9 Conclusion

Item-by-form interactions were detected in analyses of percents-correct of items that appeared in the 1984 and 1986 NAEP reading assessment instruments and samples of students administered 1984 and 1986 test forms in 1988. A quick look at historical results from previous paced NAEP reading surveys suggests that the 1984/1986 BIB item-by-form interactions, while undesirable and largely avoidable, are about the same size as corresponding effects in past NAEP assessments under paced administration.

This variation degrades the sensitivity of trend analyses. The item-by-form interactions observed in NAEP data would be negligible for comparing individual examinees or tracking an individual's performances over time, but they are large from the perspective of estimating population changes. The magnitudes of the item-by-form variances detected in the 1984 and 1986 NAEP assessments had effects comparable to cutting the number of items to one-quarter or the examinee sample size to one-fifth.

Item-by-form interactions merely reduce efficiency (albeit possibly dramatically) as long as their average effects are zero. Nonzero averages, on the other hand, can invalicate the data totally for comparing performance levels over time. It may be that controlling item-level timing and administration conditions in past NAEP assessments helped to minimize the form main effects that can cause anomalies such as the one observed in 1986. The corresponding step that is now being taken under BIB procedures is to hold some proportion of timed blocks identical across successive assessments, with respect to composition, timing, and administration conditions.

Chapter 9

EPILOGUE

Albert E. Beaton

The study of the 1988 bridges shows that the effect of changing measurement instruments can be so large that it obscures real changes in educational performance. This leads us to repeat the major lesson from the reading anomaly that was stated in Chapter 1: *When measuring change, do not change the measure*. The empirical evidence to support the wisdom of this lesson is clear enough from the results of the analyses of the measurement system changes incorporated in the 1988 bridge samples, which are summarized in Chapter 4. The work by Mislevy, shown in Chapter 8, presents further evidence by computing item-by-form interactions and showing that the amount of variance created by changing assessment forms may be substantially greater than the variance over time of student performance, which we are attempting to measure. As Mislevy shows, this variance was present even when assessment items were individually timed using a tape recorder. The lesson is clear: Changes in trend assessment methodology are fraught with danger and should be undertaken only with great care.

The pressure to make changes in ass ssment instruments and procedures is considerable. NAEP's complex consensus process involves hundreds of staff and advisors, many of whom have suggestions about how NAEP can be improved. Most, if not all, of these suggestions have merit. For example, committees of teachers reviewing items may make suggestions as to how to make individual items more precise. A printer may suggest ways to improve the artistry of a booklet. And yet, for measuring trends, these suggestions must be rejected, since they *might* render the various assessment years incomparable.

Defending the previous assessment procedures is not always easy. NAEP has been measuring trends since 1969, and there have been substantial changes in curriculum since then. For example, some formerly emphasized topics from the "new math" are no longer taught, and the pencil-and-paper computation of a square root has been de-emphasized. Over the years, NAEP has carefully removed items on such topics. Today, many believe that students should have different proficiencies, such as knowing how to use a scientific calculator. NAEP has already introduced calculator items and will use scientific calculators with the 1990 cross-sectional sample. Never changing the measurement instruments would surely make NAEP grow obsolete and uninteresting.

The tension between continuity and change is not unique to NAEP or to educational measurement. For example, as United States corporations merge, go private, or fail, the Dow Jones average must change its composition while maintaining as much continuity in interpretation as possible. Government indices such as the Consumer Price Index must also adjust to changes in popular consumption. Such changes can never be made without introducing some change in the properties of the indicator, yet the changes are necessary to keep the indicator relevant.

We believe that the proposed adjustments to the NAEP design are a prudent response to the conflicting goals of measuring trends and using up-to-date and relevant measurement. For now, we will maintain separate trend samples in which the measurement instruments and procedures are as close as possible to those used in the assessment with which the new data will be compared. Separate samples of students will be measured using the most current information about the subject area and innovative technology. Only after their properties and their relationship to the trend lines are fully understood will assessment forms and technology move from these innovative samples to the trend samples.

Investigation of the reading anomaly has reinforced the realization that no measurement is perfect, especially the measurement of changes over time. Despite applying the best available measurement technology, subtle changes in the relevance of items and small shifts in the school populations both introduce interpretive difficulties into comparisons with the past. Even holding the measurement system constant does not assure that changes in instruction and the form of learning will not affect the meaning of trends. Sampling error and other, inestimable types of error also affect the accuracy of trend estimates. The public as well as the measurement community should understand the difficulties and limitations of measurement--in education as in economics, in science, or in technology.

Despite what has been presented about the limitations of assessment, it is important to note that a national assessment is still useful, indeed indispensable, if we expect to make decisions about the path that American education should take in the future. Educational policy makers and the public want to know if there have been major shifts in educational performance, and NAEP is the best instrument we have found for measuring such changes. This study of the reading anomaly shows that it is inappropriate to overinterpret small shifts in performance that occur in a short period of time; such small shifts might be attributable to the various errors--only some of whose sizes can we estimate directly--that affect an estimate. In interpreting small changes, it is usually prudent to repeat the measurement procedure over time until the shift stabilizes as a trend or is corroborated through other sources. Although the standard error attributable to measurement may be large compared to the changes in *average* performance that has been observed over years, the standard error of a proficiency mean is quite small compared to the total variability of student performance. Put another way, the standard error is small compared to the difference between adjacent anchor points on the NAEP scales, and these anchor points represent substantial differences in student performance[1]. We have little doubt that, even in the short term, NAEP would reliably identify major shifts in educational performance, as it is intended to do. In a longer term, as it is also intended to do, it will reliably identify the cumulative effect of more or less consistent trends that are small in the short term.

Finally, although we intend to minimize changes in the assessment technology used for trend estimation, we also feel strongly that experimenting with and eventually introducing newer technology is essential for NAEP. The history of science is brimming with improvements in measurement that have resulted in better understanding of the world around us. Study of the reading anomaly has given us a fuller understanding of the strengths and weaknesses of the present NAEP assessment technology. The identification of technological limitations always presents a challenge for methodological improvement.

[1] Anchor points are used to describe what students at various levels of the NAEP scales know and can do. They are described in the reports in which the various scales are discussed and in the NAEP technical reports (Beaton, 1987, 1988b). Basically, the description of an anchor point describes what a large majority of students at that level know and can do that a majority of students at lower levels cannot.

APPENDIX A

Summary of Modifications in Reading Scale Results Used in Tables and Figures

Appendix A
Summary of Modifications in Reading Scale Results Used in Tables and Figures[a]

	Expanded Conditioning Model for 1971-1986?	Adjusted 1984 Weights?[b]	Modified 1986 Conditioning Model?[b]	1986 Results Adjusted for Context Effect?	Which Set of 1988 Results Used?[c]
Figure 1.1	Yes	No	No	No	Not applicable
Figure 1.2	Yes	Yes	Yes	Yes	Set 2
Table 4.1	[All changes and adjustments are detailed in this table]				Set 2
Figure 4.1	Yes	Yes	Yes	No	Set 1
Table 5.9	Yes	Yes	Yes	No	Set 1
Figure 5.1	[Identical to Figure 4.1]				
Figure 6.1	[Identical to Figure 1.2]				
Table 6.1	Yes	Yes	Yes	Unadjusted and adjusted results given	Not applicable
Tables 6.2-6.4	Yes	Yes	Yes	Unadjusted and adjusted results given	Set 2

[a] See Chapter 4 for further detail.

[b] Applies to ages 9 and 13 only.

[c] Two sets of results were obtained for the 1988 bridge to 1984: (1) a set that uses the same conditioning variables as the 1988 bridge to 1986, maximizing the comparability of the results for the two bridges (see Chapters 4 and 5) and (2) a set that uses an expanded conditioning model that maximizes comparability with the 1984 results. Set 2 is most appropriate for assessing trend and is used in the most recent reading trend report, *The Reading Report Card, 1971 to 1988* (Mullis & Jenkins, 1990). In the present report, figures and tables that compare the 1988 bridge to 1984 to the 1988 bridge to 1986 use Set 1; figures and tables that do not include the 1988 bridge to 1986 use Set 2.

APPENDIX B

Sampling and Weighting Procedures for the 1988 NAEP Bridges

Appendix B

SAMPLING AND WEIGHTING PROCEDURES FOR THE 1988 NAEP BRIDGES

Eugene G. Johnson
Keith F. Rust

Each of the bridge samples drawn as a part of the 1988 assessment was designed to replicate the administration of an earlier NAEP assessment. Thus the sampling and weighting procedures used in these bridges were designed to repeat as closely as feasible the procedures used previously. Some changes from the previous procedures were necessary, however. In particular, the poststratification procedures[1] used in 1988 differed somewhat from those used in 1986 and 1984; these changes are described below. The effects of these changes in procedures on proficiency scores are also given below and are shown to be relatively small.

THE 1988 BRIDGE SAMPLES

The bridge studies included in the 1988 assessment that pertain to the current report are as follows:

Bridge to 1984: This bridge consists of samples comparable to the 1984 main assessment and addresses the subject areas of reading and writing. The samples are collected by grade and by age for age 9/grade 4, age 13/grade 8,

[1] In poststratification, the sampling weights are adjusted to make sample estimates of certain subpopulation totals conform to external, more accurate, estimates.

and age 17/grade 11, using the age definitions and time of testing equivalent to those used in 1984. Six assessment booklets were administered at each age/grade, each of these booklets consisting of at least one block of reading items and at least one block of writing items. The administration of these bocklets was nonpaced (that is, no audiotape was used). Thus at all three ages a spiral, print-administered bridge of reading and writing was conducted. The booklets used formed part of the spiral assessment in 1984, when reading and writing were both administered. For the 1984 sample these assessments were weighted as part of the full spiral sample, using 39 poststratification cells for each age (although only 26 of these are relevant to the age eligibles, the group of interest across time).

The bridge samples for 1988 consist of approximately 4,000 age-eligible and approximately 5,200 age/grade-eligible students at each age class. The original 1984 spiral samples consisted of 26,000 to 29,000 age/grade-eligible students. The level of poststratification used in 1984 appears to be about the full extent possible without giving rise to reduced gains in estimation efficiency. Since the 1988 bridge samples are based on many fewer students than the 1984 spiral samples, it did not seem appropriate to use the same poststrata for the 1988 bridge samples and so some collapsing of poststrata was performed. The comparability of weighting procedures of the original and the bridge samples will be discussed later in this appendix.

<u>Bridge to 1986, Ages 9 and 13</u>: This bridge consists of samples for ages 9 and 13 comparable to those used for the measurement of trend in 1986. The samples were collected by age only and used age definitions and time of testing equivalent to those used in 1984 and in the 1986 bridge to 1984. The

subject areas addressed by this tridge are reading, mathematics, and science. Three assessment booklets were administered at each of the ages 9 and 13, and these are the same booklets as were administered in 1986. Each booklet contains one blo. of reading, one block of mathematics, and one block of sci ce exercises. As in 1986, the mathematics and science blocks were aiministered using a tape recorder while the reading blocks were administered by pencil and paper only. The three tape sessions at each age were conducted to replicate the fall and winter bridges conducted in 1986. The numbers of students for the two sets of samples are similar--around 2,000 age eligibles each in 1986 and around 1,333 each in 1988. Although time restrictions prevented the exact repetition of the poststratification procedures, comparability has been maintained as much as possible (specifically, by not using age and grade eligibility for nonresponse adjustment and poststratification). Seven poststrata were used for each age in 1988 (compared to eight in 1986), with five of the poststrata having the same definition across the two assessments.

Bridge to 1986, Age 17: This bridge consists of a sample of age 17/grade 11 students comparable to the 1986 main assessment using an equivalent age definition and time of testing to that used in that assessment and, since those definitions are also the same, for the 1984 assessment. The subject areas addressed by this bridge are reading, mathematics, science, and history. Seven assessment booklets were administered to age 17/grade 11 students. One consisted entirely of blocks of history items; the remaining six consisted of blocks of reading, mathematics and science items. The administration of these booklets was nonpaced. The books of reading, mathematics, and science were administered as part of the full spiral sessions in 1986, where their purpose was to bridge to 1984. In the 1988 bridge they were repeated in separate spiral sessions since the age definition is different from the regular Age 17 assessment in 1988. As in the other spiral bridges, it was not possible to repeat the full level of poststratification that was used on the 1986 sample, where 26 poststratification cells were used for age-eligible students, and 39 in total.

SAMPLE DESIGN

The sample of students for the 1988 NAEP assessment was selected using a complex multistage sample design involving the sampling of students from selected schools within 94 selected geographic regions, called primary sampling units (PSUs), from across the United States. All 94 PSUs were used for the main 1988 assessment and subsamples of these PSUs were used for the bridge assessments. The sample design, which is similar to that used in 1986, will be described in detail by Westat, Inc., the firm subcontracted by ETS to select the sample, in *National Assessment of Educational Progress--1988 Sampling and Weighting Procedures, Final Report*. This section will provide an overview of the design. Since the PSUs used for the bridge assessments were subsamples of those used for the main assessment, the selection of the main assessment PSUs is given first.

Primary Sampling Units for the Main Assessment

In the first stage of sampling, the United States (the 50 states and the District of Columbia) was divided into geographic primary sampling units, where each PSU met a minimum size requirement and comprised either a metropolitan statistical area (MSA), a single county, or a group of contiguous counties. Twelve subuniverses of PSUs were then defined as described below.

The 34 largest PSUs were designated as certainty units because they were so large as to be selected with probability one. The remaining, smaller, PSUs were not guaranteed to be selected into the sample. These were grouped into a number of noncertainty strata (so called because the PSUs in these strata were not included in the sample with certainty).

The PSUs were classified into four regions, each containing about one-fourth of the U.S. population. In each region, PSUs were classified as MSA or nonMSA. In the Southeast and West regions, the PSUs were further classified as high minority (at least 20 percent of the population in the 1980 Census was either Black or Hispanic) or not. The resulting subuniverses are shown below.

Table B.1

The Sampling Subuniverses
and the Number of Noncertainty Strata in Each

	MSA PSUs		NonMSA PSUs	
Region	Regular Strata	High-minority Strata	Regular Strata	High-minority Strata
Northeast	8	--	2	--
Southeast	4	6	4	6
Central	8	--	6	--
West	4	6	4	2
Total	24	12	16	8

Within each major stratum (the subuniverses), further stratification was achieved by ordering the noncertainty PSUs according to several additional socioeconomic characteristics, yielding 60 strata. One PSU was selected with probability proportional to size from each of the 60 noncertainty strata.

PSUs within the high-minority subuniverses were sampled at twice the rate of PSUs in the other subuniverses.

These 94 PSUs were used for the main assessments of all three age classes. To allow for the estimation of within-school-year growth in achievement and to match the administration times of previous assessments, the assessment sample was divided into two randomly equivalent subsamples, one subsample to be assessed in the winter and the other to be assessed in the spring. For this purpose, the 94 PSUs were designated as winter PSUs, spring PSUs, or both winter and spring PSUs according to the following scheme. The 18 largest certainty PSUs were designated as both winter and spring PSUs, to be included in the sample for both seasons (the sample of schools within each of these PSUs was randomly split in half, one subsample to be assessed in the winter and one to be assessed in the spring). The 16 smaller certainty PSUs were ordered by region and then alternately designated as winter PSUs or spring PSUs, resulting in 8 PSUs for each season. Similarly, alternate members of the set of the 60 noncertainty PSUs, arranged in stratum order within each subuniverse, were designated as winter or spring PSUs. The end result was 56 winter PSUs, 38 in which assessments were conducted only in the winter and 18 in which assessments were conducted in both winter and spring, and 56 spring PSUs, consisting of 38 in which only spring assessments were conducted plus the 18 winter and spring PSUs.

Primary Sampling Units for the Bridge Assessments

The bridge assessments used a subsample of the 94 PSUs used for the main assessment. The age 9/grade 4 bridge assessments, which were conducted in the winter, used the 56 PSUs designed as winter PSUs in the main assessment; the age 17/grade 11 bridge assessments conducted in the spring, used the 56 PSUs designated as spring PSUs. The age 13/grade 8 bridge assessments, conducted in the fall, used 64 PSUs that were selected from the complete set of 94 PSUs with probability proportional to a measure of size. As with the winter and spring subsamples, the 18 largest certainty PSUs were retained in the fall bridge sample with certainty.

Schools for Bridge Samples; the Assignment of Sessions to Schools

Schoo's to participate in the age 13/grade 8 bridge assessments (conducted in the fall) were selected from the subsample of 64 PSUs that had been designated as the age 13/grade 8 bridge PSUs. To avoid the possibility that a particular bridge session might be assigned to a school with only one or very few eligibles, small schools were clustered with other schools in the same PSU to form clusters of a specified minimum number of eligibles. Bridge sessions were then assigned within each PSU by selecting a school cluster with probability proportional to the estimated number of age and grade eligibles within the school (or school cluster).

Schools to participate in the age 9/grade 4 bridge assessments (conducted in the winter) were selected from the subsample of the PSUs designated as being for the winter assessment. The selection was such that each of the distinct booklets used in the bridge assessments would be administered at least once within each of the 56 PSUs designated as winter or both winter and spring PSUs. Clusters of schools were formed in the same manner as for age 13/grade 8; in this case, two clusters were selected per PSU.

In a like manner, schools to participate in the age 17/grade 11 bridge assessments (conducted in the spring) were selected from the subsample of the PSUs designated as being ᶜor the spring assessment such that each of the distinct booklets used in the bridge assessments would be administered once within each of the 56 spring PSUs. Two clusters of schools were selected per PSU.

For all three age/grades, sessions were assigned to bridge sample schools in the following manner. First, the number of sessions per school was established. This was the maximum, up to four, that could be administered without creating unduly small session sizes with few eligibles. Thus in most bridge sample schools four types of session were conducted, but, for example, schools with fewer than 20 eligibles were asked to conduct just a single session. The assignment of sessions to schools was performed so as to maximize the number of session types conducted within each PSU. Thus, to the extent feasible, session assignment was delayed until after it was determined that a selected school would participate in the sample. Because this happened sometimes but not always, two types of school nonresponse adjustment factor, denoted school and session, were required.

This procedure assured that each session type was assigned in each PSU at least once for the age 9/grade 4 and age 17/grade 11 samples. At age 13/grade 8, however, sometimes a PSU was represented in the sample by a single large chool. As it was not considered feasible to administer each of five different session types in a single school, not all session types were administered in all 64 PSUs, but each session type was administered in most PSUs.

Sampling Students

In the fourth stage of sampling, a consolidated list of all grade-eligible and age-eligible students was established for each school. A systematic selection of eligible students was made if necessary to provide the target sample size (otherwise all eligible students were selected) and, for bridge sample schools assigned both pencil and paper and paced-tape assessments, students were randomly assigned by Westat district supervisors to print or tape sessions using prespecified procedures. Students assigned to paced-tape sessions who were not age-eligible were dropped from the assessment.

Excluded Students

Some students selected for the sample were deemed unassessable by the school authorities because they had limited English language proficiency, were judged as being educable mentally retarded, or were functionally disabled. In these cases, an Excluded Student Questionnaire was filled out by the school staff listing the reason for excluding the student and providing some background information. The same guidelines for exclusion were employed for all bridges as well as for the main assessment. For the excluded students, unlike the assessed students, no distinction was made as to the season of the year in which their school was assessed since the timing of the assessment is unimportant for these unassessed students. Consequently, for age 9/grade 4 and age 13/grade 8, no distinction is made between students excluded from the bridge assessments and the students excluded from the main assessments since the same grade and age eligibility definitions apply in each case. Since this is not the case for the third age class, the excluded students from the bridge assessments (with an October-September age definition and modal grade of 11) are treated as separate from the excluded students from the main assessment (with a calendar-year age definition and modal grade of 12).

PROCEDURES TO DERIVE STUDENT SAMPLE WEIGHTS

The weight assigned to a particular student reflects two major components of the sample design and the population being surveyed. The first component, the student's base weight, reflects the probability of selection of the student for participation in a particular type of assessment session (i.e., a particular bridge assessment session or for the main assessment). As explained below, 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 to ensure that estimates, based on the weights, of certain subpopulation totals correspond to values reliably estimated from external sources (i.e., Census and Current Population Survey). This latter form of adjustment, known as poststratification, reduces sampling variability and may also reduce the bias resulting from noncoverage and nonresponse.

Apart from changes in the poststratification procedure, detailed below, the weighting procedures used for the 1988 bridges were essentially the same as those used in 1986 and 1984.

As mentioned above, the base weight assigned to a student is the reciprocal of the probability that the student was invited to a particular type of assessment session. The base weight for a selected student was adjusted by three nonresponse factors: one to adjust for noncooperating schools, the second, used only in the case of bridge samples, to adjust for allocated sessions that were not conducted, and the third to adjust for students who were (or should have been) invited to the assessment but did not appear either in the scheduled session or a makeup session. For spiral sessions, the student nonresponse adjustment was made separately for two classes of students in a PSU by age class: those in or above the modal grade for their age and those below. This differentiation acknowledges likely and observed differences between students in the two classes both in their assessed abilities and in their likelihood of nonresponse. For some sessions in some PSUs, these two classes were combined, since one or both was too small to form the basis for an adjustment factor. The student nonresponse adjustment for students sampled for tape sessions was similar except that, to achieve comparability with the prior assessments, the adjustment was computed within a PSU for each tape booklet across all students originally selected for that booklet.

A few students were assigned extremely large weights. One cause of large weights was underestimation of the number of eligible students in some schools leading to inappropriately low probabilities of selection for those schools. Other extremely large weights arose as the result of relatively high levels of nonresponse coupled with low-to-moderate probabilities of selection. Students with extremely large weights can have an unusually large impact on estimates such as weighted means. Since the variability in weights contributes to the variance of an overall estimate, a few extremely large weights are likely to produce large sampling variances of the statistics of interest. In such cases, a procedure of trimming the more extreme weights to values somewhat closer to the mean weight was applied in order to reduce the mean square errors of the estimates.

POSTSTRATIFICATION

As in most sample surveys, the weight assigned to a respondent is a random variable that is subject to sampling variability. If there were no nonresponse, the respondent weights would provide unbiased estimates of the various subgroup proportions. However, since unbiasedness refers to average performance over all possible replications of the sampling, it is unlikely that any given estimate, based on the sample actually obtained, will exactly equal the population value. Furthermore, the respondent weights have been adjusted for nonresponse and a number of extreme weights have been reduced in size.

To reduce the mean squared error of estimates, the sampling weights were further adjusted so that estimated population totals for a number of specified subgroups of the population, based on the sum of weights of students of the specified type, were the same as presumablv better estimates derived from other sources. This adjustment, called poststratification, reduces the mean squared error of estimates relating to student populations that span several subgroups of the population.

The poststratification procedures used for the 1988 NAEP data differ from those used for the 1984 and the 1986 assessments. To make the differences clear, the 1986 and 1984 procedures will be explained.

1986 and 1984 Poststratification Procedures

The same poststratification procedures were used for both the 1984 and 1986 assessments. For the spiral assessments, 13 subgroups were defined in terms of race, ethnicity, census region and community size (SDOC) as shown in Table B.2. Each of the 13 subgroups was further divided into three classes:

(a) students eligible for inclusion in the sample by both age and grade;

(b) students eligible for inclusion by age only;

(c) students eligible for inclusion by grade only.

Table B.2
Major Subgroups for Poststratification in 1986 and 1984

Subgroup	Race	Ethnicity	Region	SDOC*
1	White	Non-Hispanic	NE	1, 2
2	White	Non-Hispanic	NE	3, 4, 5
3	White	Non-Hispanic	SE, Central	1, 2
4	White	Non-Hispanic	SE, Central	3
5	White	Non-Hispanic	SE, Central	4, 5
6	White	Non-Hispanic	West	1, 2
7	White	Non-Hispanic	West	3, 4, 5
8	Any	Hispanic	NE,SE,Central	Any
9	Any	Hispanic	West	Any
10	Black	Non-Hispanic	NE	Any
11	Black	Non-Hispanic	SE	Any
12	Black	Non-Hispanic	Central, West	Any
13	Other	Non-Hispanic	Any	Any

*SDOC (Sample Description of Community) categories: 1--Big City; 2-- Fringe of Big City; 3--Medium City; 4--Small Place; and 5--Extreme Rural.

This resulted in 39 poststratification cells for each age class. The final weight for a student was the product of the base weight (after adjusting for nonresponse and after trimming to reduce the size of certain extremely large weights) and a poststratification factor whose denominator was the sum of those weights for the cell to which the student belongs and whose numerator was an adjusted estimate of the total number of students in the cell. This adjusted estimate was a composite of estimates from the NAEP sample and independent estimates based on projections based on Current Population Survey estimates and Census projections. The adjusted estimate was a weighted mean of the various estimates, the weights being inversely proportional to the approximate variances of the NAEP and independent estimates.

The sample of students in each of the paced-tape administered assessments was much smaller than the sample for the spiral assessments. Consequently, some subgroups in Table B.2 were collapsed for poststratification as follows:

1, 2
3
4
5
6, 7
8, 9
10, 11, 12
13

Furthermore, to achieve comparability with earlier assessments, there was no subdivision into eligibility classes (of students eligible by age, grade, or both), so there were eight poststratification cells for each age class.

1988 Poststratification Procedures

The poststratification in 1988 was done for each age/grade and separately for each of the spiral assessments and each of the tape assessments. Within each age/grade and assessment-type group, poststratification adjustment cells were defined in terms of race, ethnicity, and NAEP region as shown in Table B.3.

Table B.3
Major Subgroups for Poststratification in 1988

Subgroup	Race	Ethnicity	Region
1	White	Non-Hispanic	NE
2	White	Non-Hispanic	SE
3	White	Non-Hispanic	Central
4	White	Non-Hispanic	West
5	Any	Hispanic	Any
6	Black	Non-Hispanic	Any
7	Other	Non-Hispanic	Any

This grouping resulted in seven cells for each tape sessioii. For the spiral samples, each of the seven subgroups was further divided into the three eligibility classes:

(a) students eligible by both age and grade;

(b) students eligible by age only;

(c) students eligible by grade only.

In brief, the new poststratification procedures differ from those used for the 1984 and the 1986 assessments in three ways:

1) The 1988 poststrata totals incorporate current Census Bureau monthly population estimates by single years of age by race/ethnicity groups. Such monthly estimates were not available at the time of the poststratification of the 1984 and 1986 weights. The use in 1988 of estimates of in-school eligitles based on data relating only to the particular grade and age in question eliminated the need to derive year-to-year retention factors foi age 17 students and the need to incorporate projections from younger ages and lower grades, as was done in 1984 and 1986.

2) The number of cells used in poststratification was reduced from the 39 cells used in 1986 and 1984 to the 21 cells used in 1988. The 21 poststrata used for 1988 vary substantially in mean performance level and yet are large enough to produce reasonably stable poststratification factors. The reduction in the number of cells from 39 to 21 was made to increase the stability of the poststratification factors in an effort to reduce the sampling variance.

3) The 1988 poststrata totals were derived solely from CPS data and Census Bureau population projections and, in contrast to the method used in previous years, did not use any data from the 1988 NAEP samples.

The new procedure was adopted in order to speed up the production of the weights, since poststrata totals based only on CPS and Census data can be derived well in advance of the weighting of the data.

It is clearly important to ascertain the impact of these changes in poststratification on the estimates of subgroup proficiencies. In particular, it is important to establish that the measurement of trend in subgroup proficiencies is affected in a minimal way by this revision in procedures. The approach used to ascertain the effect of the change in poststratification procedures was to reweight the 1986 samples according to the new procedures and then compare the results with the previous results. (This approach is considerably more cost- and time-efficient than the alternative approach of reweighting the 1988 data according to the 1986 procedures.)

Tables B.4, B.5, and B.6 show the result when the age eligible students in the trend samples of the 1986 assessment of reading are reweighted using the new poststratification factors. The first two columns in each table compare the new procedure with the old in terms of the estimated relative frequencies by race/ethnicity, region, parental education, and grade. The last two columns compare the two procedures in terms of the mean reading proficiencies for those subgroups. (It should be noted that the standard errors of the proficiency estimates do not include the component due to the variability of the linear equating function--see Appendix E for a discussion.)

An examination of these tables shows that the effect of changing the poststratification procedure on mean proficiency estimates is slight: in most cases, the difference between the proficiency estimates based on the two procedures is less than one standard error (of the mean proficiency based on the old method) and in every case the difference is less than 1.25 standard errors. Since these standard errors do not include the variability due to equating and are, consequently, underestimates of the true standard errors of the mean proficiencies, the differences between estimates based on the two poststratification methods are well within the fluctuations to be expected by chance in either of the individual estimates.

We note that the standard errors of the difference between the original and revised estimates are likely to be relatively small, due to the high degree of correlation between the two sets of estimates. However, the important aspects of the change in the method are the sizes of the resulting differences in estimates, relative to the precision of the estimates themselves, as discussed above.

Table B.4

Effect of Change in Poststratification Procedures:
Relative Frequencies and Mean Reading Proficiencies, Age 9

	Relative Frequencies		Mean Reading Proficiencies	
	New Procedure	Old Procedure	New Procedure	Old Procedure
Observed Race/Ethnicity				
White	76.0%(1.0)	76.5%(1.1)	214.7(1.5)	214.9(1.3)
Black	15.5%(0.5)	14.9%(0.5)	186.4(1.6)	185.0(1.6)
Hispanic	6.0%(1.1)	6.2%(1.1)	189.0(2.9)	189.8(3.3)
Other	2.4%(0.5)	2.5%(0.5)	204.7(6.2)!	203.7(6.6)!
Region				
Northeast	20.7%(1.1)	21.1%(1.1)	212.0(3.0)	212.3(2.7)
Southeast	25.9%(2.0)	22.5%(4.7)	205.2(3.2)	202.5(2.7)!
Central	26.2%(0.9)	28.6%(4.0)	211.7(2.5)	212.9(2.7)
West	27.2%(1.6)	27.7%(1.6)	206.0(3.1)	206.5(3.0)
Grade				
< Modal Grade	34.2%(1.7)	33.9%(1.7)	188.3(1.2)	189.4(1.4)
at Modal Grade	65.5%(1.7)	65.8%(1.7)	218.9(1.3)	218.5(1.2)
> Modal Grade	0.3%(0.1)	0.3%(0.1)	238.2(8.8)!	241.9(11.3)!
Parental Education				
Not Graduated H S	4.3%(0.4)	4.2%(0.4)	190.1(2.9)	189.5(2.8)
Graduated H S	16.0%(0.8)	16.4%(0.7)	201.5(1.4)	202.2(1.9)
Post H S	44.7%(1.2)	44.4%(1.2)	219.2(1.4)	219.0(1.3)
Total			208.5(1.3)	208.6(1.2)

Note: Standard errors in parentheses (standard errors do not include equating error)

! Interpret with caution--the sampling error cannot be accurately estimated, since the coefficient of variation of the estimated total number of students in the subpopulation exceeds 20 percent.

Table B.5

Effect of Change in Poststratification Procedures:
Relative Frequencies and Mean Reading Proficiencies, Age 13

	Relative Frequencies		Mean Reading Proficiencies	
	New Procedure	Old Procedure	New Procedure	Old Procedure
Observed Race/Ethnicity				
White	77.3%(0.9)	76.8%(1.0)	260.3(0.9)	258.8(1.2)
Black	14.4%(0.8)	14.4%(0.9)	239.2(1.9)	239.3(1.6)
Hispanic	6.1%(1.0)	6.6%(1.1)	242.1(2.6)	242.2(3.1)
Other	2.2%(0.3)	2.2%(0.3)	262.3(3.6)	263.9(4.1)
Region				
Northeast	23.9%(1.6)	22.4%(1.6)	259.6(2.2)	258.7(2.1)
Southeast	23.9%(1.9)	24.7%(5.8)	254.3(1.6)	254.8(1.6)!
Central	25.6%(0.6)	24.9%(5.0)	254.6(1.3)	250.8(3.6)
West	26.7%(1.4)	28.0%(1.5)	256.1(1.8)	256.0(1.7)
Grade				
< Modal Grade	32.3%(1.6)	32.7%(2.1)	239.3(1.4)	238.4(1.4)
at Modal Grade	67.3%(1.6)	66.8%(2.1)	264.1(1.0)	263.0(0.9)
> Modal Grade	0.5%(0.1)	0.5%(0.1)	279.5(6.5)!	275.8(6.0)!
Parental Education				
Not Graduated H S	7.3%(0.5)	7.8%(1.0)	245.4(2.2)	244.2(2.9)
Graduated H S	29.6%(1.3)	30.5%(1.2)	249.8(1.2)	249.3(1.1)
Post H S	54.0%(2.0)	52.3%(2.1)	263.7(1.0)	262.7(0.9)
Total			256.2(0.8)	255.0(1.0)

Note: Standard errors in parentheses (standard errors do not include equating error)

! Interpret with caution--the sampling error cannot bc accurately estimated, since the coefficient of variation of the estimated total number of students in the subpopulation exceeds 20 percent.

Table B.6

Effect of Change in Poststratification Procedures:
Relative Frequencies and Mean Reading Proficiencies, Age 17

	Relative Frequencies		Mean Reading Proficiencies	
	New Procedure	Old Procedure	New Procedure	Old Procedure
Observed Race/Ethnicity				
White	76.6%(0.4)	78.0%(0.4)	290.9(0.9)	291.4(0.9)
Black	14.6%(0.2)	13.5%(0.2)	264.9(1.3)	265.0(1.2)
Hispanic	6.4%(0.2)	6.2%(0.2)	266.3(2.4)	267.5(2.1)
Other	2.4%(0.3)	2.4%(0.3)	274.1(4.1)	276.0(4.4)
Region				
Northeast	25.4%(1.2)	23.8%(0.3)	291.2(2.0)	293.1(2.0)
Southeast	24.0%(0.6)	21.2%(1.4)	280.0(1.0)	279.4(1.0)
Central	26.1%(0.6)	28.4%(1.5)	287.1(2.1)	288.1(2.1)
West	24.5%(0.9)	26.5%(0.5)	281.7(1.4)	282.7(1.5)
Grade				
< Modal Grade	24.9%(0.6)	21.8%(0.6)	258.0(0.9)	257.7(1.0)
at Modal Grade	65.8%(0.4)	70.3%(0.4)	293.1(0.8)	293.1(0.8)
> Modal Grade	9.3%(0.6)	7.9%(0.5)	301.2(2.0)	301.0(2.1)
Parental Education				
Not Graduated H S	9.3%(0.5)	8.9%(0.6)	265.0(1.1)	266.3(1.4)
Graduated H S	27.8%(0.9)	27.7%(0.8)	274.9(0.8)	275.9(0.8)
Post H S	58.9%(1.3)	59.4%(1.2)	295.3(0.8)	295.8(0.9)
Total			285.1(0.8)	286.0(0.9)

Note: Standard errors in parentheses (standard errors do not include equating error)

APPENDIX C

Revision of Poststratification Weights for Age 9/Grade 4 and Age 13/Grade 8, 1984 NAEP

Appendix C

REVISION OF POSTSTRATIFICATION WEIGHTS FOR AGE 9/GRADE 4 AND AGE 13/GRADE 8, 1984 NAEP

Keith F. Rust

A comparison of the proportions of 9-year-old students who were in grade 4, based on weighted data, revealed an inconsistency between the 1984 main sample results and those for bridge studies in subsequent years. In 1984, the percentage of 9-year-old students in grade 4 was 74.9. For three subsequent bridges, the percentage ranged from 62.6 to 66.1.

A consideration of the method of obtaining the separate poststratification factors for those students both grade and age eligible, those eligible by age alone, and those eligible by grade alone, used in 1984 but not for subsequent bridges, revealed the possibility of improving the approach used to derive the independent estimates which constitute the major component of the numerators of each poststratification factor. This improvement pertained to the poststratification procedure for age 9/grade 4 and age 13/grade 8, but not age 17/grade 11.

The possibility of improvement arose because the independent estimates were derived using Current Population Survey (CPS) data on the distribution over grades of the population by whole years of age. These ages are as of early October, the time each year the CPS survey in which this information is collected is conducted. The age definition for ages 9 and 13 used in 1984 means that this distribution is required as of January 1. (For age 17, and for all three ages for the main samples in 1986, the appropriate date is October 1, consistent with the CPS data.)

Evidence from the 1984 and 1988 NAEP samples shows clearly that the proportion of 9-year-olds who were in grade 4 and 13-year-olds who were in grade 8 declined between October 1 and the following January 1. That is, there were more fourth graders who had their tenth birthday during this period than there were fourth graders who had their ninth birthday. The difference was sufficiently great as to decrease the percentage of 9-year-olds who were age-eligible by about 10 percentage points. A similar but less marked decrease also occurred at age 13.

Independent estimates and the resulting poststratification factors were recomputed in a way that recognized this shift. The magnitude in the shift was estimated from NAEP data, this being the only source of information available. We note that the shift proved very consistent between the 1984 and 1988 samples, when the same age and grade definitions were used.

The 1988 poststratification procedure, which differed from that used in 1984 and 1986 in a number of ways, was performed in a manner that also accounted for this shift in the age/grade distribution. Hence, no revision of the 1988 poststratification factors is required.

APPENDIX D

Tables of Conditioning Effects and IRT Parameters for Reading, Mathematics, and Science Items

Appendix D

TABLES OF CONDITIONING EFFECTS AND IRT PARAMETERS

FOR READING, MATHEMATICS, AND SCIENCE ITEMS

Table D.1

Conditioning Effects for 1988 Reading Bridge Samples

	Age 9		Age 13		Age 17	
Conditioning	1988 Bridge		1988 Bridge		1988 Bridge	
Variable	to 1984	to 1986	to 1984	to 1986	to 1984	to 1986
1. OVERALL	-1.184954	-1.202769	0.009881	-0.001284	0.548721	0.451626
2. GENDER(F)	0.165308	0.126011	0.213211	0.195182	0.146165	0.288233
3. ETHN-BLACK	-0.438853	-0.429080	-0.259103	-0.326142	-0.272406	-0.386952
4. ETHN-HISP.	-0.412559	-0.485930	-0.274732	-0.465548	-0.345180	-0.344322
5. ETHN-ASIAN	0.357416	0.214722	0.305359	0.139659	-0.060925	-0.121725
6. HIGH METRO	-0.148497	-0.310135	-0.183941	-0.142262	-0.068281	-0.164850
7. OTHER METRO	0.147669	0.092584	0.113991	0.112262	0.152745	0.087913
8. SOUTHEAST	-0.103437	-0.132848	0.015798	-0.036619	-0.055429	-0.022728
9. CENTRAL	-0.026666	-0.093962	-0.009762	-0.061835	-0.023580	-0.041082
10. WEST	-0.154755	-0.085688	-0.037010	-0.128724	-0.078623	-0.112447
11. PAR ED1(HG)	0.291950	0.284291	0.078765	0.147679	0.273536	0.134028
12. PAR ED2(PH)	0.370200	0.477401	0.329406	0.405494	0.558189	0.443789
13. PAR ED3(COL)	0.464392	0.475387	0.324801	0.396613	0.547234	0.594699
14. PAR ED4(MIS)	0.163113	0.134590	-0.071777	0.060794	-0.106648	-0.232533
15. TV	0.235433	0.236459	-0.027696	0.086582	-0.066412	-0.014609
16 TV**2	-0.038932	-0.036129	-0.001551	-0.018358	0.000159	-0.013616

Table D.2

1986 Adjusted Reading Item Parameters, Age 9

	- A -	- B -	- C -
N001501	2.6295	-0.9528	0.3161
N001502	2.2625	-0.4435	0.1829
N001503	1.8960	-0.7052	0.2737
N001504	2.0545	-0.4687	0.2567
N002001	1.3980	-0.0249	0.1567
N002002	1.6070	-0.1793	0.2090
N002003	1.7060	-0.2887	0.2377
N002801	2.3385	-0.8188	0.1752
N002802	2.1880	-0.9522	0.1719
N003101	1.4020	-0.6171	0.2590
N003102	1.9155	-0.3275	0.2193
N003104	0.9560	2.0387	0.0000
N004101	0.9070	-1.1044	0.1996
N008601	1.9990	-0.9874	0.1979
N008602	1.7485	-0.7121	0.2457
N008603	1.7075	-0.9390	0.2074
N008901	1.4920	-0.9869	0.2490
N008902	1.5150	-1.1042	0.2452
N009401	1.4945	-1.5595	0.1307
N009801	2.1765	-2.1176	0.2378
N010201	1.4325	-1.7527	0.2057
N010301	0.8620	-2.0273	0.2212
N010401	0.7640	-1.1312	0.2283
N010402	1.2130	0.0162	0.2521
N010403	1.2350	0.4135	0.1829
N010501	2.7340	-1.2871	0.3298
N010502	1.3120	-1.0646	0.2831
N010503	1.9330	-1.2786	0.3015
N010504	2.7855	-1.0061	0.2048
N013301	1.8695	-1.8405	0.1849
N014201	1.4615	-0.8734	0.2006

Table D.3

1986 Adjusted Reading Item Parameters, Age 13

	- A -	- B -	- C -
N001501	2.3220	-1.0862	0.3161
N001502	1.9975	-0.5095	0.1829
N001503	1.6740	-0.8058	0.2737
N001504	1.8140	-0.5379	0.2567
N002001	1.2345	-0.0354	0.1567
N002002	1.4190	-0.2103	0.2090
N002003	1.5065	-0.3341	0.2377
N002801	2.0650	-0.9345	0.1752
N002802	1.9320	-1.0855	0.1719
N003001	1.1580	1.3614	0.1867
N003003	2.3540	1.3540	0.1131
N003101	1.2380	-0.7061	0.2590
N003102	1.6910	-0.3781	0.2193
N003104	0.8445	2.3017	0.0000
N004601	0.8275	-0.0688	0.1932
N004602	1.3360	-0.2043	0.2641
N004603	1.3740	-0.6796	0.2651
N005001	2.2580	1.1207	0.2340
N005002	1.1535	1.2448	0.3678
N005003	1.0380	1.5894	0.1366
N008201	3.7805	-0.4031	0.2773
N008202	1.1140	-0.1301	0.2268
N008203	1.6225	-0.3674	0.3021
N008204	2.5680	-0.2234	0.1897
N008205	3.0465	-0.1629	0.2711

Table D.4

1986 Adjusted Reading Item Parameters, Age 17

	- A -	- B -	- C -
N001501	2.5745	-0.6543	0.3161
N001502	2.2150	-0.1341	0.1829
N001503	1.8560	-0.4013	0.2737
N001504	2.0115	-0.1597	0.2567
N002001	1.3685	0.2935	0.1567
N002002	1.5730	0.1358	0.2090
N002003	1.6705	0.0241	0.2377
N002801	2.2895	-0.5174	0.1752
N002802	2.1420	-0.6536	0.1719
N003001	1.2840	1.5533	0.1867
N003003	2.6100	1.5466	0.1131
N003101	1.3730	-0.3114	0.2590
N003102	1.8750	-0.0155	0.2193
N003104	0.9360	2.4014	0.0000
N003201	1.5410	-0.1243	0.2674
N003202	1.7470	0.3168	0.2264
N003203	1.6815	0.5189	0.2064
N003204	1.8565	0.3744	0.2069
N004601	0.9175	0.2634	0.1932
N004602	1.4815	0.1412	0.2641
N004603	1.5235	-0.2875	0.2651
N005001	2.5035	1.3362	0.2340
N005002	1.2790	1.4482	0.3678
N005003	1.1510	1.7589	0.1366
N007301	1.0080	0.2102	0.2330
N007302	1.0670	0.5985	0.2181
N007303	1.4635	0.2393	0.1681
N007304	1.1395	0.3353	0.2271
N007305	0.8115	0.6117	0.2020
N007306	1.4375	0.2348	0.1601
N007401	1.3800	0.3170	0.1854
N007402	1.2790	-0.3754	0.2088
N007403	1.8645	0.2807	0.1955
N007404	1.1090	0.2393	0.2282
N007405	1.1575	1.2632	0.2522
N008201	4.1920	-0.0381	0.2773
N008202	1.2350	0.2082	0.2268
N008203	1.7990	-0.0059	0.3021
N008204	2.8470	0.1239	0.1897
N008205	3.3780	0.1785	0.2711
N013401	1.3085	0.1789	0.1540
N013402	1.7845	0.0858	0.2624
N013403	2.1115	0.3466	0.2168
N021301	1.1945	0.1155	0.0000

(continued)

Table D.4 (continued)

1986 Adjusted Reading Item Parameters, Age 17

	- A -	- B -	- C -
N021303	1.0995	-0.4062	0.1905
N021304	0.4930	0.6265	0.1725
N021305	1.0815	0.4609	0.1890
N021201	0.9520	0.1586	0.1799
N021202	0.6940	0.1285	0.1946
N021203	0.7785	0.3696	0.2039
N021204	0.8030	0.0307	0.1901
N021601	0.6850	0.0290	0.2516
N021602	0.8675	1.0329	0.1544
N021603	0.4065	1.3870	0.2163
N021604	1.5265	0.5088	0.1589
N021605	0.8780	1.2430	0.3888
N021701	1.2380	-0.1516	0.2287
N021702	1.0115	1.1863	0.1079
N021703	1.4940	1.3292	0.2894
N021801	1.3600	0.2600	0.0000
N021803	1.3090	0.5337	0.2914
N021805	1.0665	0.0178	0.0000

Table D.5

Conditioning Effects for 1986 Reading
with Adjusted Item Parameters, Age 9

Variable	Estimated Effect	Description
1 OVERALL	-0.449782	1 OVERALL CONSTANT '1' FOR EVERYONE
2 GENDER2	0.148332	2 SEX (FEMALE)
3 ETHNIC2	-0.057906	3 ETHNICITY (BLACK)
4 ETHNIC3	-0.224260	4 ETHNICITY (HISPANIC)
5 ETHNIC4	-0.027006	5 ETHNICITY (ASIAN)
6 STOC2	0.092196	6 SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7 STOC3	0.149317	7 SIZE AND TYPE OF COMMUNITY (NOT HIGH OR LOW)
8 REGION2	-0.027025	8 REGION (SOUTHEAST)
9 REGION3	0.037337	9 REGION (CENTRAL)
10 REGION4	0.030380	10 REGION (WEST)
11 PARED2	0.058072	11 PARENTS EDUCATION (HIGH SCHOOL GRAD)
12 PARED3	0.238289	12 PARENTS EDUCATION (POST HIGH SCHOOL)
13 PARED4	0.210194	13 PARENTS EDUCATION (COLLEGE GRAD)
14 PARED_	0.130707	14 PARENTS EDUCATION (MISSING, I DON'T KNOW)
15 ITEMS2	0.020836	15 ITEMS IN HOME (FOUR OF THE FIVE)
16 ITEMS3	0.045386	16 ITEMS IN HOME (FIVE OF THE FIVE)
17 TV	0.077068	17 HOURS TV WATCHING (LINEAR)
18 TV**2	-0.015100	18 HOURS TV WATCHING (QUADRATIC)
19 HW-YES	-0.253901	19 HOMEWORK (DON'T HAVE ANY & SOME AMOUNT)
20 HW-2345	-0.013257	20 HOMEWORK AMOUNT (LINEAR)
21 LM BY E3	0.044572	21 LANGUAGE MINORITY BY ETHNICITY (YES, HISPANIC)
22 LM BY E4	-0.120663	22 LANGUAGE MINORITY BY ETHNICITY (YES, ASIAN)
23 LM BY E_	-0.074779	23 LANGUAGE MINORITY BY ETHNICITY (YES, OTHER ETH)
24 LUNCH%	-0.040061	24 PERCENT IN LUNCH PROGRAM (F3.2)
25 LUNCH_	-0.046324	25 LUNCH PROGRAM (MISSING)
26 %WHITE49	-0.148280	26 PERCENT WHITE IN SCHOOL (0-49% WHITE MINORITY)
27 %WHITE79	-0.067982	27 PERCENT WHITE IN SCHOOL (50-79% INTEGRATED)
28 E2 X SEX	0.134723	29 ETHNICITY BY GENDER (BLACK FEMALE)
29 E3 X SEX	0.087811	30 ETHNICITY BY GENDER (HISPANIC FEMALE)
30 E4 X SEX	-0.001002	31 ETHNICITY BY GENDER (ASIAN FEMALE)
31 E2 X PE2	-0.179641	32 ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
32 E2 X PE3	-0.224820	33 ETHNICITY BY PARENT'S ED (BLACK, POST HS)
33 E2 X PE4	-0.098036	34 ETHNICITY BY PARENT'S ED (BLACK, COLLEGE GRAD)
34 E2 X PE_	-0.159220	35 ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
35 E3 X PE2	0.030622	36 ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
36 E3 X PE3	-0.170710	37 ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
37 E3 X PE4	-0.111656	38 ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
38 E3 X PE_	0.058495	39 ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
39 E4 X PE3	0.365161	41 ETHNICITY BY PARENT'S ED (ASIAN, POST HS)
40 E4 X PE4	0.259550	42 ETHNICITY BY PARENT'S ED (ASIAN, COLLEGE GRAD)
41 E4 X PE_	0.366435	43 ETHNICITY BY PARENT'S ED (ASIAN, UNKNOWN)
42 <MA,<MG	-0.682847	44 MODAL AGE, LESS THAN MODAL GRADE
43 MA,MG	-0.420010	45 MODAL AGE, MODAL GRADE, MISSING

(continued)

Table D.5 (continued)

Conditioning Effects for 1986 Reading
with Adjusted Item Parameters, Age 9

Variable	Estimated Effect	Description
44 SCH TYPE	0.073026	48 SCHOOL TYPE (NOT PUBLIC)
45 ASK SW?	0.056887	49 FAMILY ASKS ABOUT SCHOOLWORK (ALMOST EVERY DAY)
46 PRESCH1	0.071246	50 WENT TO PRESCHOOL (YES)
47 #PARENT1	0.102494	51 SINGLE/MULTIPLE PARENT HOME (MOTHER,FATHER HOME)
48 MOTHER	0.011320	52 MOTHER AT HOME (WORKING AND NON-WORKING)
49 MOWORK	0.009405	53 MOTHER WORKS OUTSIDE HOME (YES)
50 SCIEN123	-0.167822	54 TIME SPENT IN SCIENCE(AT LEAST ONCE A WEEK)
51 SCIEN45	-0.171251	55 TIME SPENT IN SCIENCE(<ONCE A WEEK OR NEVER)
52 COMPUTER	0.022781	56 USE COMPUTERS FOR MATH, READING, ETC. (YES)
53 SUPERVIS	0.062169	57 ADULT SUPERVISION OF STUDENT AFTER SCHOOL(YES)
54 MATH Q1	-0.298827	58 MATH 1ST QUANTILE (LINEAR -1,0,1)
55 SCI Q1	-0.229154	59 SCIENCE 1ST QUANTILE (LINEAR -1,0,1)

Table D.6

Conditioning Effects for 1986 Reading with Adjusted Item Parameters, Age 13

Variable	Estimated Effect	Description
1 OVERALL	-1.178359	1 OVERALL CONSTANT '1' FOR EVERYONE
2 GENDER2	0.150605	2 SEX (FEMALE)
3 ETHNIC2	-0.051714	3 ETHNICITY (BLACK)
4 ETHNIC3	-0.062785	4 ETHNICITY (HISPANIC)
5 ETHNIC4	0.306524	5 ETHNICITY (ASIAN)
6 STOC2	0.088873	6 SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7 STOC3	0.022681	7 SIZE AND TYPE OF COMMUNITY (NOT HIGH OR LOW)
8 REGION2	0.130568	8 REGION (SOUTHEAST)
9 REGION3	0.016751	9 REGION (CENTRAL)
10 REGION4	0.007241	10 REGION (WEST)
11 PARED2	-0.085747	11 PARENTS EDUCATION (HIGH SCHOOL GRAD)
12 PARED3	-0.034406	12 PARENTS EDUCATION (POST HIGH SCHOOL)
13 PARED4	-0.073942	13 PARENTS EDUCATION (COLLEGE GRAD)
14 PARED_	-0.086500	14 PARENTS EDUCATION (MISSING, I DON'T KNOW)
15 ITEMS2	0.091700	15 ITEMS IN HOME (FOUR OF THE FIVE)
16 ITEMS3	0.088342	16 ITEMS IN HOME (FIVE OF THE FIVE)
17 TV	0.063395	17 HOURS TV WATCHING (LINEAR)
18 TV**2	-0.009358	18 HOURS TV WATCHING (QUADRATIC)
19 HW-NO	0.245607	19 HOMEWORK (DON'T HAVE ANY)
20 HW-YES	0.189577	20 HOMEWORK (YES - SOME AMOUNT)
21 HW-3456	0.010437	21 HOMEWORK AMOUNT (LINEAR)
22 LM BY E3	0.102256	22 LANGUAGE MINORITY BY ETHNICITY (YES, HISPANIC)
23 LM BY E4	0.056094	23 LANGUAGE MINORITY BY ETHNICITY (YES, ASIAN)
24 LM BY E_	-0.061011	24 LANGUAGE MINORITY BY ETHNICITY (YES, OTHER ETH)
25 LUNCH%	0.019951	25 PERCENT IN LUNCH PROGRAM (F3.2)
26 LUNCH_	-0.055597	26 LUNCH PROGRAM (MISSING)
27 %WHITE49	0.032951	27 PERCENT WHITE IN SCHOOL (0-49% WHITE MINORITY)
28 %WHITE79	0.029780	28 PERCENT WHITE IN SCHOOL (50-79% INTEGRATED)
29 E2 X SEX	-0.044162	30 ETHNICITY BY GENDER (BLACK FEMALE)
30 E3 X SEX	0.016185	31 ETHNICITY BY GENDER (HISPANIC FEMALE)
31 E4 X SEX	0.143079	32 ETHNICITY BY GENDER (ASIAN FEMALE)
32 E2 X PE2	0.032872	33 ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
33 E2 X PE3	0.041641	34 ETHNICITY BY PARENT'S ED (BLACK, POST HS)
34 E2 X PE4	0.026823	35 ETHNICITY BY PARENT'S ED (BLACK, COLLEGE GRAD)
35 E2 X PE_	0.096667	36 ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
36 E3 X PE2	-0.114480	37 ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
37 E3 X PE3	-0.161151	38 ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
38 E3 X PE4	-0.142465	39 ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
39 E3 X PE_	-0.038051	40 ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
40 E4 X PE2	-0.618719	41 ETHNICITY BY PARENT'S ED (ASIAN, HS GRAD)
41 E4 X PE3	-0.502270	42 ETHNICITY BY PARENT'S ED (ASIAN, POST HS)
42 E4 X PE4	-0.417888	43 ETHNICITY BY PARENT'S ED (ASIAN, COLLEGE GRAD)

(continued)

Table D.6 (continued)

Conditioning Effects for 1986 Reading
with Adjusted Item Parameters, Age 13

Variable	Estimated Effect	Description
43 E4 X PE_	-0.193980	44 ETHNICITY BY PARENT'S ED (ASIAN, UNKNOWN)
44 <MA,<MG	-0.235984	45 MODAL AGE, LESS THAN MODAL GRADE
45 MA,MG	-0.134553	46 MODAL AGE, MODAL GRADE, MISSING
46 SCH TYPE	0.096220	49 SCHOOL TYPE (NOT PUBLIC)
47 ASK SW?	0.046783	50 FAMILY ASKS ABOUT SCHOOLWORK (ALMOST EVERY DAY)
48 PRESCH1	0.067139	51 WENT TO PRESCHOOL (YES)
49 #PARENT1	-0.028203	52 SINGLE/MULTIPLE PARENT HOME (MOTHER,FATHER HOME)
50 MOTHER	0.070043	53 MOTHER AT HOME (WORKING AND NON-WORKING)
51 MOWORK	-0.012856	54 MOTHER WORKS OUTSIDE HOME (YES)
52 COMPUTER	-0.087101	55 USE COMPUTERS FOR MATH, READING, ETC. (YES)
53 MATH2	0.232694	56 TYPE OF MATH CLASS (REGULAR MATH)
54 MATH3	0.259156	57 TYPE OF MATH CLASS (PRE-ALGEBRA)
55 MATH45	0.312297	58 TYPE OF MATH CLASS (ALGEBRA, OTHER)
56 SCIENCE2	0.034047	59 STUDYING IN SCIENCE THIS YEAR (LIFE SCIENCE)
57 SCIENCE3	0.077382	60 STUDYING IN SCIENCE THIS YEAR (PHYSICAL SCIENCE)
58 SCIENCE4	0.092771	61 STUDYING IN SCIENCE THIS YEAR (EARTH SCIENCE)
59 SCIENCE5	0.095069	62 STUDYING IN SCIENCE THIS YEAR (GENERAL SCIENCE)
60 GRADES	0.163220	63 GRADES IN SCHOOL (LINEAR)
61 MATH Q1	-0.192220	64 MATH 1ST QUANTILE (LINEAR -1,0,1)
62 SCI Q1	-0.253671	65 SCIENCE 1ST QUANTILE (LINEAR -1,0,1)

Table D.7

Conditioning Effects for 1986 Reading
with Adjusted Item Parameters, Age 17

Variable	Estimated Effect	Description
1 OVERALL	-0.094618	1 OVERALL CONSTANT '1' FOR EVERYONE
2 GENDER2	0.183789	2 SEX (FEMALE)
3 ETHNIC2	-0.152881	3 ETHNICITY (BLACK)
4 ETHNIC3	-0.192689	4 ETHNICITY (HISPANIC)
5 ETHNIC4	-0.267717	5 ETHNICITY (ASIAN)
6 STOC2	0.128344	6 SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7 STOC3	0.062382	7 SIZE AND TYPE OF COMMUNITY (NOT HIGH OR LOW)
8 REGION2	-0.013891	8 REGION (SOUTHEAST)
9 REGION3	0.031733	9 REGION (CENTRAL)
10 REGION4	-0.029610	10 REGION (WEST)
11 PARED2	-0.036836	11 PARENTS EDUCATION (HIGH SCHOOL GRAD)
12 PARED3	0,048126	12 PARENTS EDUCATION (POST HIGH SCHOOL)
13 PARED4	0.056826	13 PARENTS EDUCATION (COLLEGE GRAD)
14 PARED_	-0.185737	14 PARENTS EDUCATION (MISSING, I DON'T KNOW)
15 ITEMS2	0.086351	15 ITEMS IN HOME (FOUR OF THE FIVE)
16 ITEMS3	0.116690	16 ITEMS IN HOME (FIVE OF THE FIVE)
17 TV	0.018872	17 HOURS TV WATCHING (LINEAR)
18 TV**2	-0.006411	18 HOURS TV WATCHING (QUADRATIC) (F2.0)
19 HW-NO	-0.285550	19 HOMEWORK (DON'T HAVE ANY)
20 HW-YES	-0.141044	20 HOMEWORK (YES - SOME AMOUNT)
21 HW-3456	-0.000823	21 HOMEWORK AMOUNT (LINEAR)
22 LM BY E3	-0.056669	22 LANGUAGE MINORITY BY ETHNICITY (YES, HISPANIC)
23 LM BY E4	-0.209130	23 LANGUAGE MINORITY BY ETHNICITY (YES, ASIAN)
24 LM BY E_	0.012988	24 LANGUAGE MINORITY BY ETHNICITY (YES, OTHER ETH)
25 LUNCH%	-0.120100	25 PERCENT IN LUNCH PROGRAM (F3.2)
26 LUNCH_	-0.019393	26 LUNCH PROGRAM (MISSING)
27 %WHITE49	0.008171	27 PERCENT WHITE IN SCHOOL (0-49% WHITE MINORITY)
28 %WHITE79	0.033535	28 PERCENT WHITE IN SCHOOL (50-79% INTEGRATED)
29 E2 X SEX	-0.125090	30 ETHNICITY BY GENDER (BLACK FEMALE)
30 E3 X SEX	-0.007812	31 ETHNICITY BY GENDER (HISPANIC FEMALE)
31 E4 X SEX	0.045293	32 ETHNICITY BY GENDER (ASIAN FEMALE)
32 E2 X PE2	-0.011828	33 ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
33 E2 X PE3	0.063253	34 ETHNICITY BY PARENT'S ED (BLACK, POST HS)
34 E2 X PE4	-0.036463	35 ETHNICITY BY PARENT'S ED (BLACK, COLLEGE GRAD)
35 E2 X PE_	0.108169	36 ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
36 E3 X PE2	0.024990	37 ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
37 E3 X PE3	0 074898	38 ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
38 E3 X PE4	0.060779	39 ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
39 E3 X PE_	0.165982	40 ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
40 E4 X PE2	0.076386	41 ETHNICITY BY PARENT'S ED (ASIAN, HS GRAD)
41 E4 X PE3	0.251174	42 ETHNICITY BY PARENT'S ED (ASIAN, POST HS)
42 E4 X PE4	0.181287	43 ETHNICITY BY PARENT'S ED (ASIAN, COLLEGE GRAD)

(continued)

Table D.7 (continued)

Conditioning Effects for 1986 Reading
with Adjusted Item Parameters, Age 17

Variable	Estimated Effect	Description
43 E4 X PE_	0.276123	44 ETHNICITY BY PARENT'S ED (ASIAN, UNKNOWN)
44 <MA,<MG	-0.281517	45 MODAL AGE, LESS THAN MODAL GRADE
45 MA,MG	-0.053717	46 MODAL AGE, MODAL GRADE, MISSING
46 MA,>MG	0.001677	47 MODAL AGE, GREATER THAN MODAL GRADE
47 >MA,MG	-0.213755	48 GREATER THAN MODAL AGE, MODAL GRADE
48 SCH TYPE	0.064388	49 SCHOOL TYPE (NOT PUBLIC)
49 ASK SW?	-0.034331	50 FAMILY ASKS ABOUT SCHOOLWORK (ALMOST EVERY DAY)
50 PRESCH1	0.003179	51 WENT TO PRESCHOOL (YES)
51 #PARENT1	0.007493	52 SINGLE/MULTIPLE PARENT HOME (MOTHER,FATHER HOME)
52 MOTHER	-0.027155	53 MOTHER AT HOME (WORKING AND NON-WORKING)
53 MOWORK	-0.002359	54 MOTHER WORKS OUTSIDE HOME (YES)
54 GRADES	0.175612	55 GRADES IN SCHOOL (LINEAR) (F3.1)
55 HS PGM2	0.100833	56 HIGH SCHOOL PROGRAM(COLLEGE PREPARATORY)
56 HS PGM3	-0.031808	57 HIGH SCHOOL PROGRAM(VOCATIONAL, TECHNICAL)
57 NO. MATH	0.061355	58 NO. OF MATH COURSES
58 NO. SCI	0.052572	59 NO. OF SCIENCE COURSES
59 POSTSEC2	0.055939	60 POST-SECONDARY PLANS(TWO YEAR COLLEGE)
60 POSTSEC3	0.147077	61 POST-SECONDARY PLANS(FOUR YEAR COLLEGE)
61 WORKHOUR	-0.041279	62 HOURS OF OUTSIDE WORK
62 ENG.23	0.096967	63 TYPE OF ENGLISH CLASS(ADVANCED PLACEMENT&COLLEGE
63 ENGLISH5	-0.151264	64 TYPE OF ENGLISH CLASS(REMEDIAL)
64 MATH Q1	-0.152759	65 MATH 1ST QUANTILE (LINEAR -1,0,1) (F2.0)
65 SCI Q1	-0.251787	66 SCIENCE 1ST QUANTILE (LINEAR -1,0,1) (F2.0)

Table D.8

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 9

Field	Block	Item	A	SE	B	SE	C	SE
N270901	M1	1	0.894	(0.037)	-2.165	(0.098)	0.000	(0.000)
N277401	M1	2	1.026	(0.063)	-1.573	(0.114)	0.177	(0.038)
N267601	M1	3	1.268	(0.066)	-0.611	(0.049)	0.156	(0.020)
N276801	M1	4	0.490	(0.045)	-3.763	(0.353)	0.000	(0.000)
N276802	M1	5	0.725	(0.038)	-1.591	(0.090)	0.000	(0.000)
N276803	M1	6	0.621	(0.035)	0.147	(0.027)	0.000	(0.000)
N250701	M1	7	0.743	(0.044)	-0.850	(0.059)	0.139	(0.022)
N250702	M1	8	1.001	(0.048)	0.841	(0.054)	0.117	(0.011)
N250703	M1	9	1.054	(0.064)	0.015	(0.033)	0.123	(0.016)
N262201	M1	10	0.441	(0.036)	-1.218	(0.105)	0.196	(0.024)
N257201	M1	11	1.233	(0.084)	-0.533	(0.055)	0.283	(0.020)
N276101	M1	12	0.963	(0.040)	-0.758	(0.042)	0.000	(0.000)
N286101	M1	13	0.814	(0.039)	-0.521	(0.035)	0.000	(0.000)
N270001	M1	14	0.448	(0.030)	-0.727	(0.053)	0.000	(0.000)
N272102	M1	15	0.992	(0.062)	0.034	(0.039)	0.173	(0.018)
N284001	M1	16	0.981	(0.050)	-0.383	(0.033)	0.000	(0.000)
N284002	M1	17	0.792	(0.037)	2.054	(0.103)	0.000	(0.000)
N267602	M1	18	1.103	(0.057)	-0.074	(0.031)	0.104	(0.014)
N262501	M1	19	0.269	(0.031)	-0.688	(0.084)	0.227	(0.019)
N262502	M1	20	0.254	(0.062)	6.169	(1.519)	0.172	(0.008)
N265401	M1	21	1.582	(0.164)	2.224	(0.360)	0.340	(0.011)
N266101	M1	22	0.542	(0.052)	1.917	(0.192)	0.264	(0.011)
N269101	M1	23	0.540	(0.071)	2.970	(0.402)	0.238	(0.009)
N268201	M1	24	1.248	(0.058)	1.026	(0.068)	0.201	(0.010)
N252101	M1	25	0.839	(0.060)	1.752	(0.143)	0.170	(0.012)
N272301	M2	1	0.946	(0.052)	-1.947	(0.123)	0.180	(0.040)
N276601	M2	2	1.061	(0.062)	-1.010	(0.076)	0.170	(0.029)
N257801	M2	3	0.588	(0.038)	-0.909	(0.066)	0.240	(0.022)
N263401	M2	4	0.888	(0.063)	-0.701	(0.063)	0.299	(0.022)
N263402	M2	5	1.010	(0.080)	-0.203	(0.043)	0.282	(0.018)
N273501	M2	6	0.744	(0.058)	-0.684	(0.068)	0.261	(0.026)
N275401	M2	7	0.985	(0.043)	-0.478	(0.033)	0.000	(0.000)
N277501	M2	8	0.842	(0.039)	-0.421	(0.031)	0.000	(0.000)
N277601	M2	9	1.438	(0.049)	-0.522	(0.037)	0.000	(0.000)
N277602	M2	10	1.267	(0.053)	0.172	(0.029)	0.000	(0.000)
N277603	M2	11	1.507	(0.063)	-0.011	(0.030)	0.000	(0.000)
N261401	M2	12	0.509	(0.042)	-0.145	(0.037)	0.232	(0.020)
N250601	M2	13	1.097	(0.078)	-0.231	(0.045)	0.212	(0.019)
N250602	M2	14	0.791	(0.053)	-0.584	(0.054)	0.189	(0.023)
N250603	M2	15	1.366	(0.071)	0.566	(0.056)	0.158	(0.013)
N251401	M2	16	0.654	(0.042)	-0.265	(0.038)	0.151	(0.021)
N250901	M2	17	0.599	(0.040)	-0.411	(0.040)	0.178	(0.019)
N250902	M2	18	1.101	(0.051)	1.181	(0.072)	0.157	(0.010)
N250903	M2	19	0.970	(0.051)	0.685	(0.050)	0.109	(0.012)

(continued)

Table D.8 (continued)

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 9

Field	Block	Item	A	SE	B	SE	C	SE
N276001	M2	21	0.879	(0.037)	-0.975	(0.049)	0.000	(0.000)
N276002	M2	22	0.778	(0.035)	1.507	(0.074)	0.000	(0.000)
N271101	M2	24	0.626	(0.034)	-0.305	(0.028)	0.000	(0.000)
N252001	M2	25	1.244	(0.131)	2.670	(0.372)	0.196	(0.009)
N269001	M2	26	0.565	(0.087)	4.055	(0.634)	0.082	(0.007)
N272801	M3	15	0.576	(0.049)	-2.007	(0.176)	0.180	(0.036)
N267001	M3	16	0.597	(0.045)	-1.392	(0.110)	0.249	(0.026)
N272101	M3	17	0.990	(0.096)	-0.533	(0.071)	0.286	(0.024)
N262401	M3	18	0.594	(0.069)	0.928	(0.116)	0.300	(0.013)
N258501	M3	19	0.876	(0.066)	1.029	(0.092)	0.236	(0.012)

Table D.9

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 13

Field	Block	Item	A	SE	B	SE	C	SE
N281901	M1	15	0.925	(0.040)	-2.181	(0.105)	0.146	(0.034)
N254601	M1	16	1.092	(0.054)	-1.553	(0.089)	0.284	(0.030)
N276801	M1	17	0.433	(0.049)	-4.715	(0.542)	0.000	(0.000)
N276802	M1	18	0.493	(0.044)	-3.957	(0.359)	0.000	(0.000)
N276803	M1	19	0.435	(0.033)	-1.927	(0.148)	0.000	(0.000)
N277601	M1	20	0.856	(0.036)	-2.504	(0.113)	0.000	(0.000)
N277602	M1	21	0.624	(0.030)	-1.885	(0.095)	0.000	(0.000)
N277603	M1	22	0.617	(0.031)	-2.287	(0.117)	0.000	(0.000)
N267201	M1	23	0.776	(0.058)	-1.051	(0.087)	0.254	(0.026)
N286201	M1	24	0.891	(0.051)	-0.892	(0.061)	0.243	(0.021)
N250901	M1	25	0.423	(0.029)	-2.565	(0.176)	0.152	(0.027)
N250902	M1	26	1.020	(0.049)	-0.349	(0.031)	0.075	(0.014)
N250903	M1	27	0.820	(0.039)	-1.510	(0.078)	0.096	(0.025)
N262401	M1	28	0.854	(0.054)	-0.556	(0.048)	0.323	(0.017)
N274801	M1	29	0.629	(0.051)	-0.192	(0.036)	0.269	(0.018)
N265202	M1	30	0.843	(0.074)	-0.176	(0.041)	0.339	(0.018)
N266801	M1	31	0.559	(0.038)	-1.108	(0.080)	0.248	(0.021)
N252901	M1	32	1.249	(0.072)	-0.036	(0.033)	0.109	(0.015)
N262501	M1	33	0.360	(0.033)	-0.237	(0.034)	0.348	(0.015)
N262502	M1	34	1.216	(0.068)	1.974	(0.151)	0.379	(0.008)
N257601	M1	35	1.280	(0.055)	-0.538	(0.035)	0.000	(0.000)
N265201	M1	36	0.810	(0.062)	-1.548	(0.127)	0.339	(0.032)
N273901	M1	37	1.786	(0.111)	0.258	(0.047)	0.184	(0.013)
N258801	M1	38	1.273	(0.055)	1.124	(0.076)	0.397	(0.010)
N263101	M1	39	0.527	(0.027)	-0.291	(0.024)	0.000	(0.000)
N265901	M1	40	0.933	(0.060)	0.930	(0.079)	0.333	(0.012)
N252101	M1	41	0.933	(0.056)	0.623	(0.054)	0.240	(0.013)
N275001	M1	42	0.946	(0.040)	0.363	(0.027)	0.000	(0.000)
N260101	M1	43	1.299	(0.072)	0.415	(0.042)	0.160	(0.011)
N269001	M1	44	1.012	(0.053)	0.382	(0.036)	0.152	(0.011)
N286301	M1	45	1.189	(0.050)	0.660	(0.046)	0.205	(0.010)
N254602	M1	46	0.744	(0.045)	1.413	(0.095)	0.235	(0.009)
N261001	M1	47	0.833	(0.049)	1.011	(0.070)	0.219	(0.010)
N286501	M1	48	1.256	(0.042)	1.161	(0.058)	0.141	(0.008)
N278904	M1	49	1.315	(0.057)	1.487	(0.097)	0.194	(0.010)
N255701	M1	50	1.317	(0.044)	1.268	(0.063)	0.139	(0.008)
N283101	M1	51	1.579	(0.049)	1.554	(0.080)	0.148	(0.006)
N277401	M2	8	0.778	(0.056)	-2.903	(0.220)	0.145	(0.042)
N277901	M2	9	0.591	(0.033)	-3.506	(0.199)	0.000	(0.000)
N277902	M2	10	0.688	(0.036)	-3.301	(0.178)	0.000	(0.000)
N277903	M2	11	0.573	(0.030)	-2.859	(0.154)	0.000	(0.000)
N263401	M2	12	0.675	(0.046)	-2.751	(0.196)	0.257	(0.040)
N263402	M2	13	0.635	(0.045)	-2.478	(0.181)	0.263	(0.036)
N250701	M2	14	0.688	(0.035)	-2.717	(0.143)	0.106	(0.033)

(continued)

Table D.9 (continued)

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 13

Field	Block	Item	A	SE	B	SE	C	SE
N250702	M2	15	1.145	(0.051)	-0.797	(0.047)	0.102	(0.018)
N250703	M2	16	0.649	(0.031)	-2.110	(0.106)	0.110	(0.028)
N256101	M2	17	0.760	(0.033)	-1.056	(0.052)	0.000	(0.000)
N262201	M2	18	0.520	(0.037)	-1.789	(0.132)	0.361	(0.023)
N270301	M2	20	0.421	(0.031)	-1.596	(0.119)	0.126	(0.022)
N270302	M2	21	1.018	(0.047)	2.194	(0.118)	0.051	(0.005)
N253701	M2	22	0.361	(0.031)	-0.504	(0.050)	0.271	(0.016)
N286601	M2	23	1.698	(0.059)	-0.194	(0.029)	0.000	(0.000)
N286602	M2	24	1.363	(0.051)	-0.247	(0.027)	0.000	(0.000)
N286603	M2	25	1.494	(0.050)	0.405	(0.030)	0.000	(0.000)
N269101	M2	26	0.752	(0.048)	-0.384	(0.037)	0.213	(0.016)
N282201	M2	28	1.063	(0.058)	0.576	(0.051)	0.343	(0.011)
N278902	M2	29	0.720	(0.051)	1.338	(0.107)	0.216	(0.012)
N263501	M2	30	1.389	(0.092)	0.187	(0.036)	0.115	(0.012)
N258802	M2	31	1.619	(0.078)	0.484	(0.051)	0.254	(0.011)
N278901	M2	32	1.559	(0.086)	0.415	(0.051)	0.212	(0.013)
N264701	M2	33	1.175	(0.056)	0.867	(0.059)	0.206	(0.010)
N261501	M2	34	0.661	(0.056)	-0.545	(0.055)	0.141	(0.020)
N261801	M2	35	0.679	(0.053)	0.044	(0.033)	0.223	(0.017)
N261601	M2	36	0.344	(0.043)	1.903	(0.239)	0.155	(0.012)
N261301	M2	37	0.700	(0.048)	0.768	(0.062)	0.113	(0.012)
N261201	M2	38	0.525	(0.052)	1.619	(0.166)	0.219	(0.012)
N281401	M2	39	0.728	(0.050)	1.711	(0.127)	0.106	(0.009)
N252001	M2	40	1.423	(0.064)	0.832	(0.062)	0.179	(0.010)
N258803	M2	41	1.191	(0.044)	1.351	(0.068)	0.170	(0.007)
N278903	M2	42	1.338	(0.058)	1.066	(0.073)	0.169	(0.010)
N286502	M2	43	1.671	(0.054)	1.171	(0.068)	0.160	(0.008)
N275301	M3	25	0.372	(0.028)	-1.728	(0.132)	0.147	(0.022)
N282202	M3	26	0.936	(0.066)	-0.458	(0.045)	0.255	(0.017)
N266101	M3	27	0.849	(0.065)	-0.161	(0.033)	0.292	(0.014)
N254001	M3	28	1.161	(0.084)	-0.479	(0.047)	0.118	(0.017)
N269901	M3	29	0.664	(0.049)	-0.274	(0.035)	0.288	(0.015)
N256501	M3	30	0.866	(0.069)	0.581	(0.061)	0.318	(0.012)
N265902	M3	31	1.077	(0.073)	1.170	(0.103)	0.328	(0.011)
N256801	M3	32	1.051	(0.069)	0.841	(0.072)	0.312	(0.011)

Table D.10

NAEP 1988 Mathematics Trend Conditioning Variables, Age 9

	Variable	Estimated Effect	Description
1	OVERALL	-0.279547	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.047747	GENDER (FEMALE)
3	ETHNIC2	-0.706632	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	0.209298	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	0.762678	OBSERVED ETHNICITY (ASIAN)
6	STOC3	0.186615	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7	STOC1	0.087756	SIZE AND TYPE OF COMMUNITY (NOT HI&NOT LO)
8	REGION2	0.007280	REGION (SOUTHEAST)
9	REGION3	0.123942	REGION (CENTRAL)
10	REGION4	-0.035032	REGION (WEST)
11	PARED2	0.251057	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.223869	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.454556	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	0.136615	PARENTS EDUCATION (MISSING, I DON'T KNOW)
15	<MODAL GRADE	-0.728308	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	0.631198	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.239816	ITEMS IN THE HOME (YES TO 3)
18	ITEMS3	0.367498	ITEMS IN THE HOME (YES TO ALL 4)
19	E2 X SEX	0.087308	ETHNICITY BY GENDER (BLACK, FEMALE)
20	E3 X SEX	-0.066049	ETHNICITY BY GENDER (HISPANIC, FEMALE)
21	E4 X SEX	-0.231095	ETHNICITY BY GENDER (ASIAN AMERICAN, FEMALE)
22	E2 X PE2	-0.063586	ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
23	E2 X PE3	0.375105	ETHNICITY BY PARENT'S ED (BLACK, POST HS)
24	E2 X PE4	0.039552	ETHNICITY BY PARENT'S ED (BLACK, COLLEGE)
25	E2 X PE_	0.191412	ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
26	E3 X PE2	-0.354255	ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
27	E3 X PE3	0.237226	ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
28	E3 X PE4	-0.256883	ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
29	E3 X PE_	-0.246003	ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
30	E4 X PE2	-1.034833	ETHNICITY BY PARENT'S ED (ASIAN AM, HS GRAD)
31	E4 X PE3	-0.690193	ETHNICITY BY PARENT'S ED (ASIAN AM, POST HS)
32	E4 X PE4	-0.786758	ETHNICITY BY PARENT'S ED (ASIAN AM, COLLEGE)
33	E4 X PE_	-0.518339	ETHNICITY BY PARENT'S ED (ASIAN AM, UNKNOWN)
34	SCH TYP2	0.158816	SCHOOL TYPE (NOT PUBLIC)
35	SCH TYP_		SCHOOL TYPE (MISSING)
36	TV1	0.278883	0-2 HOURS OF TV WATCHING
37	TV2	0.434684	3-5 HOURS OF TV WATCHING
38	TV3	0.259356	6+ HOURS OF TV WATCHING
39	LANGHOM3	-0.283533	LANGUAGE IN HOME OTHER THAN ENGLISH? (ALWAYS)
40	LANGHOM2	0.088718	LANGUAGE IN HOME OTHER THAN ENGLISH?SOMETIMES
41	E2 X LH1	0.143997	ETHNICITY BY LANGUAGE IN HOME (BLACK, OFTEN)
42	E2 X LH2	0.080093	ETHNICITY BY LANG IN HOME (BLACK, SOMETIMES)
43	E3 X LH1	0.390581	ETHNICITY BY LANGUAGE IN HOME(HISPANIC,OFTEN)
44	E3 X LH2	-0.117348	ETHNICITY BY LANG IN HOME(HISPANIC,SOMETIMES)

(continued)

Table D.10 (continued)

NAEP 1988 Mathematics Trend Conditioning Variables, Age 9

	Variable	Estimated Effect	Description
45	E4 X LH1	0.411867	ETHNICITY BY LANGUAGE IN HOME(ASIAN AM,OFTEN)
46	E4 X LH2	0.238582	ETHNICITY BY LANG IN HOME(ASIAN AM,SOMETIMES)
47	TIME ASS		TIME OF ASSESSMENT(APPLICABLE FOR Y17, N/AY19)
48	STUDYCMP	-0.057134	ARE YOU STUDYING COMPUTERS? B004501 (YES)
49	DRACE2	-0.069875	DERIVED RACE/ETHNICITY (BLACK)
50	DRACE3	-0.341651	DERIVED RACE/ETHNICITY (HISPANIC)
51	DRACE4	0.185246	DERIVED RACE/ETHNICITY (ASIAN AMERICAN)

Table D.11

NAEP 1988 Mathematics Trend Conditioning Variables, Age 13

	Variable	Estimated Effect	Description
1	OVERALL	-1.504811	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.228401	GENDER (FEMALE)
3	ETHNIC2	-0.242682	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	0.086195	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	0.378006	OBSERVED ETHNICITY (ASIAN)
6	STOC3	0.534516	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7	STOC1	0.298905	SIZE AND TYPE OF COMMUNITY (NOT HI&NOT LO)
8	REGION2	-0.121025	REGION (SOUTHEAST)
9	REGION3	-0.063070	REGION (CENTRAL)
10	REGION4	-0.107134	REGION (WEST)
11	PARED2	0.140058	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.197777	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.278975	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	0.021061	PARENTS EDUCATION (MISSING, I DON'T KNOW)
15	<MODAL GRADE	-0.480949	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	0.541153	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.122176	ITEMS IN THE HOME (YES TO 3)
18	ITEMS3	0.177230	ITEMS IN THE HOME (YES TO ALL 4)
19	E2 X SEX	0.020985	ETHNICITY BY GENDER (BLACK, FEMALE)
20	E3 X SEX	0.099927	ETHNICITY BY GENDER (HISPANIC, FEMALE)
21	E4 X SEX	-0.096259	ETHNICITY BY GENDER (ASIAN AMERICAN, FEMALE)
22	E2 X PE2	-0.181870	ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
23	E2 X PE3	-0.179468	ETHNICITY BY PARENT'S ED (BLACK, POST HS)
24	E2 X PE4	-0.397062	ETHNICITY BY PARENT'S ED (BLACK, COLLEGE)
25	E2 X PE_	0.090978	ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
26	E3 X PE2	-0.033586	ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
27	E3 X PE3	-0.035114	ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
28	E3 X PE4	-0.359408	ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
29	E3 X PE_	-0.150307	ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
30	E4 X PE2	-0.412270	ETHNICITY BY PARENT'S ED (ASIAN AM, HS GRAD)
31	E4 X PE3	-1.023135	ETHNICITY BY PARENT'S ED (ASIAN AM, POST HS)
32	E4 X PE4	0.005724	ETHNICITY BY PARENT'S ED (ASIAN AM, COLLEGE)
33	E4 X PE_	-0.148864	ETHNICITY BY PARENT'S ED (ASIAN AM, UNKNOWN)
34	SCH TYP2	0.019369	SCHOOL TYPE (NOT PUBLIC)
35	SCH TYP_		SCHOOL TYPE (MISSING)
36	TV1	-0.192841	0-2 HOURS OF TV WATCHING
37	TV2	-0.259867	3-5 HOURS OF TV WATCHING
38	TV3	-0.391540	6+ HOURS OF TV WATCHING
39	HW-NO	0.143508	HOMEWORK (NONE ASSIGNED)
40	HW-YES	0.295564	HOMEWORK (YES - SOME AMOUNT)
41	HW-345	-0.046762	HOMEWORK (LINEAR AMOUNT)
42	LANGHOM3	-0.142210	LANGUAGE IN HOME OTHER THAN ENGLISH? (ALWAYS)
43	LANGHOM2	0.050961	LANGUAGE IN HOME OTHER THAN ENGLISH(SOMETIMES)

(continued)

Table D.11 (continued)

NAEP 1988 Mathematics Trend Conditioning Variables, Age 13

	Variable	Estimated Effect	Description
44	E2 X LH1	0.100579	ETHNICITY BY LANGUAGE IN HOME (BLACK, OFTEN)
45	E2 X LH2	0.051984	ETHNICITY BY LANG IN HOME (HISP., SOMETIMES)
46	E3 X LH1	0.032823	ETHNICITY BY LANGUAGE IN HOME (HISP., OFTEN)
47	E3 X LH2	-0.081489	ETHNICITY BY LANG IN HOME (HISP., SOMETIMES)
48	E4 X LH1	-0.295872	ETHNICITY BY LANGUAGE IN HOME (ASIAN AM,OFTEN)
49	E4 X LH2	-0.351225	ETHNICITY BY LANG IN HOME (ASIAN AM,SOMETIMES)
50	GRADES	0.329379	GRADES IN SCHOOL
51	TYPEMAT2	0.557133	TYPE OF MATH CLASS (REGULAR MATH)
52	TYPEMAT3	0.860079	TYPE OF MATH CLASS (PRE-ALGEBRA)
53	TYPEMAT4	1.067878	TYPE OF MATH CLASS (ALGEBRA, OTHER)
54	TIME ASS		TIME OF ASSESSMENT (APPLICABLE Y17, N/A Y19)
55	STUDYCMP	0.000685	ARE YOU STUDYING COMPUTERS? B004501 (YES)
56	DRACE2	0.021696	DERIVED RACE/ETHNICITY (BLACK)
57	DRACE3	-0.262241	DERIVED RACE/ETHNICITY (HISPANIC)
58	DRACE4	0.239560	DERIVED RACE/ETHNICITY (ASIAN AMERICAN)

Table D.12

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 17

Field	Block	Item	A	SE	B	SE	C	SE
N256101	M1	15	1.003	(0.029)	-1.407	(0.051)	0.000	(0.000)
N260601	M1	16	1.499	(0.035)	-1.136	(0.043)	0.000	(0.000)
N262401	M1	17	0.920	(0.040)	-1.326	(0.068)	0.255	(0.025)
N258804	M1	18	0.682	(0.037)	-1.852	(0.105)	0.254	(0.029)
N286001	M1	19	0.766	(0.035)	-0.944	(0.051)	0.169	(0.020)
N286002	M1	20	0.855	(0.032)	-1.658	(0.071)	0.121	(0.027)
N286302	M1	22	1.088	(0.056)	-0.439	(0.044)	0.289	(0.018)
N278501	M1	23	1.030	(0.033)	-0.759	(0.035)	0.000	(0.000)
N278502	M1	24	0.895	(0.032)	-0.559	(0.030)	0.000	(0.000)
N278503	M1	25	0.900	(0.030)	-0.831	(0.037)	0.000	(0.000)
N258802	M1	26	1.728	(0.089)	-0.175	(0.042)	0.256	(0.014)
N254602	M1	27	1.575	(0.070)	-0.024	(0.036)	0.211	(0.012)
N259901	M1	28	1.235	(0.066)	-0.225	(0.037)	0.289	(0.014)
N287101	M1	29	1.358	(0.060)	-0.382	(0.037)	0.202	(0.014)
N270301	M1	30	0.942	(0.036)	-1.403	(0.063)	0.140	(0.026)
N270302	M1	31	1.586	(0.059)	0.119	(0.031)	0.067	(0.009)
N255701	M1	32	1.451	(0.061)	-0.609	(0.045)	0.201	(0.018)
N254301	M1	33	1.035	(0.051)	0.084	(0.033)	0.258	(0.013)
N286502	M1	34	1.797	(0.097)	-0.123	(0.038)	0.181	(0.013)
N260901	M1	35	2.210	(0.113)	0.086	(0.045)	0.157	(0.012)
N256801	M1	36	1.299	(0.062)	-0.268	(0.038)	0.265	(0.015)
N258803	M1	37	0.992	(0.045)	0.250	(0.033)	0.222	(0.011)
N262601	M1	38	0.756	(0.039)	0.432	(0.038)	0.233	(0.012)
N253901	M1	39	1.647	(0.083)	0.011	(0.041)	0.259	(0.013)
N253902	M1	40	0.930	(0.057)	1.032	(0.084)	0.479	(0.011)
N253903	M1	41	1.168	(0.048)	0.915	(0.060)	0.322	(0.011)
N253904	M1	42	1.576	(0.062)	0.700	(0.058)	0.359	(0.011)
N263001	M1	43	0.664	(0.026)	0.707	(0.035)	0.000	(0.000)
N278905	M1	44	1.178	(0.046)	1.053	(0.063)	0.283	(0.010)
N287301	M1	45	0.793	(0.030)	0.120	(0.022)	0.000	(0.000)
N287302	M1	46	0.994	(0.031)	1.226	(0.048)	0.000	(0.000)
N264301	M1	47	0.800	(0.028)	0.888	(0.040)	0.000	(0.000)
N282801	M1	48	1.806	(0.054)	1.310	(0.075)	0.206	(0.010)
N251101	M1	49	1.166	(0.035)	0.949	(0.041)	0.000	(0.000)
N254601	M2	15	1.300	(0.049)	-1.815	(0.089)	0.237	(0.037)
N262301	M2	17	0.517	(0.035)	-1.239	(0.089)	0.233	(0.023)
N263201	M2	18	0.973	(0.050)	-1.348	(0.080)	0.361	(0.026)
N263202	M2	19	0.659	(0.042)	-0.434	(0.041)	0.352	(0.016)
N260101	M2	20	1.460	(0.055)	-0.973	(0.054)	0.195	(0.023)
N254001	M2	21	0.923	(0.044)	-0.847	(0.050)	0.186	(0.020)
N269001	M2	22	0.938	(0.046)	-0.398	(0.034)	0.169	(0.016)
N278901	M2	23	1.129	(0.056)	-0.229	(0.034)	0.232	(0.015)
N261501	M2	24	0.775	(0.036)	-2.237	(0.113)	0.166	(0.034)
N261801	M2	25	0.589	(0.032)	-1.985	(0.114)	0.211	(0.029)

(continued)

Table D.12 (continued)

NAEP 1988 IRT Parameters, Mathematics Trend Items, Age 17

Field	Block	Item	A	SE	B	SE	C	SE
N261201	M2	26	0.510	(0.031)	-1.518	(0.097)	0.215	(0.024)
N261601	M2	27	0.472	(0.032)	0.708	(0.055)	0.209	(0.012)
N261301	M2	28	0.581	(0.031)	-1.299	(0.074)	0.153	(0.022)
N281401	M2	29	0.685	(0.032)	-0.245	(0.027)	0.109	(0.015)
N280401	M2	30	0.550	(0.026)	-1.313	(0.067)	0.000	(0.000)
N259001	M2	31	1.188	(0.045)	-0.218	(0.025)	0.000	(0.000)
N287102	M2	32	1.114	(0.050)	-0.556	(0.040)	0.172	(0.018)
N286301	M2	33	1.350	(0.071)	-0.450	(0.043)	0.221	(0.017)
N286501	M2	34	1.142	(0.049)	-0.847	(0.049)	0.149	(0.021)
N262501	M2	35	0.878	(0.060)	0.217	(0.043)	0.477	(0.013)
N262502	M2	36	0.598	(0.045)	1.756	(0.141)	0.365	(0.010)
N263101	M2	37	0.754	(0.032)	-0.569	(0.033)	0.000	(0.000)
N258801	M2	38	1.904	(0.110)	0.216	(0.048)	0.284	(0.012)
N264701	M2	39	1.578	(0.082)	-0.033	(0.038)	0.216	(0.015)
N261001	M2	40	0.806	(0.046)	-0.734	(0.052)	0.216	(0.022)
N251701	M2	41	0.892	(0.046)	0.005	(0.029)	0.147	(0.015)
N278902	M2	42	1.162	(0.065)	0.014	(0.036)	0.236	(0.016)
N260801	M2	43	1.301	(0.044)	0.388	(0.030)	0.000	(0.000)
N278903	M2	44	1.921	(0.092)	0.365	(0.051)	0.227	(0.013)
N255601	M2	45	1.248	(0.059)	1.576	(0.107)	0.332	(0.011)
N255301	M2	46	1.539	(0.052)	1.503	(0.086)	0.219	(0.009)
N268901	M2	47	1.691	(0.062)	0.639	(0.054)	0.184	(0.012)
N268801	M2	48	0.917	(0.039)	1.654	(0.085)	0.102	(0.009)
N255801	M2	49	0.679	(0.030)	1.668	(0.080)	0.000	(0.000)
N266501	M3	31	0.775	(0.060)	-0.326	(0.041)	0.244	(0.017)
N271301	M3	32	1.374	(0.120)	0.185	(0.048)	0.261	(0.014)
N255501	M3	33	0.808	(0.054)	0.668	(0.059)	0.232	(0.013)
N256001	M3	34	1.055	(0.068)	0.066	(0.027)	0.000	(0.000)
N257101	M3	35	0.579	(0.054)	1.853	(0.181)	0.254	(0.011)

Table D.13

NAEP 1988 Mathematics Trend Conditioning Variables, Age 17

	Variable	Estimated Effect	Description
1	OVERALL	0.466202	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.227644	GENDER (FEMALE)
3	ETHNIC2	-0.326424	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	-0.125207	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	-0.542147	OBSERVED ETHNICITY (ASIAN)
6	STOC3	0.355679	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
7	STOC1	0.268174	SIZE AND TYPE OF COMMUNITY (NOT HI&NOT LO)
8	REGION2	-0.035567	REGION (SOUTHEAST)
9	REGION3	0.092946	REGION (CENTRAL)
10	REGION4	0.041544	REGION (WEST)
11	PARED2	-0.009106	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.276562	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.215802	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	0.039054	PARENTS EDUCATION (MISSING, I DON'T KNOW)
15	<MODAL GRADE	-0.212266	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	-0.091063	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.032057	ITEMS IN THE HOME (YES TO 3)
18	ITEMS3	0.089343	ITEMS IN THE HOME (YES TO ALL 4)
19	E2 X SEX	0.130167	ETHNICITY BY GENDER (BLACK, FEMALE)
20	E3 X SEX	0.294555	ETHNICITY BY GENDER (HISPANIC, FEMALE)
21	E4 X SEX	-0.190247	ETHNICITY BY GENDER (ASIAN AMERICAN, FEMALE)
22	E2 X PE2	-0.014269	ETHNICITY BY PARENT'S ED (BLACK, HS GRAD)
23	E2 X PE3	-0.186204	ETHNICITY BY PARENT'S ED (BLACK, POST HS)
24	E2 X PE4	-0.163440	ETHNICITY BY PARENT'S ED (BLACK, COLLEGE)
25	E2 X PE_	-0.256462	ETHNICITY BY PARENT'S ED (BLACK, UNKNOWN)
26	E3 X PE2	0.037801	ETHNICITY BY PARENT'S ED (HISPANIC, HS GRAD)
27	E3 X PE3	-0.197622	ETHNICITY BY PARENT'S ED (HISPANIC, POST HS)
28	E3 X PE4	-0.148578	ETHNICITY BY PARENT'S ED (HISPANIC, COLLEGE)
29	E3 X PE_	0.076608	ETHNICITY BY PARENT'S ED (HISPANIC, UNKNOWN)
30	E4 X PE2	1.148569	ETHNICITY BY PARENT'S ED (ASIAN AM, HS GRAD)
31	E4 X PE3	0.548141	ETHNICITY BY PARENT'S ED (ASIAN AM, POST HS)
32	E4 X PE4	-0.003476	ETHNICITY BY PARENT'S ED (ASIAN AM, COLLEGE)
33	E4 X PE_	0.555852	ETHNICITY BY PARENT'S ED (ASIAN AM, UNKNOWN)
34	SCH TYP2	-0.130104	SCHOOL TYPE (NOT PUBLIC)
35	SCH TYP_		SCHOOL TYPE (MISSING)
36	TV1	-1.980878	0-2 HOURS OF TV WATCHING
37	TV2	-1.992986	3-5 HOURS OF TV WATCHING
38	TV3	-2.079726	6+ HOURS OF TV WATCHING
39	HW-NO	-0.243494	HOMEWORK (NONE ASSIGNED)
40	HW-YES	0.104266	HOMEWORK (YES - SOME AMOUNT)
41	HW-345	-0.024606	HOMEWORK (LINEAR AMOUNT)
42	LANGHOM3	-0.306630	LANGUAGE IN HOME OTHER THAN ENGLISH? (ALWAYS)
43	LANGHOM2	-0.027324	LANGUAGE IN HOME OTHER THAN ENGLISH(SOMETIMES)

(continued)

Table D.13 (continued)

NAEP 1988 Mathematics Trend Conditioning Variables, Age 17

	Variable	Estimated Effect	Description
44	E2 X LH1	0.234334	ETHNICITY BY LANGUAGE IN HOME (BLACK, OFTEN)
45	E2 X LH2	-0.085786	ETHNICITY BY LANG IN HOME (HISP, SOMETIMES)
46	E3 X LH1	0.372056	ETHNICITY BY LANGUAGE IN HOME (HISP, OFTEN)
47	E3 X LH2	0.068137	ETHNICITY BY LANG IN HOME (HISP, SOMETIMES)
48	E4 X LH1	0.542742	ETHNICITY BY LANG IN HOME (ASIAN AM, OFTEN)
49	E4 X LH2	0.390736	ETHNICITY BY LANG IN HOME (ASIAN AM,SOMETIMES)
50	NMATH1	-0.221100	HIGHEST LEVEL MATH TAKEN (PRE-ALGEBRA)
51	NMATH2	0.252774	HIGHEST LEVEL MATH TAKEN (ALGEBRA)
52	NMATH3	0.354687	HIGHEST LEVEL MATH TAKEN (GEOMETRY)
53	NMATH4	0.700470	HIGHEST LEVEL MATH TAKEN (ALGEBRA-2)
54	NMATH5	1.208891	HIGHEST LEVEL MATH TAKEN (CALCULUS)
55	COMPUTER	-0.009892	COMPUTER CLASS TAKEN ? (YES)
56	GRADES	0.293596	GRADES IN SCHOOL
57	HSPROG2	0.196396	HIGH SCHOOL PROGRAM (COLLEGE PREP)
58	HSPROG3	-0.090029	HIGH SCHOOL PROGRAM (VOC/TECH)
59	DRACE2	0.119675	DERIVED RACE/ETHNICITY (BLACK)
60	DRACE3	-0.202548	DERIVED RACE/ETHNICITY (HISPANIC)
61	DRACE4	-0.056777	DERIVED RACE/ETHNICITY (ASIAN AMERICAN)

Table D.14

NAEP 1988 IRT Parameters, Science Trend Items, Age 9

Field	Block	Item	A	SE	B	SE	C	SE
N400001	S1	6	0.650	(0.056)	-1.173	(0.109)	0.237	(0.030)
N400301	S1	8	0.993	(0.113)	-0.130	(0.055)	0.340	(0.021)
N400401	S1	9	1.246	(0.092)	-1.214	(0.117)	0.417	(0.035)
N400402	S1	10	1.829	(0.126)	-0.733	(0.089)	0.280	(0.027)
N400403	S1	11	0.566	(0.063)	-1.941	(0.223)	0.422	(0.036)
N400404	S1	12	1.164	(0.098)	-0.651	(0.078)	0.322	(0.026)
N400405	S1	13	1.012	(0.095)	-0.748	(0.090)	0.390	(0.027)
N400501	S1	14	0.545	(0.063)	0.593	(0.083)	0.330	(0.018)
N400101	S1	15	0.294	(0.069)	2.732	(0.643)	0.460	(0.016)
N400102	S1	16	0.455	(0.076)	1.909	(0.329)	0.424	(0.015)
N400601	S1	17	0.647	(0.062)	-0.202	(0.044)	0.225	(0.021)
N400701	S1	18	0.741	(0.066)	0.070	(0.040)	0.202	(0.019)
N400901	S1	19	0.333	(0.049)	1.804	(0.268)	0.253	(0.015)
N401001	S1	20	0.542	(0.053)	0.729	(0.082)	0.210	(0.016)
N401101	S1	21	0.292	(0.048)	1.737	(0.288)	0.275	(0.016)
N401201	S1	22	0.851	(0.080)	2.036	(0.215)	0.243	(0.011)
N401301	S1	23	0.504	(0.060)	1.478	(0.183)	0.269	(0.014)
N401501	S2	1	0.260	(0.047)	0.249	(0.060)	0.347	(0.019)
N401601	S2	2	0.599	(0.058)	-1.492	(0.150)	0.207	(0.033)
N401702	S2	4	0.304	(0.059)	0.556	(0.118)	0.452	(0.018)
N401703	S2	5	0.299	(0.059)	1.035	(0.209)	0.443	(0.017)
N401801	S2	6	0.686	(0.109)	-0.035	(0.057)	0.447	(0.021)
N401802	S2	7	0.570	(0.082)	-0.962	(0.147)	0.432	(0.028)
N401803	S2	8	0.455	(0.075)	-0.279	(0.068)	0.440	(0.023)
N401804	S2	9	0.346	(0.068)	1.698	(0.338)	0.424	(0.016)
N401901	S2	10	0.469	(0.072)	1.855	(0.291)	0.318	(0.015)
N402001	S2	11	0.935	(0.091)	-1.045	(0.118)	0.381	(0.032)
N402002	S2	12	1.224	(0.106)	-1.036	(0.115)	0.386	(0.034)
N402005	S2	15	0.712	(0.103)	-0.510	(0.091)	0.411	(0.026)
N402101	S2	16	0.562	(0.061)	-0.332	(0.051)	0.206	(0.022)
N402201	S2	17	0.231	(0.039)	0.333	(0.067)	0.245	(0.019)
N402401	S2	18	0.253	(0.051)	2.764	(0.561)	0.235	(0.015)
N402501	S2	19	0.622	(0.090)	2.692	(0.407)	0.258	(0.011)
N402602	S2	21	0.401	(0.063)	-0.686	(0.117)	0.439	(0.022)
N402701	S2	23	0.453	(0.058)	1.980	(0.261)	0.199	(0.013)
N402801	S2	24	1.084	(0.083)	2.031	(0.189)	0.161	(0.009)
N402901	S2	25	0.373	(0.094)	4.734	(1.194)	0.185	(0.010)
N403001	S3	12	0.422	(0.062)	-5.043	(0.745)	0.238	(0.053)
N403101	S3	13	0.638	(0.062)	-3.422	(0.342)	0.232	(0.051)
N403201	S3	14	0.404	(0.048)	-3.042	(0.368)	0.212	(0.039)
N403202	S3	15	0.291	(0.038)	-1.195	(0.161)	0.238	(0.024)
N403301	S3	16	0.624	(0.056)	-1.079	(0.105)	0.218	(0.029)
N403401	S3	17	0.234	(0.047)	0.435	(0.095)	0.331	(0.019)
N403501	S3	18	0.563	(0.067)	0.257	(0.057)	0.400	(0.019)

(continued)

Table D.14 (continued)

NAEP 1988 IRT Parameters, Science Trend Items, Age 9

Field	Block	Item	A	SE	B	SE	C	SE
N403502	S3	19	0.551	(0.059)	-1.918	(0.211)	0.404	(0.034)
N403503	S3	20	0.412	(0.060)	0.152	(0.054)	0.409	(0.020)
N403601	S3	21	0.811	(0.069)	0.534	(0.065)	0.254	(0.016)
N403701	S3	22	3.290	(0.390)	-0.287	(0.108)	0.312	(0.021)
N403702	S3	23	3.150	(0.247)	-0.496	(0.118)	0.374	(0.023)
N403703	S3	24	2.076	(0.204)	-0.326	(0.077)	0.302	(0.021)
N403801	S3	25	0.359	(0.057)	1.082	(0.180)	0.428	(0.017)
N403803	S3	27	0.497	(0.056)	-0.991	(0.119)	0.393	(0.026)
N403804	S3	28	0.484	(0.063)	-0.506	(0.080)	0.408	(0.023)
N403901	S3	29	0.653	(0.056)	-0.309	(0.046)	0.193	(0.023)
N404001	S3	30	0.203	(0.036)	1.764	(0.317)	0.223	(0.016)
N404201	S3	31	0.425	(0.050)	1.363	(0.165)	0.216	(0.015)
8105013-A001/001			0.504	(0.068)	-0.150	(0.105)	0.220	(0.056)
8202072-A001/001			0.606	(0.175)	2.669	(0.840)	0.233	(0.026)
8204035-A001/001			0.547	(0.062)	-0.698	(0.112)	0.187	(0.051)
8204085-A001/001			0.412	(0.056)	-0.292	(0.100)	0.189	(0.055)
8303086-A001/001			0.308	(0.060)	-1.816	(0.376)	0.483	(0.071)
8C17C04-A001/001			0.686	(0.118)	1.246	(0.276)	0.196	(0.034)
8C21C08-A001/001			0.894	(0.112)	0.963	(0.182)	0.131	(0.025)
8C23C11-A001/001			0.881	(0.084)	-1.138	(0.135)	0.180	(0.048)
8C24C07-A001/001			0.512	(0.075)	0.138	(0.117)	0.230	(0.055)
8C52C03-A001/001			0.369	(0.105)	3.181	(0.934)	0.188	(0.036)
8C52C04-A001/001			1.116	(0.119)	0.051	(0.095)	0.201	(0.035)
8C54C10-A001/001			0.522	(0.097)	1.528	(0.325)	0.163	(0.037)
8C55C03-A001/001			0.398	(0.079)	1.170	(0.281)	0.232	(0.053)
8C56C02-A001/001			0.704	(0.094)	0.766	(0.162)	0.167	(0.036)
8C58C10-A001/001			0.665	(0.098)	1.036	(0.205)	0.162	(0.033)
8C61C09-A001/001			0.503	(0.065)	0.221	(0.098)	0.173	(0.046)
8C63C13-A001/001			0.370	(0.061)	-0.388	(0.135)	0.327	(0.064)
8C71C09-A001/001			0.578	(0.066)	-0.546	(0.105)	0.194	(0.051)
8C71C12-A001/001			0.947	(0.103)	-2.060	(0.261)	0.198	(0.055)
8C71C13-A001/001			0.622	(0.095)	-2.270	(0.372)	0.465	(0.071)
8C71C13-A002/002			0.605	(0.091)	-2.207	(0.358)	0.459	(0.070)
8C71C13-A003/003			0.546	(0.168)	1.822	(0.655)	0.546	(0.037)
8C71C13-A004/004			0.498	(0.075)	-0.987	(0.192)	0.432	(0.065)
8C82C08-A001/001			0.522	(0.081)	0.754	(0.172)	0.184	(0.045)

Table D.15

NAEP 1988 IRT Parameters, Science Trend Items, Age 13

Field	Block	Item	A	SE	B	SE	C	SE
N404501	S1	12	1.153	(0.055)	-2.021	(0.119)	0.164	(0.042)
N404601	S1	13	0.318	(0.039)	-0.641	(0.084)	0.228	(0.021)
N404701	S1	14	0.601	(0.043)	-1.538	(0.117)	0.194	(0.029)
N404702	S1	15	0.449	(0.041)	-0.140	(0.033)	0.201	(0.018)
N400201	S1	16	0.464	(0.041)	-1.666	(0.151)	0.206	(0.029)
N404901	S1	17	0.691	(0.051)	-0.629	(0.057)	0.209	(0.022)
N404801	S1	20	1.372	(0.085)	-1.624	(0.136)	0.422	(0.043)
N404802	S1	21	1.610	(0.140)	-0.514	(0.077)	0.360	(0.022)
N404803	S1	22	0.956	(0.078)	0.240	(0.049)	0.321	(0.016)
N405001	S1	23	0.349	(0.037)	0.200	(0.037)	0.214	(0.017)
N405101	S1	24	0.794	(0.052)	0.968	(0.077)	0.199	(0.012)
N405201	S1	25	0.315	(0.036)	-0.124	(0.033)	0.182	(0.019)
N405301	S1	26	0.623	(0.049)	1.251	(0.107)	0.199	(0.012)
N405401	S1	27	0.801	(0.053)	1.138	(0.087)	0.181	(0.011)
N401201	S1	28	0.544	(0.049)	0.415	(0.051)	0.249	(0.016)
N405501	S1	29	0.628	(0.052)	-0.031	(0.035)	0.197	(0.019)
N405601	S1	30	0.233	(0.034)	1.041	(0.153)	0.198	(0.016)
N405701	S1	31	1.012	(0.067)	0.715	(0.065)	0.185	(0.013)
N405801	S1	32	0.493	(0.044)	1.324	(0.124)	0.166	(0.012)
N405901	S1	33	0.637	(0.049)	1.658	(0.137)	0.158	(0.011)
N406001	S1	34	0.455	(0.107)	4.846	(1.148)	0.174	(0.008)
N406101	S1	35	0.531	(0.120)	4.384	(1.008)	0.207	(0.008)
N406201	S1	36	0.360	(0.089)	5.620	(1.399)	0.099	(0.007)
N406301	S2	10	0.356	(0.052)	-1.563	(0.231)	0.430	(0.026)
N406302	S2	11	0.386	(0.051)	-0.408	(0.069)	0.428	(0.021)
N406303	S2	12	0.606	(0.063)	1.470	(0.166)	0.392	(0.013)
N406304	S2	13	0.471	(0.066)	1.354	(0.200)	0.419	(0.015)
N406401	S2	14	0.504	(0.066)	-0.157	(0.050)	0.461	(0.020)
N406402	S2	15	0.861	(0.090)	0.303	(0.062)	0.405	(0.018)
N406403	S2	16	0.753	(0.074)	-1.328	(0.142)	0.419	(0.031)
N406404	S2	17	0.910	(0.111)	-0.305	(0.067)	0.457	(0.022)
N406405	S2	18	0.628	(0.066)	-0.528	(0.075)	0.402	(0.025)
N406501	S2	19	0.495	(0.043)	0.628	(0.064)	0.170	(0.016)
N406601	S2	20	0.491	(0.044)	-0.855	(0.082)	0.175	(0.023)
N406701	S2	21	0.576	(0.049)	0.093	(0.034)	0.240	(0.016)
N406801	S2	22	1.128	(0.074)	-1.417	(0.114)	0.396	(0.036)
N406802	S2	23	0.342	(0.047)	0.687	(0.104)	0.445	(0.015)
N406803	S2	24	0.816	(0.074)	-0.660	(0.074)	0.382	(0.022)
N406804	S2	25	1.057	(0.073)	-1.014	(0.086)	0.371	(0.027)
N406805	S2	26	1.037	(0.097)	1.523	(0.181)	0.550	(0.011)
N406806	S2	27	0.440	(0.053)	0.226	(0.050)	0.423	(0.017)
N406901	S2	28	0.613	(0.052)	0.019	(0.034)	0.231	(0.017)
N407001	S2	29	0.263	(0.035)	0.158	(0.038)	0.182	(0.019)
N407101	S2	30	0.817	(0.055)	2.218	(0.168)	0.126	(0.009)

(continued)

Table D.15 (continued)

NAEP 1988 IRT Parameters, Science Trend Items, Age 13

Field	Block	Item	A	SE	B	SE	C	SE
N407201	S2	31	0.470	(0.041)	0.437	(0.050)	0.207	(0.015)
N407301	S2	32	0.319	(0.039)	1.672	(0.208)	0.234	(0.013)
N407302	S2	33	0.346	(0.046)	1.817	(0.245)	0.270	(0.014)
N408001	S2	34	0.848	(0.050)	1.268	(0.087)	0.176	(0.010)
N407601	S2	35	0.453	(0.044)	1.743	(0.173)	0.180	(0.012)
N407701	S2	37	0.564	(0.044)	1.273	(0.107)	0.144	(0.012)
N407801	S2	38	0.666	(0.055)	2.158	(0.189)	0.199	(0.010)
N407901	S2	39	0.383	(0.037)	0.849	(0.089)	0.168	(0.015)
N408201	S2	40	0.567	(0.070)	3.245	(0.415)	0.206	(0.009)
N408301	S3	10	0.788	(0.061)	0.970	(0.089)	0.298	(0.013)
N408302	S3	11	0.708	(0.065)	-1.545	(0.152)	0.408	(0.033)
N408303	S3	12	0.647	(0.060)	-1.607	(0.163)	0.415	(0.031)
N408304	S3	13	0.971	(0.079)	-1.384	(0.129)	0.414	(0.034)
N408401	S3	14	0.240	(0.032)	-1.476	(0.199)	0.223	(0.022)
N408501	S3	15	0.733	(0.056)	-0.896	(0.077)	0.205	(0.025)
N408502	S3	16	0.390	(0.040)	1.337	(0.140)	0.154	(0.013)
N408601	S3	17	0.388	(0.035)	-1.071	(0.102)	0.153	(0.022)
N408701	S3	18	0.346	(0.038)	-0.101	(0.031)	0.212	(0.018)
N408801	S3	19	0.174	(0.030)	0.655	(0.117)	0.234	(0.017)
N408901	S3	20	0.743	(0.079)	0.274	(0.055)	0.445	(0.015)
N408902	S3	21	0.889	(0.069)	-1.740	(0.149)	0.410	(0.038)
N408903	S3	22	0.656	(0.066)	0.434	(0.062)	0.404	(0.015)
N408904	S3	23	0.540	(0.060)	0.877	(0.107)	0.411	(0.014)
N409001	S3	24	0.599	(0.045)	-0.364	(0.040)	0.163	(0.019)
N409101	S3	25	0.635	(0.045)	-1.494	(0.113)	0.239	(0.029)
N409102	S3	26	0.556	(0.047)	0.178	(0.036)	0.229	(0.016)
N409103	S3	27	0.518	(0.059)	2.017	(0.235)	0.306	(0.011)
N409201	S3	28	0.292	(0.039)	0.393	(0.061)	0.261	(0.017)
N409301	S3	29	0.706	(0.056)	-0.145	(0.035)	0.165	(0.018)
N409501	S3	33	0.607	(0.052)	2.148	(0.191)	0.134	(0.009)
N409601	S3	34	0.708	(0.061)	1.717	(0.162)	0.290	(0.011)
N409701	S3	35	0.633	(0.060)	2.485	(0.248)	0.165	(0.009)

Table D.16

NAEP 1988 IRT Parameters, Science Trend Items, Age 17

Field	Block	Item	A	SE	B	SE	C	SE
N400201	S1	12	0.543	(0.116)	-1.669	(0.370)	0.196	(0.049)
N404601	S1	13	0.542	(0.118)	-0.565	(0.150)	0.197	(0.043)
N410003	S1	16	0.509	(0.121)	-1.988	(0.486)	0.400	(0.055)
N410004	S1	17	0.499	(0.129)	-1.225	(0.334)	0.401	(0.050)
N409901	S1	18	0.867	(0.168)	-0.931	(0.209)	0.191	(0.046)
N408601	S1	19	0.426	(0.091)	-1.329	(0.295)	0.164	(0.043)
N409301	S1	20	0.625	(0.122)	-1.324	(0.274)	0.149	(0.044)
N406301	S1	21	0.334	(0.091)	-1.322	(0.371)	0.410	(0.047)
N406302	S1	22	0.420	(0.105)	-0.246	(0.118)	0.401	(0.042)
N406303	S1	23	0.506	(0.129)	0.383	(0.147)	0.397	(0.037)
N406304	S1	24	0.511	(0.138)	-0.276	(0.132)	0.395	(0.044)
N410101	S1	25	0.626	(0.158)	-0.700	(0.207)	0.394	(0.046)
N410102	S1	26	0.433	(0.113)	-0.401	(0.144)	0.404	(0.043)
N410103	S1	27	0.566	(0.139)	-1.408	(0.365)	0.396	(0.053)
N406601	S1	28	0.547	(0.115)	-0.915	(0.210)	0.151	(0.042)
N405001	S1	29	0.462	(0.098)	-0.305	(0.102)	0.198	(0.039)
N401201	S1	30	0.613	(0.124)	-0.226	(0.097)	0.229	(0.040)
N405201	S1	31	0.444	(0.095)	-0.703	(0.168)	0.152	(0.040)
N410201	S1	32	0.491	(0.119)	1.890	(0.476)	0.199	(0.030)
N406001	S1	33	0.471	(0.115)	2.129	(0.536)	0.197	(0.026)
N409501	S1	34	0.714	(0.129)	1.100	(0.225)	0.133	(0.028)
N406101	S1	35	0.494	(0.142)	2.885	(0.854)	0.214	(0.025)
N406201	S1	37	0.658	(0.130)	2.184	(0.457)	0.116	(0.022)
N408101	S1	38	0.625	(0.124)	1.626	(0.344)	0.142	(0.026)
N406401	S2	10	0.632	(0.158)	-0.678	(0.205)	0.395	(0.049)
N406402	S2	11	0.676	(0.189)	-0.075	(0.124)	0.391	(0.043)
N406403	S2	12	0.815	(0.189)	-1.522	(0.388)	0.395	(0.059)
N406404	S2	13	0.721	(0.182)	-1.204	(0.333)	0.393	(0.056)
N406405	S2	14	0.637	(0.171)	-0.963	(0.292)	0.397	(0.054)
N410401	S2	15	0.396	(0.093)	0.086	(0.088)	0.244	(0.039)
N406801	S2	16	0.672	(0.157)	-1.921	(0.471)	0.396	(0.059)
N406802	S2	17	0.452	(0.121)	1.281	(0.367)	0.403	(0.033)
N406803	S2	18	0.575	(0.136)	-1.248	(0.314)	0.396	(0.051)
N406804	S2	19	0.709	(0.147)	-1.539	(0.344)	0.391	(0.056)
N406805	S2	20	0.458	(0.117)	0.473	(0.164)	0.408	(0.037)
N406806	S2	21	0.396	(0.105)	0.270	(0.128)	0.406	(0.040)
N410501	S2	22	0.415	(0.088)	-0.420	(0.118)	0.150	(0.039)
N410601	S2	23	1.057	(0.208)	2.077	(0.507)	0.229	(0.025)
N410602	S2	24	0.430	(0.122)	-2.476	(0.714)	0.405	(0.058)
N410603	S2	25	0.768	(0.170)	1.333	(0.336)	0.338	(0.029)
N410604	S2	26	0.414	(0.110)	-2.139	(0.577)	0.405	(0.055)
N406901	S2	27	0.500	(0.109)	-0.531	(0.142)	0.196	(0.042)
N407401	S2	28	0.388	(0.052)	-0.059	(0.042)	0.486	(0.017)
N407403	S2	30	0.581	(0.151)	-0.258	(0.137)	0.393	(0.046)

(continued)

Table D.16 (continued)

NAEP 1988 IRT Parameters, Science Trend Items, Age 17

Field	Block	Item	A	SE	B	SE	C	SE
N407404	S2	31	0.714	(0.166)	-1.370	(0.349)	0.395	(0.057)
N407201	S2	32	0.500	(0.106)	0.120	(0.084)	0.153	(0.035)
N407001	S2	33	0.333	(0.079)	-0.920	(0.232)	0.155	(0.042)
N410701	S2	34	0.542	(0.120)	0.833	(0.209)	0.201	(0.033)
N407701	S2	35	0.450	(0.097)	0.898	(0.214)	0.152	(0.032)
N407301	S2	36	0.346	(0.083)	0.510	(0.147)	0.204	(0.036)
N407302	S2	37	0.445	(0.110)	0.917	(0.249)	0.246	(0.035)
N407101	S2	38	0.614	(0.126)	1.878	(0.410)	0.150	(0.026)
N410801	S2	39	0.542	(0.124)	1.554	(0.376)	0.193	(0.030)
N410901	S2	40	0.707	(0.134)	1.777	(0.367)	0.155	(0.024)
N411001	S2	41	0.545	(0.145)	2.730	(0.751)	0.193	(0.024)
N408301	S3	10	0.834	(0.186)	-0.241	(0.125)	0.381	(0.041)
N408302	S3	11	0.457	(0.119)	-1.685	(0.456)	0.401	(0.056)
N408303	S3	12	0.543	(0.122)	-2.012	(0.470)	0.398	(0.057)
N408304	S3	13	0.640	(0.162)	-1.585	(0.426)	0.396	(0.058)
N405101	S3	14	0.595	(0.111)	0.272	(0.103)	0.235	(0.035)
N408901	S3	15	0.769	(0.168)	-1.203	(0.292)	0.393	(0.053)
N408902	S3	16	0.836	(0.165)	-1.922	(0.422)	0.395	(0.062)
N408903	S3	17	0.563	(0.127)	-0.172	(0.112)	0.392	(0.041)
N408904	S3	18	0.586	(0.135)	-0.374	(0.135)	0.398	(0.043)
N405401	S3	19	0.619	(0.104)	0.631	(0.138)	0.145	(0.031)
N411301	S3	20	0.469	(0.139)	3.814	(1.163)	0.120	(0.023)
N405501	S3	21	0.584	(0.121)	-0.295	(0.108)	0.196	(0.041)
N411101	S3	22	0.507	(0.096)	0.255	(0.093)	0.150	(0.035)
N411201	S3	23	0.566	(0.105)	0.490	(0.126)	0.195	(0.033)
N408801	S3	24	0.505	(0.101)	-0.340	(0.105)	0.198	(0.039)
N411401	S3	25	0.846	(0.151)	0.534	(0.137)	0.152	(0.030)
N411501	S3	26	0.860	(0.125)	1.749	(0.300)	0.179	(0.024)
N411502	S3	27	0.619	(0.131)	-1.037	(0.240)	0.237	(0.048)
N411601	S3	28	0.609	(0.108)	1.227	(0.244)	0.184	(0.030)
N411701	S3	29	0.745	(0.119)	1.395	(0.256)	0.169	(0.027)
N411801	S3	30	1.069	(0.175)	0.650	(0.161)	0.167	(0.031)
N411901	S3	31	0.762	(0.122)	1.429	(0.261)	0.142	(0.025)
N412001	S3	32	0.572	(0.119)	2.048	(0.453)	0.187	(0.029)

Table D.17

NAEP 1988 Science Trend Conditioning Variables, Age 9

	Variable	Estimated Effect	Description
1	OVERALL	-0.167629	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.160032	GENDER (FEMALE)
3	ETHNIC2	-0.716027	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	-0.677694	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	-0.143962	OBSERVED ETHNICITY (ASIAN)
6	STOC2	-0.400385	SIZE AND TYPE OF COMMUNITY (LOW METRO)
7	STOC3	0.114765	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
8	REGION2	0.105314	REGION (SOUTHEAST)
9	REGION3	0.202669	REGION (CENTRAL)
10	REGION4	0.081810	REGION (WEST)
11	PARED2	0.200699	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.279235	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.435635	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	0.172272	PARENTS EDUCATION (MISSING, DON'T KNOW)
15	<MODAL GRADE	-0.498134	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	1.050936	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.289243	ITEMS IN THE HOME (YES TO 3)
18	ITEHS3	0.478227	ITEMS IN THE HOME (YES TO ALL 4)
19	SCH TYP2	0.076284	SCHOOL TYPE (NOT PUBLIC)

Table D.18

NAEP 1988 Science Trend Conditioning Variables, Age 13

		Estimated	
	Variable	Effect	Description
1	OVERALL	-0.048884	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.267412	GENDER (FEMALE)
3	ETHNIC2	-0.719052	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	-0.524609	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	0.161636	OBSERVED ETHNICITY (ASIAN)
6	STOC2	-0.395130	SIZE AND TYPE OF COMMUNITY (LOW METRO)
7	STOC3	-0.007911	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
8	REGION2	-0.077003	REGION (SOUTHEAST)
9	REGION3	0.046762	REGION (CENTRAL)
10	REGION4	-0.102571	REGION (WEST)
11	PARED2	0.107733	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.357308	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.412279	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	-0.047971	PARENTS EDUCATION (MISSING, DON'T KNOW)
15	<MODAL GRADE	-0.530171	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	0.969538	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.222418	ITEMS IN THE HOME (YES TO 3)
18	ITEMS3	0.404732	ITEMS IN THE HOME (YES TO ALL 4)
19	SCH TYP2	0.128735	SCHOOL TYPE (NOT PUBLIC)

Table D.19

NAEP 1988 Science Trend Conditioning Variables, Age 17

	Variable	Estimated Effect	Description
1	OVERALL	-0.018353	OVERALL CONSTANT '1' FOR EVERYONE
2	GENDER2	-0.422265	GENDER (FEMALE)
3	ETHNIC2	-0.675393	OBSERVED ETHNICITY (BLACK)
4	ETHNIC3	-0.028940	OBSERVED ETHNICITY (HISPANIC)
5	ETHNIC4	0.105174	OBSERVED ETHNICITY (ASIAN)
6	STOC2	-0.215624	SIZE AND TYPE OF COMMUNITY (LOW METRO)
7	STOC3	0.200910	SIZE AND TYPE OF COMMUNITY (HIGH METRO)
8	REGION2	-0.078230	REGION (SOUTHEAST)
9	REGION3	-0.145136	REGION (CENTRAL)
10	REGION4	-0.156447	REGION (WEST)
11	PARED2	0.277744	PARENTS EDUCATION (HIGH SCHOOL GRAD)
12	PARED3	0.506933	PARENTS EDUCATION (POST HIGH SCHOOL)
13	PARED4	0.724225	PARENTS EDUCATION (COLLEGE GRAD)
14	PARED_	-0.353136	PARENTS EDUCATION (MISSING, DON'T KNOW)
15	<MODAL GRADE	-0.540566	MODAL GRADE (LESS THAN MODAL GRADE)
16	>MODAL GRADE	0.345666	MODAL GRADE (GREATER THAN MODAL GRADE)
17	ITEMS2	0.091730	ITEMS IN THE HOME (YES TO 3)
18	ITEMS3	0.208488	ITEMS IN THE HOME (YES TO ALL 4)
19	SCH TYP2	-0.094395	SCHOOL TYPE (NOT PUBLIC)

APPENDIX E

Estimation of the Standard Errors of the Adjusted 1986 NAEP Results

Appendix E

ESTIMATION OF THE STANDARD ERRORS OF THE ADJUSTED 1986 NAEP RESULTS

Eugene G. Johnson
Robert J. Mislevy
Rebecca Zwick

Common-population linear equating of the results from the 1988 bridges was used to link the 1986 results to the 1984 reading scale. The procedures described below were carried out for each age cohort independently. Let $\hat{\mu}_1$ and $\hat{\sigma}_1$ be, respectively, the estimated mean and standard deviation of the proficiency scores from the 1988 bridge to 1986, these values being in the 86-P provisional metric[1] (see Chapter 6). Let $\hat{\mu}_2$ and $\hat{\sigma}_2$ be the estimated mean and standard deviation of the proficiency scores from the 1988 bridge to 1984; these values being on the same metric as the 1984 reading scale[2]. The common-population linear equating of the two sets of bridge values comes about by matching the estimated moments for the two bridges, producing the following equating function for going from the 1986 (86-P) metric to the 1984 metric:

$$f(\theta, \hat{A}, \hat{B}) = \hat{A}\,\theta + \hat{B} \tag{1}$$

[1] that is, with reference to the item parameters estimated from the 1986 data only.

[2] that is, with reference to item parameters estimated from the 1984 data only.

where

θ is a proficiency value in the 1986 metric,

$\hat{A} = \hat{\sigma}_2 / \hat{\sigma}_1$, and

$\hat{B} = \hat{\mu}_2 - \hat{\mu}_1 \hat{\sigma}_2 / \hat{\sigma}_1$.

Equation (1) is used to produce adjusted proficiency values for the 1986 assessment. In particular, let $\overline{X}$ be the estimated 1986 mean proficiency for some subgroup of the population (or for the population as a whole), this estimate based on the proficiency values in the provisional 1986 metric. The adjusted estimate of the 1986 mean proficiency of the subgroup, in the metric of the 1984 reading scale, is

$$\overline{X}_{adj} = f(\overline{X}, \hat{A}, \hat{B}) = \hat{A}\overline{X} + \hat{B}. \qquad (2)$$

If $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\mu}_2$, $\hat{\sigma}_2$ and, consequently, $\hat{A}$ and $\hat{B}$ were known without error, the variance of $\overline{X}_{adj}$ would be simply

$$\hat{A}^2 \mathrm{Var}(\overline{X}). \qquad (3)$$

However, since $\hat{A}$ and $\hat{B}$ are based on estimates from samples of the 1988 population, they are subject to variability. Ignoring this variability by using (3) as the estimate of the variance of the adjusted subgroup mean will result in an underestimate of the true variability of $\overline{X}_{adj}$.

A large sample approximation to the variance of $\overline{X}_{adj}$ is

$$\left[\frac{\partial(f)}{\partial\overline{X}} \; \frac{\partial(f)}{\partial\hat{A}} \; \frac{\partial(f)}{\partial\hat{B}} \right] \Sigma \left[\frac{\partial(f)}{\partial\overline{X}} \; \frac{\partial(f)}{\partial\hat{A}} \; \frac{\partial(f)}{\partial\hat{B}} \right]^T$$

$$= \quad [\ \hat{A}\ \overline{X}\ 1\ ]\ \Sigma\ [\ \hat{A}\ \overline{X}\ 1\ ]^T$$

where

$$\Sigma = \mathrm{Var}([\,\overline{X}\ \hat{A}\ \hat{B}\,]) = \begin{bmatrix} \Sigma_{XX} & 0 & 0 \\ 0 & \Sigma_{AA} & \Sigma_{AB} \\ 0 & \Sigma_{BA} & \Sigma_{BB} \end{bmatrix},$$

with

$$\Sigma_{XX} = \mathrm{Var}(\overline{X}),\ \Sigma_{AA} = \mathrm{Var}(\hat{A}),\ \Sigma_{BB} = \mathrm{Var}(\hat{B}) \text{ and } \Sigma_{AB} = \Sigma_{BA} = \mathrm{Cov}(\hat{A},\hat{B}).$$

Since the factors $\hat{A}$ and $\hat{B}$ are derived from the 1988 bridge assessments, and are consequently independent of the value $\overline{X}$ from the 1986 assessment, the covariances between $\overline{X}$ and $\hat{A}$ and $\overline{X}$ and $\hat{B}$ are zero.

Thus

$$\mathrm{Var}(\overline{X}_{adj}) \approx \hat{A}^2 \Sigma_{XX} + \overline{X}^2 \Sigma_{AA} + 2\,\overline{X}\,\Sigma_{AB} + \Sigma_{BB}. \qquad (4)$$

An estimate of Σ_{XX} comes by applying the jackknife technology (E. G. Johnson, 1987b) to the estimate $\overline{X}$. Since the factors $\hat{A}$ and $\hat{B}$ are each functions of the bridge sample means and standard deviations, estimates of Σ_{AA}, Σ_{AB} and Σ_{BB} can be obtained by expressing $\hat{A}$ and $\hat{B}$ in terms of the vector

$$\Psi = [\,\hat{\mu}_1,\ \hat{\sigma}_1,\ \hat{\mu}_2,\ \hat{\sigma}_2\,]$$

and applying the delta method to the result. This produces

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \approx \begin{bmatrix} \frac{\partial A}{\partial \Psi} & \frac{\partial B}{\partial \Psi} \end{bmatrix} \Sigma_{\Psi\Psi} \begin{bmatrix} \frac{\partial A}{\partial \Psi} & \frac{\partial B}{\partial \Psi} \end{bmatrix}^T \qquad (5)$$

where $\Sigma_{\Psi\Psi}$ is the 4×4 variance-covariance matrix of Ψ. (In $\Sigma_{\Psi\Psi}$, the covariances between the terms based on the bridge to 1986 and the terms based on the bridge to 1984 are taken to be 0, since these two bridges are independent samples.)

Estimates of the various terms in $\Sigma_{\Psi\Psi}$ can be obtained by the jackknife repeated replications technique. However, it is known (Mosteller & Tukey, 1977) that the jackknife procedure has relatively poorer performance in estimating the variance of a statistic with a markedly nonsymmetric distribution. Consequently, the jackknife estimates of the variances of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ would be expected to be of lower quality than the jackknife variance estimates of $\hat{\mu}_1$ and $\hat{\mu}_2$.

Since the jackknife performs better when the distribution of the statistic in question is symmetric, it is preferable to apply a symmetrizing transformation to the standard deviations $\hat{\sigma}_1$ and $\hat{\sigma}_2$, obtain the jackknife variance estimates of the transformed statistics, and reexpress (5) to account for the transformation. A transformation of a variance statistic which promotes symmetry is the Wilson-Hilferty cube-root transformation (Kendall & Stuart, 1977, p. 400).

Let

$$\Xi = [\,\hat{\mu}_1,\ \omega_1,\ \hat{\mu}_2,\ \omega_2\,]$$

where $\omega_1 = \hat{\sigma}_1^{\,2/3}$ and $\omega_2 = \hat{\sigma}_2^{\,2/3}$ are the Wilson-Hilferty transformed values of $\hat{\sigma}_1$ and $\hat{\sigma}_2$. Then

$$\hat{A} = (\omega_2 / \omega_1)^{3/2} \quad \text{and}$$

$$\hat{B} = \hat{\mu}_2 - \hat{\mu}_1 (\omega_2 / \omega_1)^{3/2}$$

so that

$$\begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \approx \begin{bmatrix} \frac{\partial A}{\partial \Xi} & \frac{\partial B}{\partial \Xi} \end{bmatrix} \Sigma_{\Xi\Xi} \begin{bmatrix} \frac{\partial A}{\partial \Xi} & \frac{\partial B}{\partial \Xi} \end{bmatrix}' \qquad (6)$$

where

$$\Sigma_{\Xi\Xi} = \begin{bmatrix} \Sigma_{\mu_1\mu_1} & \Sigma_{\mu_1\omega_1} & 0 & 0 \\ \Sigma_{\omega_1\mu_1} & \Sigma_{\omega_1\omega_1} & 0 & 0 \\ 0 & 0 & \Sigma_{\mu_2\mu_2} & \Sigma_{\mu_2\omega_2} \\ 0 & 0 & \Sigma_{\omega_2\mu_2} & \Sigma_{\omega_2\omega_2} \end{bmatrix}.$$

Since $\frac{\partial \hat{A}}{\partial \Xi} = \begin{bmatrix} 0 & -\frac{3\hat{A}}{2\,\omega_1} & 0 & \frac{3\hat{A}}{2\,\omega_2} \end{bmatrix}$

and $\frac{\partial \hat{B}}{\partial \Xi} = \begin{bmatrix} -\hat{A} & \frac{3\hat{A}\hat{\mu}_1}{2\,\omega_1} & 1 & -\frac{3\hat{A}\mu_2}{2\,\omega_2} \end{bmatrix}$,

from (3) we have

$$\Sigma_{AA} \approx (9/4)\,\hat{A}^2(\,\Sigma_{\omega_1\omega_1} / \omega_1^2 + \Sigma_{\omega_2\omega_2} / \omega_2^2\,)$$

$\Sigma_{AB} \approx -(9/4)\,\hat{A}^2\,\hat{\mu}_1\,(\,\Sigma_{\omega_1\omega_1} / \omega_1^2 + \Sigma_{\omega_2\omega_2} / \omega_2^2\,)$, (taking $\Sigma_{\mu_1\omega_1}$ and $\Sigma_{\mu_2\omega_2}$ to be zero), and

$$\Sigma_{BB} \approx \hat{A}^2\,\Sigma_{\mu_1\mu_1} + \Sigma_{\mu_2\mu_2} + (9/4)\,\hat{A}^2\,\hat{\mu}_1^2\,(\,\Sigma_{\omega_1\omega_1} / \omega_1^2 + \Sigma_{\omega_2\omega_2} / \omega_2^2\,).$$

Inserting these approximations into (4) produces the following estimate of the variance of the 1986 adjusted means:

$$\begin{aligned}\mathrm{Var}(\,\overline{X}_{adj}\,) \approx \hat{A}^2\,(\,\Sigma_{XX} + \Sigma_{\mu_1\mu_1}\,) + \Sigma_{\mu_2\mu_2} &+ \\ (9/4)\,\hat{A}^2\,(\overline{X} - \hat{\mu}_1)^2\,(\,\Sigma_{\omega_1\omega_1} / \omega_1^2 &+ \Sigma_{\omega_2\omega_2} / \omega_2^2\,) \qquad (7)\end{aligned}$$

and the standard error of the adjusted mean is the square root of this variance estimate.

An idea of the effect of using the square root of equation (7) as an estimate of the standard error of an adjusted mean, rather than the traditional estimate based on equation (3) can be obtained by comparing the two standard error estimates. Table E.1 does this for the standard errors of the 1986 adjusted mean scores for students of age 9, 13 and 17.

Table E.1

Comparison of Standard Errors for the 1986 Adjusted Mean Scores

	SE_7 *	SE_3 **	Ratio
Age 9	1.9	1.2	1.58
Age 13	1.6	1.0	1.60
Age 17	1.7	1.1	1.55

* Standard error computed using equation 7
** Standard error computed using equation 3

We see that the effect of acknowledging that the parameters in the equating function (1) are subject to variability is to multiply the estimate of the standard error of the population mean proficiency estimate by about 1.6. Viewed in another way, the traditional estimate of the standard error of an adjusted mean proficiency value may underestimate the standard error by a factor of 1.6 so that the variance estimate based on (7) is two and a half times the size of that based on (3). This is largely because the traditional variance estimate only considers the variance of $\bar{X}$ while the more proper variance is essentially the sum of the (appropriately scaled) variances of the three means: $\bar{X}$, $\hat{\mu}_1$, and $\hat{\mu}_2$.

REFERENCES

REFERENCES

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1988). *Who reads best? Factors related to reading achievement in grades 3, 7, and 11.* (No. 17-R-01) Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Baratz-Snowden, J. C., Rock, D., Pollack, J., & Wilder, G. (1988). *The educational progress of language minority children: Findings from the NAEP 1985-86 special study.* Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report.* (No. 15-TR-20) Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1988a). *The NAEP 1985-86 reading anomaly: A technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1988b). *Expanding the new design: The NAEP 1985-86 technical report.* (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Cochran, W. G., et al. (1954). *Statistical problems of the Kinsey report on the sexual behavior in the human male.* Washington, DC: American Statistical Association.

Cronbach, L. J., Gleser, G. C., Nanda, R., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Dossey, J. A., Mullis, I. V. S., Lindquist, M. M., & Chambers, D. L. (1988). *The mathematics report card: Are we measuring up? Trends and achievement based on the 1986 National Assessment.* (No. 17-M-01) Princeton, NJ: National Assessment of Educational Progress.

Ferris, J. J. (1988). Quality control of data entry. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 143-146). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Haertel, E. (Chair). (1989). *Report of the NAEP technical review panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level NAEP comparisons.* National Center for Education Statistics Technical Report CS 89-499. Washington, DC: U. S. Department of Education, Office of Educational Research and Development.

Hedges, L. V. (1989). The NAEP/ETS report on the 1986 reading data anomaly: A technical critique. In E. Haertel (Chair), *Report of the NAEP technical review panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level NAEP comparisons* (pp. 75-84). National Center for Education Statistics Technical Report CS 89-499. Washington, DC: U. S. Department of Education, Office of E‿ucational Research and Development.

Johnson, E. G. (1987a). Design effects. In A. E. Beaton, *Implementing the new design : The NAEP 1983-84 technical report* (pp. 545-562). (No. 15-TR-20) Princeton NJ: Educational Testing Service.

Johnson, E. G. (1987b). Estimation of uncertainty due to sampling variability. In A. E. Beaton, *Implementing the new design : The NAEP 1983-84 technical report* (pp. 505-512). (No. 15-TR-20) Princeton NJ: Educational Testing Service.

Johnson, E. G. (1988a). The NAEP populations and samples. In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 21-34). Princeton, NJ: Educational Testing Service.

Johnson, E. G. (1988b). Attributes of low-scoring students. In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 35-43). Princeton, NJ: Educational Testing Service.

Johnson, E. G. (1988c). Item data analyses. In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 65-77). Princeton, NJ: Educational Testing Service.

Johnson, E. G. (1988d). Mathematics data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 215-240). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Johnson, E. G., Burke, J., Braden, J., Hansen, M. H., Lago, J. A., and Tepping, B. J. (1988). Weighting procedures and variance estimation. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 273-291). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Johnson, J. R. (1987). Instrument and item information. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp. 119-134). (No. 15-TR-20) Princeton, NJ: Educational Testing Service.

Johnson, J. R. (1988). Booklet format and administration. In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 53-57). Princeton, NJ: Educational Testing Service.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics. Volume 1, 4th edition.* New York: MacMillan.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413.

Messick, S. J., Beaton A. E., & Lord, F. M. (1983). *NAEP reconsidered: A new design for a new era.* (NAEP Report 83-1) Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359-381.

Mislevy, R. J. (1988). Scaling procedures. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 177-204). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). (No. 15-TR-20) Princeton, NJ: Educational Testing Service.

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

Mullis, I. V. S., & Jenkins, L. B. (1988). *The science report card: Elements of risk and recovery. Trends and achievement based on the 1986 National Assessment.* (No. 17-S-01) Princeton, NJ: National Assessment of Educational Progress.

Mullis, I. V. S., & Jenkins, L. B. (1990) *The reading report card, 1971 to 1988: Trends from the Nation's Report Card.* Princeton, NJ: Educational Testing Service.

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

Schrader, W. B. (1988). [Review of NAEP 1986 database] In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 97-98). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1985) *M-Group: Estimation of group effects in multivariate models* [Computer program]. Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1988). The IRT linking procedure used to place the 1986 intermediary scaling results onto the 1984 reading calibration scale. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 555-565). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Sheehan, K. M , & Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures.* (ETS Research Report RR-88-38-ONR) Princeton, NJ: Educational Testing Service.

Slobasky, R. (1988). [Field administration factors and the 1986 NAEP reading scores] In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 99-101). Princeton, NJ: Educational Testing Service.

Wise, L. L., Chia, W. J., and Park, R. K. (1989). *Item position effects for test of word knowledge and arithmetic reasoning.* Paper presented at the annual meeting of the American Educational Research Association, March 1989, San Francisco.

Yamamoto, K. (1988). Science data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 243-255). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.

Zwick, R. (1988a). Block and booklet analyses. In A. E. Beaton, *The NAEP 1985-86 reading anomaly: A technical report* (pp. 79-86). Princeton, NJ: Educational Testing Service.

Zwick, R. (1988b). Reading data analysis. In A. E. Beaton, *Expanding the new design: The NAEP 1985-86 technical report* (pp. 207-212). (No. 17-TR-20) Princeton, NJ: Educational Testing Service.