Report of the Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons

Edward Haertel, Chair

Pascal D. Forgione, Jr., Chair,
Subpanel on State Comparisons

Herbert J. Walberg, Chair,
Subpanel on Anomaly and Trends

Janet Baldwin, R. Darrell Bock, Leigh Burstein, Dale Carlson,
Jeanne S. Chall, John T. Guthrie, Larry V. Hedges, Dan Melnick,
Mark D. Musick, Tej Pandey, William H. Schmidt, David E. Wiley

TABLE OF CONTENTS

Joan Boycoff Baron and Pascal D. Forgione, Jr., Collecting and Profiling School/Instructional Variables as Part of the State-NAEP Results Reporting: Some Technical and Policy Issues

Leigh Burstein, Technical and Policy Issues Related to Reporting of State Level NAEP in a Fair and Credible Manner

Edward Haertel, Within-State Comparisons: Suitability of State Models for National Comparisons

Dan Melnick, Evaluation for National and State-Level NAEP

INTRODUCTION

History and Charge of Panel

Last December, a Technical Review Panel was formed by the Center for Education Statistics to conduct an external review of the National Assessment of Educational Progress (NAEP). The panel was charged with examining three broad issues:

--The apparent lack of comparability between the findings of the 1984 and 1986 reading assessments

--The accuracy of NAEP trend data, particularly in reading and mathematics, and apparent inconsistencies between NAEP trend data and those from other major tests

--Problems and possible solutions in the expansion of NAEP to include a state-by-state assessment

The panel was organized into two subpanels to carry out its investigations. One subpanel, chaired by Herbert J. Walberg, addressed the issues of the 1986 reading anomaly and of the accuracy of NAEP trends. The other, chaired by Pascal D. Forgione, Jr., addressed issues in the expansion of NAEP to permit state-level reporting and comparisons. In addition to separate deliberations by the two subpanels, the entire group met to exchange views on all three issues, and to reach agreement on its recommendations and conclusions. The chair for the entire panel was Edward Haertel.

The panel held two-day meetings in December, January, and February, and a final, one-day meeting late in April. Based on discussions, data provided by the Educational Testing Service (ETS), interviews with ETS personnel, and other information, the panel formulated a set recommendations for the conduct of the NAEP,

designed to minimize the probability of a reoccurrence of the reading anomaly, to assure the accuracy and continuity of NAEP trends, and to address concerns that arise in the expansion of NAEP to provide state-level achievement estimates and comparisons.

Most members of the panel contributed individually authored papers addressing particular issues within the charge of the panel. To varying degrees, these papers reflect the results of discussions and deliberations by the panel as a whole, but each represents primarily a single author's position. This is as it should be. The members of the panel were deliberately chosen for their varied areas of expertise and their varied perspectives, and in their respective papers, each addressed areas in which she or he was especially well qualified.

Following the preparation of the separate papers, Dr. Haertel reviewed their findings and recommendations, and drafted the Report of the Panel, which summarizes and supports the findings presented in the separately authored papers. This draft was circulated to all of the panel members, and revised in response to the comments received. The Report addresses each of the panel's three charges in turn, summarizing major points from all of the relevant background papers, and from the panel's deliberations. At many points, the reader is referred back to the separate papers for more extended discussion.

This document is the final product of the panel's deliberations. It includes an Executive Summary, the panel's Recommendations and Conclusions, an Appendix in which one panel member qualifies her endorsement of these Recommendations and Conclusions, the Report of the Panel, and the separately authored papers.

EXECUTIVE SUMMARY

The NAEP Technical Review Panel was convened last December by the Center for Education Statistics and charged with examining three broad issues:

--The apparent lack of comparability between the findings of the 1984 and 1986 reading assessments (reading anomaly)
--The accuracy of NAEP trend data, particularly in reading and mathematics, and apparent inconsistencies between NAEP trend data and those from other major tests
--Problems and possible solutions in the expansion of NAEP to include a state-by-state assessment

The panel has produced a set of joint recommendations and conclusions and a report summarizing its deliberations. These are supported by individual papers on particular topics. The panel reached consensus on its recommendations, conclusions, and report, with the exception that one panel member had reservations concerning one recommendation and one conclusion (see Appendix to the Report).

The 1986 Reading Anomaly

While acknowledging that real declines in reading ability may have occurred, the panel was nearly unanimous in concluding that the bulk of the apparent declines in 9- and 17-year-old reading scores was probably artifactual. In reaching this conclusion, the panel concurred with the Technical Report on the anomaly by Educational Testing Service. The panel generally endorsed the ETS investigation of the anomaly, but criticized the almost exclusive focus on declines in mean scores. More attention should have been paid to the substantial increases from 1984 to 1986 in the variances of score distributions at all three age/grade levels. The panel also criticized the ETS report for considering possible hypotheses in isolation from one another, when a combination of two or more might easily have explained the bulk of the observed score declines. The panel also suggested some additional hypotheses that may merit consideration.

The Accuracy of NAEP Trends

The panel concluded that despite its imperfections, NAEP is a better barometer of national achievement trends than any available alternative. The only other national, longitudinal achievement data

collected over a comparable span of years come from college admissions tests, and these cover relatively limited domains of content, test only at the high school level, and examine self-selected groups of students that are not nationally representative of their age cohorts. At the same time, the panel concluded that the quality of NAEP trend reporting could be improved considerably.

The panel's three principal recommendations for improving the accuracy and authoritativeness of NAEP trends were first, to assure greater consistency over assessment cycles in the objectives covered; second, to assure greater care in revising the format of assessment materials or testing sessions, or any other aspects of NAEP procedures that might impact the continuity of NAEP trends; and third, to provide for an ongoing statistical evaluation and audit of NAEP data collection and reported findings, independent of the NAEP contractor.

State-Level NAEP

State-level assessments should be managed by a separate program unit within the National-NAEP organization, and should be parallel to the National-NAEP in most respects. In designing state-NAEP procedures, comparability among states and between state-level and national data is paramount. This implies a centralized administration plan. State samples should be drawn by the same contractor responsible for national-NAEP samples, and the National-NAEP organization should be responsible for training state-NAEP examiners, probably personnel provided by states for the 6- to 8-week period of training and data collection. The 1990 and 1992 pilots authorized by the Hawkins-Stafford law should be used to explore alternative administration procedures. The panel recommended an expanded NAEP data collection, covering more learning outcomes and more background questions. Results at both national and state levels should be reported at a greater level of specificity. Sufficient data should be collected from each student to permit accurate estimation of individual scores (although anonymity would be preserved). The amount of individual student time devoted to NAEP should be expanded to accomodate these changes. State-level results should be released promptly. In addition to reports on absolute levels of achievement, state results should be referenced to the performance of comparable states, national samples of students matched to state characteristics, or in other ways that account for demographic differences. A variety of comparison methods should be explored and reported in 1990 and 1992.

NAEP Technical Review Panel
Recommendations and Conclusions

May 21, 1988

RECOMMENDATION 1

The frameworks that have been used to organize NAEP objectives are inadequate in terms of comprehensiveness, specificity, and stability over assessment cycles. Knowledge and skills assessed should be drawn from an explicit, comprehensive, detailed, and stable domain. Content changes over assessment cycles should be specified in terms of this domain, and should be undertaken only after careful evaluation.

Each organizing domain would include descriptions of the knowledge, skills, or other possible learning outcomes that might be intended in a content area such as reading or mathematics. Domains would almost certainly encompass more than the range of learning outcomes represented by present NAEP exercise pools, and there would be no expectation that future assessments would necessarily attempt coverage of all of the particular learning outcomes within a domain. Domains would provide the basis for the more detailed reporting of NAEP results proposed under Recommendation 8.

NAEP exercises would be referenced to these domains. However, the domains would not be simple classification schemes for exercises, nor would they specify particular forms of test items corresponding to different learning outcomes. Test item responses are a consequence of more than one skill or ability, some intended and some not. The linkage of NAEP exercises to a skill domain will involve significant issues of item validity, including exercise formats appropriate for respondents at different ages, and more generally, the ancillary skills exercises may require. These ancillary skill requirements can reduce the validity of exercises as indicators of the learning outcomes they are designed to measure.

RECOMMENDATION 2

NAEP should broadly assess important learning outcomes. This implies increased emphasis on higher level learning outcomes now considered critical for all students. Priorities should be guided by expert subject matter perspectives as well as current substantive and methodological research.

A major purpose of NAEP is to inform and focus discussion of education policy and practice. It follows that, at both the national and state levels, NAEP must represent a full and rich conception of important cognitive learning outcomes in the domains assessed. These include the learnings covered by the typical curricula of the Nation's schools, but should reach beyond the typical. The content of NAEP must not be reduced to the intersection of the several states' curriculum frameworks, nor to any "lowest common denominator." Assessment objectives should reflect the best thinking of subject matter specialists, and should focus on emerging views of learning and knowledge in all content subjects, such as history and literature. The NAEP objectives should lay a solid foundation for discussion among professionals and practitioners about such issues as curriculum, teaching, inservice education, school organization, and policy alternatives available to state leaders.

The design of NAEP exercises should capitalize on methodological advances to assure the valid assessment of complex, higher level learning outcomes as well as important factual knowledge, basic skills, and other lower level outcomes usually achieved earlier and considered prerequisite for higher level learnings. Attention should be given to the measurement of processes such as reading, writing, and problem solving in the context of content subjects. Finally, various modes of assessment should provide ample opportunity for students to display their productive abilities. This will require increased emphasis on writing, speaking, and interacting in both real-world and school tasks.

RECOMMENDATION 3

State-level NAEP should collect information on student and teacher background factors and on schooling processes as well as achievement. In particular, a core set of student background questions should be included on both national-level NAEP and state-level NAEP. State-level NAEP should also explore the feasibility of collecting information on opportunity to learn and other schooling processes possibly linked to achievement. Questions should be used over a series of assessments so that trends can be observed.

NAEP achievement data become more meaningful and more useful if they can be linked to carefully chosen school, teacher, and student characteristics, and schooling processes; as well as community, home, and family

characteristics, and students' out-of-school activities. Judicious selection of background questions is essential. Priority should be given to those that are (a) important for describing patterns in NAEP achievement data, (b) plausibly related to achievement, or (c) reflective of other valued schooling outcomes.

Background questions may be addressed to students, teachers, and principals. In general, the same questions should be used in state-level NAEP and in national-level NAEP, and arbitrary changes in background questions from one assessment cycle to the next should be avoided. Changes may be necessary to assure coverage of schooling process variables important to particular content areas assessed concurrently. Information about changes in curriculum and instructional methods can be critical to the interpretation of NAEP trends. If a reliable, efficient, and unobtrusive method can be found for collecting data on students' opportunity to learn the content of concurrently administered achievement exercises, this information might be especially valuable.

RECOMMENDATION 4

The amount of individual student time devoted to the conduct of NAEP should be expanded.

Recommendations 2 and 3 call for increased amounts of information about individual examinees. In order to accomplish this, increased time is required for responding to achievement exercises and survey questions on student background and experiences.

RECOMMENDATION 5

Assessment procedures impacting the continuity of national trends should not be altered unless there is a compelling reason to do so. Changes should be made only after systematic examination of the likely consequences and justification in terms of NAEP priorities. Old and new procedures should be carried in parallel for at least one assessment cycle, on a scale sufficient to assure continuity of national trends.

In The Nation's Report Card, Alexander and James place the highest priority on the maintenance of continuity in the trend lines of NAEP achievement assessments (p. 7). We strongly concur with this priority. Undoubtedly, there will be profound changes in future NAEP data collections, especially in light of recommendations to extend NAEP to permit state-by-state comparisons. However, whatever modifications are made in the overall program design, it is mandatory that the procedures used to collect the data for national trend estimates be parallel in every important respect. During transitions when old and new procedures are carried in parallel, not

only the assessment exercises themselves but also the data collection procedures should remain the same. Only this will assure that scale scores with the same meaning as those available before 1986 can be calculated for 1988 and beyond.

We strongly endorse the current (1988) ETS replications of the 1984 and 1986 procedures as a vital source of information for future design decisions. However, we do not believe that ad hoc investigations are sufficient to assure continuity of NAEP time series. A new process should be developed to ensure adequate and systematic evaluation of proposed procedural changes. The technical advisory process to NAEP should comprehensively incorporate considerations of procedural design and audit as well as sample design and analysis. This implies formal review of on-site administration conditions and procedures, instructions and student conformity to them, etc., as well as timing and booklet design. This also implies that the technical advisory body(ies) should be composed of individuals representing all relevant areas of expertise.

RECOMMENDATION 6

NAEP data must be collected so as to assure comparability across states. Sampling procedures, core instrumentation, and conditions of administration must be uniform. Although states may choose to augment their data collection, the minimum design must be sufficient to provide comparable estimates of achievement levels and distributions for each state. A centralized administration plan will best serve the ends of comparability. As part of the pilot assessments authorized by the Hawkins-Stafford law, NCES should study the effects of alternative administration procedures on comparability.

State-level NAEP samples should be drawn following the same procedures as for the national sample, with the possible exception that schools may be drawn as the primary sampling units rather than being clustered within counties or groups of counties. The present practice of returning questionnaires on excluded students should be continued, and the percent of students excluded should be reported along with state assessment results.

For those states that wish to participate in the assessment more extensively, one or more options should be developed for an expanded assessment, which might include an expanded student sample, additional background or achievement questions, or both. The national core instrumentation must precede any supplementation.

RECOMMENDATION 7

Scores of individual students should be estimated and made available for analysis. However, consistent with confidentiality restrictions in the law, particular students shall not be identified.

Accurate score estimates for individual examinees would permit NAEP to report the estimated score distributions referred to in Recommendation 9 below, and would greatly simplify other analyses by the NAEP contractor as well as secondary users. In recent National Assessment data collections, the way in which the Balanced Incomplete Block (BIB) spiraled booklet assignments were designed made it impossible to generate accurate score estimates for individual examinees using conventional Item Response Theoretic (IRT) procedures. In future assessments, the data obtained from students should permit the accurate estimation of their individual performance levels. Individual scores could be obtained using either a BIB-Spiraled design or alternative designs.

RECOMMENDATION 8

NAEP results should be reported at a greater level of specificity. This reporting should permit distinctions among important parts of the domain. This will imply the use of multiple scores or scales within domains.

The skill domains discussed in Recommendation 1 must comprise important and conceptually distinct core components. NAEP should, in terms of these components, identify the subdomains on which trends in achievement will be reported. In the past, although a consensus approach for defining objectives has been followed for each assessment, little attention appears to have been paid to the continuity of consensus over time. Appropriate subdomain specification requires a stable skill domain and therefore this recommendation is critically dependent on the domain specification of Recommendation 1.

We distinguish four decision processes. The first, discussed above, is the specification of a comprehensive and stable domain of skills and knowledge. A second is the selection from that domain of those learning outcomes to be included in a particular assessment. The third is the specification of how these domain components will be grouped for the calculation of subdomain scores. A fourth decision process, also referred to under Recommendation 1, is the specification of the relationship between NAEP exercises and the learning outcomes to be assessed. We are recommending that each of these decision processes be formalized and articulated.

RECOMMENDATION 9

NAEP should extend the systematic reporting of distributions of achievement as well as average levels. The impact of changes over assessment cycles in society, in schooling, and in NAEP procedures on these distributions should be routinely evaluated.

NAEP reporting has been concentrated on averages or central tendency. Some aspects of the Nation's educational attainment are better informed by an examination of the entire distribution of scores. For example, changes over time in average scores may have quite different implications depending upon whether all or only a part of the score distribution is changing. To assist in the examination of changes in distribution, NAEP should report trends for important quantiles of the score distribution. Further consideration should also be given to other methods of representing changes throughout the score distribution.

To assist in interpreting changes in score distributions, it is desirable to isolate the demographic subgroups that contribute to changes in distributions. This may require collection of additional information on schools, students, and variations in administrative conditions. The 1988 report by Beaton, et al. on the 1986 reading anomaly includes a partitioning analysis that reveals the relative contributions to score declines of changes in the mixture of student subgroups and of changes in performance by particular subgroups. This partitioning analysis focuses exclusively on changes in mean performance levels. Such an analysis should be done on a routine basis, and should be focused on the full distribution, not just the means.

RECOMMENDATION 10

The expansion of NAEP to provide data at the level of individual states will entail careful study of methods for making and reporting state comparisons. In the 1990 and 1992 pilot studies, a variety of methods should be explored and reported.

Where feasible, state results should be reported for major process and content categories, using the same proficiency scales as are used for national-NAEP. In many content areas, age-specific proficiency scales may be more useful and appropriate than scales spanning different age/grade levels. In addition to reporting absolute levels of achievement on these scales, each state's performance might be referenced to that of a small group of comparable states, or to nationally representative samples of students matched to state population characteristics. Additional alternatives may also be explored and reported.

RECOMMENDATION 11

The reporting of cross-sectional and trend results for state-level NAEP should characterize both the level and distributions of student attainment within each state. This reporting should include (a) demographic subgroup and community differences, (b) variation in performance across major domains of learning outcomes, and (c) distributions of school-level performance within the state.

As discussed in Recommendations 8 and 9, reporting score distributions for major subdomains is more informative than reporting means for broad content areas. This is true at the state level as well as the national level. State and national score distributions for major subdomains should be reported in ways that facilitate their direct comparison to one another.

In addition to distributions for entire states, performance should be reported for demographic subgroups and types of communities within states, whenever such reporting is feasible. Feasibility may be limited by smaller sample sizes for groups or areas within states, or by the legal requirement that results not be reported for schools or districts in the 1990 and 1992 pilot assessments.

Because the school is an important locus of educational policy, we recommend that distributions of school means as well as distributions of individual scores be reported. Where samples of schools are sufficiently large and representative, distributions of school means should be reported for states, and for different types of schools within states. By law, particular schools would not be identified.

RECOMMENDATION 12

Evaluation of NAEP results, and in particular the sources and magnitudes of errors in estimated achievement distributions, should be undertaken routinely, and not just in response to anomalous findings. Funding should be assured for an ongoing NAEP evaluation, in some way independent of the conduct of the assessment.

To achieve its goals, NAEP should contain a strong evaluation component. This should involve experiments embedded in NAEP that would provide a basis for empirically resolving outstanding issues. The evaluation should include a program of basic and applied research to identify sources of error and model relationships among them. NCES should report NAEP errors on a regular basis rather than limiting their investigation to apparent anomalies. In conducting the evaluations, particular attention should be paid (a) NAEP's sensitivity to alternative administration procedures, (b) the impact of saliency of assessment results on individual student performance, (c) the consistency of

NAEP results with other measures of achievement, (d) methods of norming NAEP to relate it to actual performance, (e) the influence of curricular decisions on NAEP outcomes with particular attention to the problem of "teaching to the test", and (f) year-to-year consistency.

CONCLUSION 1

The anomalous declines in the estimated reading abilities of 9- and 17-year-olds found by NAEP between 1984 and 1986 are much larger than improvements or declines over comparable past periods. After careful study of available evidence, the panel was not able to identify the particular reasons for the reported drop in NAEP reading scores. However, we believe that the most likely primary causes of declines so large and so rapid are changes in testing procedures, and that the 1986 assessment results do not accurately reflect declines in the real abilities of American pupils. Real declines in reading ability may have occurred, but their magnitudes are likely obscured by factors which do not validly reflect changes in pupil learning. The primary causes of the observed decline are still unclear, although we believe that they are probably located in modifications of assessment procedures between 1984 and 1986. New studies currently being conducted by NAEP should help clarify the extent to which such procedural changes were responsible.

CONCLUSION 2

We believe that differences in college entrance examinations versus NAEP in (1) the populations represented by those tested and (2) the content tested are large enough so that reported trends in college entrance examination scores are not directly comparable to those from NAEP. NAEP was established to serve as the most accurate barometer of achievement for America's young people. Despite its imperfections, we believe that it has and will continue to serve this function better than college entrance examination scores.

APPENDIX

Qualification of Endorsement of Recommendations and Conclusions by Jeanne S. Chall

Comments regarding Recommendation 2 as worded in the May 7 memo
(Use only if the "essence" of my suggested changes cannot be made)

I dissent from Recommendation 2 in the May 7 memo because it places almost total emphasis on testing higher level learning outcomes. If NAEP focuses primarily on higher level learnings, the influence of the "lower" on the "higher" learnings will be difficult to determine.

For understanding the course of development in learning such important skills as reading, it would be helpful to assess carefully and specifically the "lower" and "middle," as well as the "higher level" learnings, and also the school, home, and community conditions that affect them.

Jeanne Chall
5/17/88

Statement on Conclusion 1

I dissent with conclusion #1 for the following reasons:

1. If the 1986 reading scores stem from anomalies in testing procedures, one might expect similar declines across all ages tested — 9, 13 and 17. Yet while the 9 and 17 year olds dropped considerably in 1986 over comparable past periods, the 13 year olds did not.

If the "most likely primary causes" for the 1986 decline are the testing procedures, then it would be necessary to show how the "changed" testing procedures in 1986 affected only the 9 and 17 year olds and not the 13 year olds.

2. At the present state of the inquiry into the 1986 reading score declines, I find it difficult to agree with the following statement in Conclusion 1: "the 1986 assessment results do not accurately reflect declines in the real abilities of American pupils" (memo of May 7). As far as I know, no analyses have so far been undertaken to warrant such a statement. While some of the decline may have resulted from changes in testing procedures, it is premature at this point to say that the scores do not accurately reflect the real abilities of American pupils.

Since the possibility of a real decline was considered by only one member of the committee, the following statement is also questionable: "Real declines may have occurred, but their magnitudes are likely obscured by factors which do not validly reflect changes in pupil learning" (memo of May 7). It would be more reasonable to state that procedural effects

Statement on Conclusion 1, cont'd.

were found but that they do not rule out changes that may have taken place in pupil learning.

See in this connection, "Could the Decline Be Real?," the individual report for the subcommittee on the 1986 reading score declines. Based on analyses of NAEP reading trends, it was hypothesized that the increases and decreases in NAEP reading scores from 1970 to 1980, from 1980 to 1984, and from 1984 to 1986 could be explained, in part, by the strengthening and weakening of reading instruction provided by schools, particularly in the early grades, by remedial support when needed, and by support from home and community. When school instruction and support are provided, the scores rise for 9 year olds (as ~~in the 1970's~~ they did from 1970 to 1980), and they tend to hold up when the same students are tested at ages 13 and 17. When instruction is not as strong in the early grades (as for those tested in 1984 ~~and 1980 students~~), the scores tend to decline and will probably remain low when students reach ages 13 and 17, unless additional measures are undertaken.

The decline in the 1986 reading scores as compared to the 1984 scores tends to follow these trends. While changes in testing procedures may have resulted in the large declines, the possibility of a real decline cannot be dismissed since the 1986 reading scores follow similar trends as those for ~~the~~ 1980 and 1984.

For a fuller explication of the hypothesis that the 1986 declines may be real, see "Could the Decline Be Real? Recent Trends in Reading Instruction and Support in the U.S.," paper prepared for the Subcommittee on the 1986 Reading Data of the NAEP Technical Review Panel. See also

Statement on Conclusion 1, cont'd.

"Literacy: Trends and Explanations," Educational Researcher, November 1983, pp. 3-8; "New Reading Trends: The NAEP Report Card," Curriculum Review, March/April 1986, pp. 42-44; and "School and Teacher Factors and the NAEP Reading Assessments," paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The Nation's Report Card, August, 1986 (ERIC Document ED 279 667).

Jeanne Chall
5/17/88

Report of the Panel

Edward Haertel, Chair

Report of the Panel

Table of Contents

REPORT OF THE PANEL

Edward Haertel, Chair

INTRODUCTION

The National Assessment of Educational Progress (NAEP) is the only regularly conducted national survey of achievement at the elementary, middle, and high school levels. For the past two decades, it has provided periodic assessments of student proficiencies in reading, writing, mathematics, science, and social studies, and less frequently, citizenship, computer literacy, history, literature, art, music, and career development. In addition to charting patterns and trends in student achievement, NAEP has collected background information that has helped to chronicle changes in educational conditions and practices. NAEP data have been provided freely to researchers interested in conducting secondary analyses, and have supported studies in curriculum, educational policy, methodological research, and research on educational productivity.

As discussed in Dr. Walberg's background paper, the purposes and methods of the National Assessment have evolved over time, and may change even more dramatically in the future. NAEP has evolved in response to new needs and purposes, and to capitalize on new methodologies for data analysis and reporting. With the signing of the Hawkins-Stafford law and the advent of state-level reporting and comparisons, more is expected of NAEP today than ever before. The next several years will bring significant changes in the assessment.

Through its consideration of the 1986 reading anomaly, the accuracy of NAEP trend data, and the challenges that will arise in

expanding NAEP to provide state-by-state results and comparisons, the Technical Review Panel has come more than ever to regard the National Assessment of Educational Progress as an invaluable national resource. Problems and deficiencies in NAEP have been identified, but these can be remedied. As it enters its third decade, NAEP is positioned to serve better than ever before as "The Nation's Report Card."

FINDINGS REGARDING THE READING ANOMALY

Introduction

One of the most frequently assessed and carefully attended areas assessed by NAEP is reading. The reading abilities of 9-, 13-, and 17-year-olds were assessed in 1971, 1975, 1980, 1984, and most recently in 1986. Findings from the most recent reading assessment appeared strikingly different from those of earlier assessments. As stated in the NAEP report Who Reads Best (Applebee, Langer, & Mullis, 1988, pp. 56-57), "The results of the 1986 reading assessment seemed to be out of line with previous NAEP reading assessment results. In particular, they indicated precipitous declines in average reading proficiency at ages 17 and 9. The nature of these drops across only a two-year period, taken in the context of only modest changes in reading proficiency across a succession of four-year periods since 1971, was simply not believable."

Declines in scores at both age 9 and age 17 were pervasive, affecting both boys and girls in all geographic regions, racial and ethnic groups, and types of communities. The magnitudes of declines were slightly greater among traditionally lower-performing

subgroups, including blacks and hispanics, children in disadvantaged urban areas, children whose parents have less education, and children who were themselves below the modal grade level. Scores declined more in the southeastern and western regions of the country than in the northeast and central regions.

The absolute magnitudes of the declines at ages 9 and 17 were quite small. Declines in scaled score values were about 3 percent of the 1984 values, reflecting declines in the overall percent correct on reading items of about 3.6 percent for 9-year-olds and 3.3 percent for 17-year-olds. There was a slight improvement in the average scores of 13-year-olds, and there were no concomitant changes from earlier assessments in the 1986 science or mathematics assessments, which were conducted concurrently with reading.

At the same time as the reading means declined at ages 9 and 17, there were striking increases in the dispersion of scores at all three age/grade levels. The standard deviation of reading proficiency scores for 9- and 13-year-olds increased about 10 percent over 1984 values, and at age 17 the increase was about 25 percent. At all three grade levels, the proportions of students at both the highest and the lowest score levels increased from 1984 to 1986. As Hedges observes in his paper, an adequate explanation for the anomaly must explain both the changes in means and the changes in variability.

We invited staff of the Educational Testing Service (ETS) who are working on NAEP to meet with us and discuss the anomaly; we reviewed the ETS technical report on their investigations of the anomaly (Beaton, Ferris, Johnson, Johnson, Mislevy, & Zwick, 1988); and we examined additional materials, including the actual reading

exercises on which the declines occurred, detailed statistical tables not included in the Beaton, et al. report, and trends derived from other data sources.

Two of the panel's background papers, by Hedges and Chall, focused on the reading anomaly. Hedges has provided a technical critique of the ETS report on the anomaly. Chall presents arguments that real declines may well have occurred due to changes in reading curriculum and instruction. The background paper by Wiley also provides analyses that helped to inform the panel's conclusions, and Schmidt's paper touches on related concerns. Walberg's observations should also be noted, that the reading anomaly hardly implies that national and state assessments are unmanageable. Too many well intentioned procedural changes may have been made too quickly, but as pointed out by Chall and others, actual declines have certainly not been ruled out. More systematic consideration of even apparently minor procedural changes should make future anomalies much less likely, so that unusual performance changes can be more confidently attributed to real changes in ability.

The Panel's Conclusions Concerning the Anomaly

As described in Wiley's paper in this volume, the possible explanations for distributional changes fall into three categories: (a) methodological artifacts, (b) changes in population (e.g., increases in the relative sizes of traditionally low-scoring groups), and (c) changes in student learning. Population changes occur slowly, and would be unlikely to lead to substantial changes over just two years. In any case, Beaton's analysis of declines by subgroup (Beaton, et al., 1988, Chap. 6) appears to rule out population changes as the cause of

the 1986 score declines. Such rapid and dramatic changes in student learning--either in school or out of school--also appear quite unlikely in the absence of dramatic, simultaneous program changes in schools across the nation. School curricula and instructional practices do evolve, but such changes are usually gradual, and seem unlikely to have resulted in massive score declines over just two years. This analysis leaves methodological artifacts as the most likely primary cause of the observed declines. Specific methodological hypotheses are discussed below.

We were unable to determine the cause of the 1986 reading score declines from the available evidence, but with the exception of Dr. Chall, we agreed that the most likely primary causes were procedural, and that, although real declines in reading ability may have occurred, their magnitudes are likely obscured by the effects of changes in testing procedures. Dr. Chall explains her reasons for dissenting from this conclusion in her statement following the Panel's Recommendations and Conclusions in this volume, and in her paper. Her argument is also summarized below, in the section on Evidence for an Actual Decline in Reading Scores.

Technical and Procedural Explanations for the 1986 Reading Anomaly

In response to the 1986 reading anomaly, ETS systematically developed and examined a number of possible explanations. These investigations and their results are described in The NAEP 1985-86 Reading Anomaly: A Technical Report by Beaton, et al. (1988). In his paper, Larry Hedges undertook a systematic technical critique of the ETS report, and raised four general criticisms. First, the report focuses almost exclusively on changes in means, largely ignoring

changes from 1984 to 1986 in the shape and variability of achievement distributions at all three age levels. Second, the ETS investigation was organized around the idea that the 1986 reading declines resulted from some single effect, giving little attention to the fact that in combination the different effects considered might well account for the observed declines. Third, the 1984 results were taken as a valid base line against which to judge the magnitude of declines, largely ignoring the possibility that 1984 results were inflated, and focusing attention almost exclusively on the 1986 assessment as the locus of possible problems. Finally, the ETS analyses may have overstated the precision of the NAEP results. A more complete accounting of sources of error in both the 1984 and 1986 results might have made the 1986 declines appear less remarkable.

Changes in the the variance and shape of achievement distributions, as well as means. Changes in the variability in reading achievement scores at all three grade levels were at least as striking as changes in the means. The standard deviation of reading proficiency scores for 9- and 13-year-olds increased by about 10 percent and for 17-year-olds the increase was about 25 percent. The proportions of very high-scoring pupils at all three age levels were actually slightly larger in 1986 than in 1984, but the proportions of very low-scoring pupils were considerably larger among 9- and 17-year-olds. The changes were more complex than simple shifts in means and increases in dispersions. The upper tails of score distributions in 1984 and 1986 are quite similar, but the lower tails of the 1986 distributions are heavier, suggesting declines among

some of the students who had been scoring about average.

In his paper in this volume, Wiley presents tabulations made from unpublished ETS data of the score levels corresponding to a series of percentiles in 1984 and again in 1986, for each of the three age groups. He finds that at sufficiently low percentile ranks, there were declines at all three ages, and at sufficiently high percentile ranks, there were increases. The "crossover" point at which 1984 and 1986 scores were the same was at roughly the 78th percentile for 9-year-olds and the 75th percentile for 17-year-olds. For 13-year-olds, the crossover was below the 10th percentile. Thus, the median scores declined at ages 9 and 17, and increased at age 13, reflecting the pattern shown by the means. Wiley's tabulations highlight distributional changes from 1984 to 1986 that are common to all three age/grade groups, and reinforce the importance of attending to changes in the shapes as well as means of the score distributions.

Separate hypotheses that together could account for anomalous declines. The hypotheses ETS was best equipped to investigate concerned their own procedures for processing the data after they were collected. Failures of quality control and artifacts of scaling were thoroughly investigated, and we concur in the conclusion of ETS personnel that these types of problems are very unlikely to have caused the 1986 reading declines. The investigations undertaken by ETS also appear to largely rule out gross problems in sampling or weighting. There are, however, several classes of explanations that merit closer attention, as discussed in Hedges's paper.

By and large, nine-year-olds were assessed slightly earlier in

the school year in 1986 than in 1984, a difference amounting to an average of 22 days. Seventeen-year-olds were also assessed earlier (18 days) and thirteen-year-olds were assessed slightly later than in 1984 (4 days). Given that time in school is probably more important than chronological age in determining reading performance, and given that achievement growth is probably nonlinear over the course of the school year, Beaton, et al. may have substantially underestimated the possible magnitude of the time of testing effect, especially among nine-year-olds.

The hypothesis that declines reflected administration difficulties in some but not all schools was investigated by Beaton, et al., but more systematic investigation would be desirable. A search for outliers in the distribution of school means suggested that the anomaly could not be accounted for by difficulties at just a few isolated sites, but more pervasive problems related to overall management of the data collection, especially increases in the size of testing sessions for 17-year-olds, were not thoroughly investigated. At age 17, the average size of the groups tested increased from 20 in 1984 to 35 in 1986, and the maximum permissible session size was increased from 200 in 1984 to 250 in 1986. A comparison of variance components at the school versus individual level in 1984 and in 1986 might have been more informative concerning the overall effects of these changes. If the use of large testing sessions (often at the insistence of school personnel) or disruptions of these sessions were correlated with student background, this effect might have led to the observed pattern of declines for different student subgroups, as well as the overall increase in score variability at age 17.

Changes in administration conditions, including the design of exercise booklets, the mix of different content areas assessed, the timing of exercise blocks, the sequence of activities carried out in testing sessions, and the size of the testing sessions, may have accounted for a substantial portion of the test score declines. Hedges notes in particular that the 1986 assessment used a "fill in the oval" format for responses which were then machine scored, while the 1984 assessment used a "circle the letter" format for responses which were then entered via keyboard. Any effect of this change would be expected to operate at age 13 as well as ages 9 and 17, but other effects may have served to increase 13-year-olds' scores, masking negative effects at that age level. Difficulties with "fill in the oval" may have been greatest for younger children and for traditionally low-performing subgroups, contributing to the increased variance of 9-year-olds' scores as well as the observed patterns of score declines across subgroups.

In testing sessions for 9-year-olds, the initial block of background questions was increased from 6 minutes of testing time in 1984 to 15 minutes in 1986, and according to a memorandum from WESTAT (Beaton, et al., 1988, Appendix C), "field staff reported that many of the 9-year-olds were frustrated by the amount of time spent on the block of background questions ... Perhaps their concentration or motivation for the reading section was affected." Frustration with the task of providing background information may have been greater among traditionally low-performing subgroups, which would have helped to account for both the pattern of score declines across groups and the increased variance of scores in 1986.

Taken together, if these and other effects are approximately additive, several of them could jointly explain most or all of the anomaly. Hedges summarizes these effects in his Table 1. Hedges also considers the possible magnitudes of these effects for 13-year-olds, and finds that the effects of date of assessment, changing patterns of nonresponse, and artifacts associated with scaling all would have tended to increase scores of 13-year-olds while contributing to declines at ages 9 and 17.

Possibility of positive bias in 1984 assessment results. By focusing almost exclusively on the 1986 assessment, Beaton, et al. (1988) may have overlooked factors that led to inflated score estimates in 1984. Any such inflation would, of course, magnify the apparent decline in 1986. In 1984, 9- and 13-year-olds were assessed on reading and writing together. The WESTAT memorandum (Beaton, et al., 1988, Appendix C) mentions that both reading and writing exercises in 1984 were self-paced, and there were some reports that children spent more than the allotted time on the reading items and less than the allotted time on writing. In 1986, reading exercises were again self-paced, but were administered concurrently with mathematics and science exercises paced by tape recorder. Thus, children were constrained to spend no more than the allotted time on reading. At age 17, Hedges notes that 1984 scores were higher than the trend of earlier assessments would have indicated. If the 1984 results were simply extrapolated from the 1971-1980 linear trend, more than 25 percent of the 1986 anomaly would disappear.

The 1984 assessment results for 17-year-olds may have been

inflated if, in response to the widespread adoption of minimum competency testing requirements and the general increase in academic rigor in the early 1980s, there was a temporary increase in the dropout rate, leaving fewer low-scoring 17-year-olds in school. Such an effect would also have tended to reduce the variance of the score distribution in 1984, exaggerating the apparent increase in score variability in 1986.

Accuracy of NAEP achievement estimates. The 1984 to 1986 declines in reading scores at ages 9 and 17 are clearly much too large to be explained by the statistical sampling of respondents or by measurement error. However, such random effects may have contributed to the apparent decline, and it is important to estimate their probable magnitudes. Hedges notes that the standard errors estimated and reported in the Beaton, et al. (1988) are cross-sectional, and do not reflect all of the sources of random error that may have contributed to the apparent score declines.

Evidence for an Actual Decline in Reading Scores

A balanced consideration of the 1986 NAEP reading score declines requires consideration of the likelihood that some or even all of the observed score declines at ages 9 and 17 are the result of real declines in the reading abilities of American school children. If logical or empirical support for real achievement declines exists, it may be that the *absence of* a score decline at age 13, rather than the presence of declines at ages 9 and 17, constitutes the anomaly. In her paper, Dr. Jeanne Chall refines and extends her analyses of trends through earlier NAEP reading assessments (Chall, 1983, 1986a, 1986b) and presents arguments in support of this position.

Before turning to these arguments, it should be noted that the panel as a whole did not take issue with Dr. Chall's arguments. Her paper presents what may be plausible arguments for the direction of changes in pupil abilities from 1984 to 1986, but in the judgment of the majority of panel members the magnitudes of these declines over a period of just two years are larger than would be expected from changes in curriculum and instruction alone. For that reason, most panel members concur that the primary causes of the declines are probably located in modifications of assessment procedures between 1984 and 1986.

Performance trends may be expressed on the NAEP reading proficiency scale, with an effective range of about 100 to 400. Up to 1984, the largest gains between successive assessments, expressed as points per year on the reading scale, were +.9, +.5, and +.8 point at ages 9, 13, and 17, respectively. The largest declines to 1984 were -.5, -.1, and -.06 points per year. In contrast, the annualized changes from 1984 to 1986 were -2.8, +1.2, and -5.7 points at these three age levels. Conclusion 1, which was endorsed by the remaining panel members, explicitly acknowledges that real declines may have occurred, but holds that their magnitudes are likely obscured by factors related to changes in the booklet design, administration procedures, or related factors.

Influence of early reading instruction. Children learning to read pass through a series of different stages. At different stages, they profit most from different kinds of informal experience and formal instruction. In particular, a too-early school emphasis on comprehension and inference, before children have acquired sufficient

skill in phonics and other fundamentals, may be of little value. If such instruction takes teaching time away from word recognition, phonics, and the reading of stories and other connected texts, it may even be detrimental.

Chall argues that beginning in the late 1960s and continuing through the 1970s, beginning reading programs were stronger than before or since. The 1970s were a time of earlier formal instruction in reading, more challenging reading textbooks grade for grade, earlier and more systematic teaching of phonics, Sesame Street and the Electric Company, more remedial help for those who needed it, and Head Start. Since that time, funding for remedial reading instruction has declined, and Dr. Chall contends that a misplaced emphasis on comprehension and inference at early grade levels ("meaning-emphasis") has led to less time for connected reading--and to declining scores. She also cites evidence for the continued importance of early reading instruction to later reading performance.

Patterns of findings from earlier NAEP reading assessments are consistent with the hypothesis that (1) children who were in the primary grades from the late 1960s through the 1970s profited from improved beginning reading instruction; and (2) as these children moved through higher grade levels, they maintained their early advantage relative to other age cohorts. Children in first grade in 1968, 1972, and 1977 were tested as fourth graders in the 1971, 1975, and 1980 NAEP assessments, which showed steady improvement over time. First graders in 1981, tested as fourth graders in the 1984 assessment, did not do as well as those tested four years before. Children in first grade in 1964, 1968, 1973, and 1977, tested as

eighth graders in the first four reading assessments, showed a slight improvement from 1971 to 1975, a larger improvement from 1975 to 1980, then virtually no change from 1980 to 1984. Among 17-year-olds, performance was essentially flat from 1971 through 1980, then improved from 1980 to 1984.

Dr. Chall extrapolated these trends to predict a continued decline at age 9, stable performance at age 13, and improvement at age 17. At age 9, the direction of changes in both the mean and the variance are in accord with her predictions. According to Dr. Chall, a meaning-emphasis approach to beginning reading (as opposed to a code-emphasis program) would have resulted in general declines, and would have been most detrimental to lower ability youngsters. The 13-year-olds were in the first and second grades in the late 1970s. Dr. Chall suggests that these years were "characterized by a stronger emphasis on word recognition and phonics," and goes on to argue that because they benefited from these early code-based programs, the 1986 13-year-olds "were more prepared to benefit from the emphasis on reading comprehension that they may have received when they were in the intermediate and upper elementary grades."

Concerning the 17-year-old test score decline, Dr. Chall acknowledges that an explanation based *mainly* on the beneficial effects of stronger beginnings does not seem to hold. Since the 1986 17-year-old cohort was in the primary grades during the 1970s, they would have been expected on that basis alone to show gains. In considering other factors that may have brought about actual declines for 17-year-olds, Dr. Chall proposes as one possibility the publication and wide influence of A Nation at Risk (National Commission on

Excellence in Education, 1983) and other "reform reports," published around 1983 and 1984. She observes that these reports called for raising standards and curriculum requirements, increasing the difficulty level of textbooks, and placing more emphasis on higher mental processes. Although some of these reports suggested remedial instruction for the lowest achievers, Chall questions whether much was provided. If changes in the direction of higher standards and more difficult textbooks were implemented, they may have been detrimental for students having difficulty meeting even the lower standards, unless these students received remedial instruction. Dr. Chall goes on to suggest several out-of-school factors that may also have contributed to the sharp two-year score decline among 17-year-olds. The direction of changes in both the mean and the variance of the 17-year-old score distributions from 1984 to 1986 are in conformity with her explanations.

Efforts by ETS to Resolve the Question of the 1986 Reading Anomaly

As part of the 1988 NAEP data collection, ETS is conducting parallel data collections replicating as closely as possible the procedures followed in the 1984 reading assessment, and also the procedures followed in 1986. If 1986 reading scores were influenced by one or more of the changes in testing procedures introduced in 1986, and if these changes act in the same way to influence respondents in 1988 as they did in 1986, then comparing the results of the 1984 replication and the 1986 replication will yield an estimate of the adjustment that must be made to the original 1986 results to make them comparable to results from earlier assessments. As stated in the discussion of Recommendation 5, the panel strongly

endorses these efforts as a vital source of information for future design decisions. At the same time, we note that *ad hoc* investigations of apparent anomalies do not provide adequate quality control to assure the reliability and validity of inferences from NAEP about trends in achievement.

It is unfortunate that the results of the procedural comparisons being carried out as part of the 1988 assessment are not scheduled to be made available until sometime in 1989. One initial reaction is that a small, quick study should be mounted to compare the 1984 and 1986 procedures and get some answers. However, given that the effect to be detected amounts to a change of only a few percent in item difficulties, and given that replication implies the use of BIB spiralled booklets, a large, systematic study may be the only way to get satisfactorily definitive answers.

FINDINGS REGARDING TRENDS

Introduction

NAEP was established to provide accurate information at the national level about school achievement, including changes over time. Indeed, the title "National Assessment of Educational Progress" expresses that intent. In their 1987 Study Group report, The Nation's Report Card, Alexander and James reaffirmed this historic commitment, and this present Technical Review Panel also concurs strongly in the priority for NAEP of providing accurate and trustworthy information about achievement trends.

Even though NAEP was explicitly designed to provide accurate longitudinal information concerning student achievement, the public,

educational policy makers, and even scholars have often relied on other information sources when drawing conclusions about achievement trends. The widely publicized test score decline from the middle of the 1960s until the beginning of the 1980s was more generally associated with the SAT than with NAEP, and the question often arises whether the SAT, with its much larger and perhaps better motivated samples of examinees tested annually, provides better trend information for some purposes than NAEP, with its small samples, biannual schedule, and assurances of student anonymity.

Concerns over the accuracy of NAEP data for charting trends and making comparisons have been heightened by the 1986 NAEP reading anomaly, and also by the increasing interest in state-level achievement comparisons. The Secretary of Education's "wall chart" presently uses SAT and ACT scores for state-level achievement comparisons, and under the Hawkins-Stafford law, the 1990 and 1992 NAEP assessments will report achievement and achievement comparisons for participating states at selected grade levels in mathematics (1990 and 1992) and reading (1992).

In the light of these concerns, this panel was charged with addressing the question of whether NAEP was the most accurate barometer of trends in the achievement of American school children. Four of the panel's background papers, by Baldwin, Pandey, Wiley, and Schmidt, specifically addressed issues of achievement trends. Dr. Baldwin's paper focused on trends in reading, analyzing changes in NAEP reading objectives and in analysis and reporting procedures across the five reading assessments from 1971 through 1986. Dr. Pandey's paper addresses changes in the framework of NAEP

mathematics objectives across the four mathematics assessments from 1973 through 1986, and also presents systematic comparisons between trends reported from NAEP and trends derived from other sources. In Dr. Wiley's paper, he systematically compares the content of the SAT verbal subtests and the NAEP reading exercises, compares the populations represented by NAEP versus SAT examinee samples, and provides tentative comparisons between 1984 to 1986 changes in SAT scores and 17-year-old NAEP reading results for the highest achieving examinees. Dr. Schmidt's paper reviews a range of inconsistencies in NAEP procedures that may have compromised trend reporting, and calls for a more systematic procedure for considering any changes from one assessment cycle to the next.

The Panel's Conclusions Concerning NAEP Trends

The panel concluded that in comparison to other national, longitudinal data sources, NAEP provided the most accurate and useful information available. At the same time, we found significant deficiencies in NAEP, and have recommended several changes to improve the quality of NAEP trend data for the future. As stated in our Conclusion 2, college entrance examinations and NAEP sample different examinee populations and different domains of content. These differences virtually preclude valid comparisons between reported trends from NAEP versus college entrance tests, although it is possible that valid comparisons might be constructed for higher ability 17-year-olds through careful analyses of portions of the data from these two sources. Although some types of items appear on the SAT and not in NAEP (e.g., vocabulary items), we believe that NAEP is the better barometer of national trends. It more faithfully represents

the entire school population, tests younger as well as older children, samples a broader range of content areas and of objectives within content areas, and provides more detailed score reporting. Moreover, NAEP is designed specifically to assure the continuity of trend lines. In contrast, SAT and ACT trends are merely by-products of examinations designed for a very different purpose.

That being said, NAEP trend reporting nonetheless presents some difficulties. The panel noted that in both reading and mathematics, there have been substantial inconsistencies over the years in the content of NAEP exercises, in the frameworks of objectives used to organize those objectives, and in forms of reporting. Successive assessments within a given content area (e.g., reading or mathematics) have been linked using common items, but the different sets of linking items used over the years have sometimes represented quite different mixes of subdomains. (Such subdomains include, among others, Literal Comprehension, Inferential Comprehension, and Reference Skills in reading; or Numbers and Operations--Knowledge, Fundamental Methods, and Measurement in mathematics.) Our concern with these issues is reflected especially in Recommendations 1 and 5.

In addition to noting inconsistencies in the content of successive assessments, the panel was also concerned about the overall number, scope, and quality of NAEP exercises. Different panel members noted an overreliance on exercise formats calling for selection rather than production of correct responses, and insufficient coverage of higher level learning outcomes, as well as failure to measure and distinguish important fundamental skills

usually prerequisite to higher learnings. In Dr. Pandey's paper, he observed that most free-response exercises in mathematics looked like multiple-choice questions with the answers removed. These concerns are expressed in our Recommendations 2 and 6.

Comparison of Trends Reported from NAEP Versus Other Data Sources

Before considering the relative accuracy of achievement trends revealed by NAEP versus other data sources, it is well to consider the extent to which they agree. Pandey's paper presents comparisons of NAEP with other data sources in mathematics, and Wiley's paper discusses comparisons of reading trends from NAEP versus the SAT.

Pandey's comparison of mathematics achievement trends from different data sources indicates substantial agreement in the directions of performance changes over time, although the relative magnitudes of changes shown by different data sources are difficult to compare directly. Comparability is limited by the definition of NAEP samples prior to 1983 according to age rather than grade level; and by the fact that alternative longitudinal data sources generally are not nationally representative. Other differences noted by Pandey include the time of year of testing, test administration procedures, and the content of different tests. Despite these limitations, NAEP trends were compared to data sources including the SAT, American College Testing (ACT) program, General Educational Development (GED) examination program, National Longitudinal Study (NLS) of the High School Class of 1972, High School and Beyond (HSB) study, and the Iowa Tests of Educational Development (ITED). The data summaries and tabulations made by Koretz (1986, 1987) were used extensively in making these comparisons. Within the limitations of the data, trends

revealed by NAEP appear consistent with those derived from other data sources.

Wiley's paper offers a more detailed analysis of the limitations of comparisons between NAEP and SAT reading trends, focusing especially on 1984 to 1986 performance changes among 17-year-olds. The principal limitations on such comparisons are differences in the populations sampled and in the kinds of items included. Differences between populations can be accounted for if one assumes that above some ability level, virtually all examinees would have taken the SAT. Under this assumption, and using data on the proportion of all in-school 17-year-olds taking the SAT, it is possible to derive corresponding percentile ranks in the SAT and NAEP achievement distributions. This amounts to comparing the most able students in the NAEP sample with the most able students in the SAT sample. The calculations required are presented in Wiley's paper.

Content differences can be reduced but not eliminated by using only the reading comprehension subscale of the SAT verbal scale (i.e., excluding the vocabulary subscale). Compared to SAT reading comprehension items, the NAEP reading exercises span a broader range of difficulty levels. A relatively small proportion of NAEP exercises are as difficult as those in the SAT. However, these more difficult NAEP exercises are probably the most discriminating for high-ability examinees, and so functionally the tests may measure similar skills at high ability levels. Careful consideration of the difficulties in validly comparing NAEP and SAT trends highlights the limitations of the SAT as a general barometer of educational achievement.

When the procedures developed in Wiley's paper are applied to

NAEP and SAT data in 1984 and in 1986, both data sources show improvements from 1984 to 1986. (Recall that the increase in variance of achievement scores led to increases from 1984 to 1986 in the proportions of very high-scoring examinees as well as very low-scoring examinees at all three age levels.) The magnitudes of these examinees' improvements according to the SAT versus NAEP probably cannot be validly compared. In any case, no such comparison has been attempted.

Accuracy and Interpretability of NAEP Trends

In both reading and mathematics, the accuracy and interpretability of NAEP trends have been diminished by changes over assessment in the frameworks used to organize objectives, and by changes in the mix of exercises used to link successive assessments. In Baldwin's paper, she tabulates the proportions of Literal Comprehension, Inferential Comprehension, and Reference Skills exercises included each of the five NAEP reading assessments from 1971 through 1986, and also the numbers of exercises in each category that were common to more than one assessment. She finds that trends in 17-year-olds' achievement differ from one item type to another, and that a different mix of subdomains was used in linking the 1984 and 1986 assessments than had been used to link earlier assessments. In particular, the proportions of Literal and of Inferential Comprehension exercises each dropped by about ten percent, and the proportion of Reference Skills exercises increased from 15 percent to 34 percent.

These findings have two major implications. First, the meaning of trend comparisons has not been entirely stable over time. Second,

exclusive reliance on the NAEP reading scale in examining and interpreting achievement trends might obscure important differences in performance changes for different kinds of reading skills. This is not to say that summaries like that provided by the NAEP reading scale are without value. Such scales can provide readily interpretable summaries of broad trends, and can be useful in guiding educational policy. However, they must be supplemented with more refined scales focused on component skills, as discussed in the panel's Recommendation 8. The unidimensional models on which the reading and other NAEP scales are based are only approximations. Important differences exist among subdomains in reading and other content areas. For that reason, it is important that the mix of exercise types comprised in such scales be held constant over time.

In Pandey's paper, he presents an analysis of NAEP mathematics objectives through time that shows even greater variation over time than in reading. Pandey reclassifies exercises from earlier assessments in terms of the most recent objectives framework, and tabulates the numbers of exercises in each category that were common to two or more of the 1978, 1982, and 1986 assessments. His tables show, for example, that at the 9-year-old level, between 1978 and 1982, 28 percent of the common items were from the categories Numbers and Operations--Knowledge and Numbers and Operations--Applications, and 13 percent were from the category Fundamental Methods. Between 1982 and 1986, the corresponding percentages had changed from 28 percent to 41 percent, and from 13 percent to 7 percent. Less dramatic changes also occurred for other content categories, and at the 13- and 17-year-old levels. These

comments should not be taken to imply that the content and meaning of NAEP scales should never change, only that such change should be planful and deliberate, not accidental. An unintended consequence of the consensual process used to arrive at sets of objectives for each separate assessment may have been diminished attention to consistency through time.

RECOMMENDATIONS FOR NAEP IN 1990 AND BEYOND

Introduction

The panel's final charge was to consider issues that will arise in the expansion of NAEP to provide state-level achievement comparisons. The Hawkins-Stafford law calls for state-level comparisons on a pilot basis in 1990 and 1992, with states participating on a voluntary basis. In 1990, state-level comparisons will be made in mathematics at a single age/grade level. In 1992, pilot state-level assessments will be conducted in mathematics at two age/grade levels, and in reading at one level. The expectation is that at some point beyond 1992, state-level comparisons will be expanded well beyond these pilot studies.

State-level comparisons have already been made on a regional basis by the Southern Regional Education Board (SREB). With the cooperation and assistance of the Educational Testing Service, participating states conducted assessments modeled after NAEP and compared their performance among states and against national performance levels. In addition to states participating in the SREB comparisons, ETS has provided state-level assessments to several other states, augmenting national NAEP samples and collecting

additional data as part of the regular national NAEP data collection. These experiences with state-level assessments have helped to highlight technical and administrative issues likely to arise in the anticipated NAEP expansion, but the 1990 and 1992 pilot state assessments will provide far more information. The panel's deliberations, background papers, and recommendations are intended first of all to guide the 1990 and 1992 pilot assessments. Although we believe that our conception of an assessment supporting state-level comparisons is sound, we anticipate that decisions about the shape of the assessment after 1992 will be informed by the results of pilot studies over the next several years.

Design and Administration of National and State-Level NAEP

There is no fundamental difference between the organization and activities of an assessment designed to provide national achievement estimates and one reporting at the level of the separate states. However, differences in the intended uses of the data collected, in concomitant incentives on the part of students, teachers, and test administrators, and in scale and cost have implications for the design of an expanded assessment program.

Even apart from the expansion of NAEP to accommodate state assessments, the panel would call for some changes in the design of the assessment. The 1986 reading anomaly as well as limitations in the validity and interpretability of NAEP trends suggest a need for improvements, even if the scope and purposes of the National Assessment were to remain as they are now. The design recommendations in this section address both the improvement of NAEP at the national level and the expansion of NAEP to provide

state-level comparisons.

A first, major question in the expansion of NAEP is whether separate data collections should inform national versus state-level achievement estimates, or whether a single data collection should serve both purposes. Ultimately, it may be that national estimates will be obtained from the union of state-level data collections. For 1990 and 1992, however, it seems clear that a national-level data collection will be designed following essentially the same procedures as in the last two or three assessments, with augmentations to the national sample in states electing to participate in the pilot state-level comparisons. A sample designed to provide good estimates for the nation would not be the same as one designed to provide good estimates for the separate states. Moreover, abrupt changes in the national NAEP sampling plan could jeopardize the continuity of national trends. Most importantly, state participation in NAEP will be voluntary in 1990, 1992, and beyond. Valid national estimates could not be assured if NAEP were to rely on data from some arbitrary subset of states electing to participate.

In this section, the terms national-NAEP and state-level NAEP will be used to refer respectively to the national NAEP data collection (and national-level data analysis and reporting) and to the state-level augmentations (along with state-by-state analysis, reporting, and comparisons). The precision of national-NAEP estimates may be enhanced by using state-level data, but these data cannot be combined in any simple way unless they are collected using instruments that are identical in every important respect, administered under carefully standardized conditions to samples having a known relationship to the

national-NAEP sampling frame.

Whether or not state-level NAEP data are used in formulating national achievement estimates, comparability between state and national data collections is critical. A primary purpose of state-level data collections will be to enable comparisons between state and national achievement levels, as well as comparisons among states. Mechanisms to ensure such comparability are addressed most directly in the background papers by Musick and by Bock, although most of the papers touch upon these concerns to a greater or lesser extent. The evaluation and quality control mechanisms set forth in the paper by Melnick would also go a long way toward assuring the dependability of such comparisons.

Recommendations concerning state-level NAEP procedures. The two papers by Musick and by Bock each propose specific procedures for a state-level assessment. Dr. Musick draws on his experience in state-level comparisons with the Southern Regional Education Board, and Dr. Bock draws on his experience in working with several states on designs for assessment programs. The two papers are complementary--Musick addresses primarily issues of administration, governance, and the logistics of data collection. Bock addresses primarily technical issues in the design, analysis, and reporting of assessment results. The proposals expressed in these two papers have evolved through the course of the panel's deliberations. By and large, the panel has reached consensus on at least the broad outlines of a design, the general features of which are sketched below. Justifications and supporting details may be found in the background papers, especially those just cited.

The panel conceives of state-level NAEP as a program unit under the parent National-NAEP organization. It would have its own staff director and would be supported by its own advisory structure, reflecting state interests and concerns. State-level assessments would be conducted in conjunction with National-NAEP assessments, within the same time frame, which should be shorter than the present 12-week testing schedule for national biannual assessments.

Present NAEP data collection procedures are designed to be minimally intrusive for participating schools. Once a school is contacted, however, the incremental costs of testing more students within that school or of increasing the testing time are relatively small. The more direct involvement of the states in NAEP offers an opportunity to reconsider decisions about testing burden. For a variety of reasons, the panel believes that testing time should be increased. (See Recommendation 4.) A two-stage testing approach might further increase the efficiency of the assessment, helping to assure that the best possible use was made of students' time. Possible methods of implementing this and related procedures are presented in Bock's paper.

The utility of the state-level and national-NAEP can be dramatically increased by providing methods of linking to them the autonomous testing and assessment programs of individual states. States that had suitable testing programs and chose to carry out such linking could then report results of their own testing for schools or districts, in terms of the NAEP scales. In Bock's paper, he describes feasible methods for accomplishing such linkages while assuring the confidentiality of NAEP results for schools and students, as required

by law.

Assuring comparability. As stated in our Recommendation 6, it is critical that the administration procedures for state and national NAEP be the same in every important respect. However, these need not be the same as present NAEP procedures. For example, the quality of both state and national data could probably be increased if students were tested in groups of no more than 30, and if two adults were present at each testing session. One would be an external examiner trained under the direction of the National-NAEP contractor, and the other would be someone from the local school with whom the students were familiar. This would help prevent disruptions which might depress the scores of entire groups of students, and would help to minimize the "substitute teacher effect," especially with younger children. The external examiners would probably be persons provided by the state for the time required for training and test administration. Personnel might be recruited from the field staffs of large sample survey firms, from the ranks of professional substitute teachers, or from the faculty of community colleges, for example. External examiners would be chosen to minimize travel and overnight lodging expenses, but would not manage any assessments in the school organizations by which they were employed. Note that changes in the size of testing sessions or in the number of adults present would call for systematic review, and for bridge studies to assure the continuity of trends, as discussed in Recommendation 5.

In addition to administration procedures for state-level and national-NAEP that are the same in all important respects, comparability requires the use of common instrumentation. The

occurrence of the 1986 reading anomaly suggests that the some items may possibly function differently in the context of different exercise booklets, assumptions of item response theory notwithstanding. To the extent possible, state-level NAEP exercise booklets should be identical to national-NAEP booklets. If the content of state-level NAEP is less comprehensive than that of national-NAEP, then the state-level NAEP booklets should correspond to a subset of the national-NAEP booklets. As mentioned in Recommendation 6, states may choose to supplement the core state-level NAEP data collection, but any supplemental questions should follow the core instrumentation.

Common administration procedures and common instrumentation are two of the four critical requirements for assuring state-level and national-NAEP comparability. The remaining concerns are first, comparable sampling of schools and students within schools; and second, uniform procedures for determining which students should be excused from testing because they are of limited English proficiency, educable mentally retarded, or functionally disabled.

Samples of schools and of students within schools should be drawn under the supervision of the national-NAEP contractor or subcontractor responsible for the national-NAEP sample. The same sampling frame should be used, although of course the selection of numbers and proportions of schools selected within strata may differ. States may assist in the sampling by providing technical or clerical assistance or, if necessary, lists of schools. Selection of students within schools should likewise follow the same procedures for

state-level as for national-NAEP.

Exclusion criteria for limited English proficiency, educable mentally retarded, or functionally disabled students must be defined in the same way for state-level and national-NAEP, and must also be applied in a consistent fashion. Ultimately, these criteria will be interpreted by hundreds or thousands of individuals at the local school level. Detailed written procedures and careful training can help to assure an acceptable degree of uniformity, but in addition, individually written justifications for each student excluded and random audits may be helpful in assuring compliance.

As procedures are developed for all aspects of the state-level NAEP data collection, state testing and assessment personnel should be involved. Procedural manuals should be written to provide a common authoritative reference and to help assure compliance.

Cognitive Items

From its inception, NAEP has espoused the goal of measuring the full range of significant learning objectives in different content areas. The NAEP exercises, the tasks set for students to find out what they know or can do, are the heart of the assessment. No refinements in sampling or administration or statistical methodology can compensate for deficiencies in the scope or quality of the exercise pools.

In the light of recent concerns over the quality of NAEP trends, and in the light of the new purposes that will accompany NAEP's expansion to provide state-level estimates, the quality of NAEP exercises has taken on even greater significance. The panel has serious concerns about the size and scope of the NAEP exercise pools

and about the stability over time of their organizing frameworks. Even our consideration of the 1986 reading anomaly was impeded by the impossibility of distinguishing 1984-to-1986 changes with respect to different constructs within the area of reading. Several of the background papers address the quality of the NAEP exercise pools, including the papers by Pandey and Baldwin, but these issues are most extensively considered in the paper by Guthrie and Hutchinson.

Continued attention to fundamentals, and increased emphasis on higher level learning. The advent of state-by-state comparisons will bring NAEP more than ever into the public eye, and will increase pressures to "teach to the test." In itself, this need not be a bad thing. Many states, districts, and schools have consciously used testing to shape curriculum and instruction. But if NAEP's influence on curriculum and instruction is to be salutary and not detrimental, then its exercise pools must be comprehensive. As discussed in our Recommendation 2, the NAEP exercises must embrace both fundamentals and higher level learning. NAEP must assess the intended learning outcomes of typical American school programs, but must also reach beyond the typical to point directions for improvement. Decisions about the learning outcomes assessed should reflect the best thinking of scholars and subject matter specialists.

Testing process learning outcomes in concert. In most content areas, particularly reading, writing, and mathematics, several distinct cognitive processes may be logically distinguished and identified with different kinds of exercises. These may include lower level and higher level processes (e.g., word attack versus inferential comprehension) or processes at a comparable level (e.g., inferential

comprehension and interpretation). There is a tension between using exercises that call for these processes separately (e.g., exercises to assess word attack skills) versus exercises that call for their use in concert. Careful, scholarly deliberation will be called for to resolve questions about the granularity of both exercises and reporting scales in each content area assessed.

One primary consideration is the intended scoring and reporting of assessment outcomes. Once skills are combined at the level of the NAEP exercises, it becomes difficult if not impossible to report them separately, or to disentangle their relative contributions to possible low performance. It is neither appropriate nor useful for NAEP to aim for detailed, diagnostic profiles of learning strengths and weaknesses. At the same time, the assessment should be sensitive to and able to distinguish broad changes over time in curriculum focus and emphasis, for example, the direct instruction of high-level strategies in reading, or greater problem solving focus in mathematics.

A second consideration is the factorial structure of the processes involved. If important component processes have low intercorrelations, then they should probably be assessed separately, but if they are highly intercorrelated, separate tests and test scores may be redundant. It must be recognized that high intercorrelations among items in an exercise pool may result from instructional practices and curricular organizations, as well as the logical, substantive structure of the exercises themselves, and the inherent nature of children's cognitive development and information processing. However, it is likely that exercises could be organized

into two or more independent scales at each age/grade level in reading, mathematics, or writing, acknowledging the complexity of these areas while providing sufficient statistical independence that scores would be separately interpretable.

Consideration of efficiency may argue in favor of more complex, integrative tasks that assess lower-level processes as components of more complex performances, making separate tests of those processes unnecessary. A long division problem may test division, subtraction, and multiplication all at once. At the 13- and 17-year-old levels, and probably as early as the 9-year-old level, reading is a critical tool to support learning in other school subject areas, and mathematics may be important for representing, understanding, and manipulating concepts throughout the curriculum. As a consequence, exercises that assess the ability to use reading and mathematics as "tools" for problem solving in content areas should be included.

Finally, findings from cognitive psychology may also argue in favor of more integrated assessments of process. As discussed in Guthrie's and Hutchinson's paper, logically separable processes develop in concert, and support one another. Full mastery of one process may be impossible without partial mastery of others. Moreover, the ability to apply different processes appears to be context bound. It is notoriously difficult to achieve transfer of learned skills to different applications. If reading processes like monitoring for understanding or using previous knowledge to understand new ideas are tested in isolation, an incentive is created to teach them in isolation, and for many children, their concerted

application in actual reading may be far from automatic.

Separation of content and process. The knowledge or skills an exercise is designed to assess may be referred to as its intended requirements. The intent of the exercise is to distinguish those that do or do not possess the attributes it is designed to measure. But every exercise also calls for additional knowledge, skills, and dispositions. At the very least, valid assessment of the intended requirements relies on knowledge about the test-taking situation, skill in marking answers accurately, and a disposition to attempt each exercise seriously. These and other additional skills may be referred to as an exercise's ancillary requirements. Valid assessment of intended knowledge and skills is only possible if an exercise's ancillary requirements are held to a level that can be presumed nearly universal among the group tested.

Content knowledge is often ancillary to the measurement of process, and processes like reading and writing are often ancillary to the measurement of content. Thus, exercises intended to assess content knowledge and not reading ability should be written at a reading level at least two years below the grade level of the examinees. Likewise, exercises intended to assess reasoning or problem solving should have children apply these skills to material or situations with which the great majority should be very familiar. An assessment that purports to show the extent of historical knowledge should not be confounded with students' reading ability, although reading in the content area of history might in itself be a worthwhile thing to assess. Likewise, scores on an assessment of critical thinking should not be confounded with content that is known to some

students and unknown to others.

Background Items

Exercises measuring learning outcomes may be at the heart of the assessment, but valid and reliable measurement of learning outcomes alone is not enough. NAEP achievement data become useful for policy when they can be related to other variables. Background questionnaires for students, teachers, and principals are administered concurrently with student achievement exercises. These permit the reporting of NAEP results for major subpopulations (e.g., gender and race/ethnicity) and in recent years have also provided limited information on instructional processes, so that these could also be related to schooling outcomes. In the panel's deliberations, some of the same concerns were raised about these background instruments as about the cognitive exercises. The panel saw a need for greater consistency through time in the questions asked, and for a stable and coherent framework to guide the selection of background questions. As stated in Recommendations 3 and 4, we believe that the time allocated to background data collection should be increased, especially in the light of new purposes accompanying the advent of state-by-state comparisons.

Several of the background papers address these concerns, but the panel's positions are developed most extensively in the paper by Baron and Forgione. In their paper, Dr. Baron and Dr. Forgione draw upon their experience with the Connecticut Assessment of Educational Progress, and make extensive use of Dr. Jeannie Oakes's framework for organizing alterable variables in education. Their paper includes appendices giving some useful classifications of variables, although

they caution that not all of these are important to assess. Baron and Forgione also call attention to the need for coordination and triangulation of questions on student, teacher, and principal questionnaires, bearing in mind that these classes of respondents bring different perspectives to bear on schooling processes.

NAEP background questions serve a variety of purposes in the assessment. There is no end of interesting questions to ask, and so a judicious, disciplined selection of background questions is essential. The panel proposes that each background question should represent at least one of three broad categories. First would be "unalterable," or demographic variables important for describing patterns in NAEP achievement data. Second would be variables plausibly related to achievement, including indicators of curriculum content and orientation or of instructional practices. Third would be variables reflective of valued schooling outcomes that cannot be directly captured by NAEP achievement exercises. Each type of variable is described below.

Variables important for describing NAEP achievement patterns. Examples of unalterable values (the first category) are gender, race/ethnicity, socioeconomic status, and size and type of community. These variables are "unalterable," but many educational policies and philosophies are predicated on the well founded assumption that their relations to schooling outcomes are alterable.

Variables plausibly related to achievement. Variables plausibly related to achievement include questions about instructional practices, for example, a "whole language" approach to reading, writing, speaking, and listening. Baron and Forgione report limited

success with such questions in their own experience, in part due to the validity of the questions themselves, and in part due to the complexity of the relationships of these variables to learning. For example, teachers may give more feedback on papers to low achieving students, so a simple correlation appears to show that teacher feedback is negatively related to achievement. A more promising focus for questions in this category may be on student opportunity to learn. If both the quality and the quantity of relevant content coverage can be addressed, such background questions may emerge as powerful predictors of learning outcomes. Questions about homework and out-of-school pursuits can also help to inform the sum total of students' educative experiences. Specific examples of background questions in a range of content areas are provided in Barons' and Forgione's paper. Finally, this second category of background questions plausibly related to achievement might include a few concomitant measures of achievement that might be used to validate patterns of NAEP findings, e.g., "What grades do you usually get in school?" (or in some particular content area being assessed).

Variables reflective of other valued schooling outcomes. The third category of variables to be represented among background questions include, for example, amount of leisure reading, or participation in student elections. A range of affective or attitudinal schooling outcomes should also be sampled in the NAEP background questions. In their paper, Baron and Forgione offer by way of illustration two statements from a Connecticut state testing program, with which students were asked to agree or disagree: "Careers in science are more appropriate for men than for women" and

"My knowledge of science will be of little value to me in my day-to-day life."

Educational indicators as ends in themselves. Once any indicator is assessed and reported, improvement with respect to that indicator may become an end in itself. This is true of both achievement exercises and background questions. Thus, background questions should be selected such that direct efforts to improve a school's standing with respect to those questions would be salutary for education. If counts of courses taken are reported, for example, there may be an incentive to offer a greater number of "watered down" courses, with no concomitant improvement in student learning. Following Murnane (1987), Baron and Forgione argue that this corruptibility of indicators can be diminished if they are specifically defined and include a qualitative as well as a quantitative dimension.

Analysis and Reporting of Results for the Nation and for Participating States

The expansion of NAEP to provide state-by-state comparisons raises a number of new issues in the analysis and reporting of results. Perhaps foremost among these is the problem of reporting interstate comparisons, but state-to-national comparisons, state trends over time, within-state comparisons among regions or student subpopulations, and the reporting of distributions of school means as well as student-level achievement distributions all call for attention. A few of these issues may be set aside, for the present, in the light of the Hawkins-Stafford law's prohibition against reporting results for identifiable units below the state level of aggregation. The panel did not give detailed consideration to this entire range of issues, but did

consider a number of them.

Recommendations 7, 8, 9, 10, and 11 all address analysis and reporting issues. At both the state and national levels, we recommend that assessment design and analysis permit the accurate estimation of scores for individual students. Specifically, we call for a retreat from the "plausible values" used as the basis for recent NAEP reports. We call for increased reporting of subdomain scores using scales specific to the content appropriate to specific age/grade levels where such scales are more appropriate than scales common to two or more levels. The panel also calls for fuller reporting of score distributions than is provided by means alone. Where feasible, these recommendations should be implemented in parallel fashion at the state and national levels. Finally, the panel calls for systematic study of alternative methods for making and reporting state comparisons. The panel recognizes that taken together, these recommendations imply increases in the amount of data collected (cf. Recommendation 4).

Issues in reporting at both state and national levels. Many of the reporting issues raised by the panel apply equally to national-NAEP and state-level NAEP. In addition to concerns addressed earlier in this Review of Findings, panel members addressed the use of NAEP proficiency scales that span the range from age 9 through age 17 (e.g., the present NAEP scales in reading and in mathematics); the value of reporting distributional summaries at the level of school means as well as score distributions for individuals; the timeliness of NAEP reporting; and the importance of relating NAEP performance to "real-world" schooling outcomes.

The use of common reporting scales across age/grade levels is problematical for at least two reasons. First, the range of such scales must necessarily be so broad that important within-grade variations are obscured because they occur over a narrow range of scale values. Second, such scales are difficult to interpret when they represent qualitatively different kinds of content at different age/grade levels. In mathematics, for example, the topics taught and tested for 17-year-olds may overlap little with those for 9-year-olds. The fact that an item response theoretic (IRT) model can be applied to data from three age levels combined does not assure that the results will be sensible.

Data analysis and reporting should take cognizance of the hierarchical nature of educational data. Methods for multilevel analysis and reporting should be used to present results for schools as well as for individuals both for states and for the entire nation.

It was argued in several papers that the present 18-month turnaround between data collection and reporting of results will be unacceptable for purposes of state-level comparison. The NAEP contractor should make basic statistical data available as soon as reasonably possible, before interpretative reports are written. At the same time or shortly thereafter, public use data tapes should be made available for secondary analysis. At the same time, as Dr. Burstein observes, raw, unelaborated columns of numbers may be inaccessible to important policy audiences. Concurrent release of data and interpretations is an important means of retaining control over the meanings imputed to the data and the kinds of recommendations they are used to support.

As discussed in Dr. Bock's background paper, the meaningfulness of NAEP reporting scales could be enhanced significantly by empirical studies relating performance on the NAEP scales to more directly measured, practical schooling outcomes. Measurement of students' ability to perform real-world tasks is far more costly than collecting data using paper-and-pencil measures. However, valid inferences about the population's performance on such tasks may be based on large-scale assessment using NAEP exercises, together with much smaller studies to determine the relation between NAEP scales and such real-world outcomes. Such studies would significantly advance the kind of construct validation envisioned by Guthrie and Hutchinson, and the statistical validation called for by Melnick.

Reporting of state-level results and interstate comparisons. State-by-state comparisons are addressed in several of the background papers, but especially the papers by Dr. Burstein and Dr. Haertel. Burstein observes that the panel's consideration of issues in reporting is complementary to recent work by the Council of Chief State School Officers (CCSSO) in modeling the consensus planning process recommended by the Alexander and James Study Group report, The Nation's Report Card. The panel's recommendations and those of the CCSSO should be largely compatible. We concur with the CCSSO that NAEP must measure a range of important learning outcomes, and that the system developed must not merely provide gross, simplistic state comparisons of the kind often seen with comparative school achievement data, but must place achievement patterns in the context of possibly different educational goals, demographics, and other contextual factors. Specific recommendations from this panel versus

the CCSSO Consensus Planning Project are contrasted in Dr. Burstein's paper.

Burstein's paper briefly reviews the context and assumptions surrounding the panel's consideration of analysis and reporting issues, and then turns to the purposes of state-level NAEP reporting. NAEP must provide a reliable and valid assessment, making efficient use the student time taken for data collection. Reporting must take account of the different needs and circumstances of the several states. To be useful in guiding policy, it should relate achievement to alterable variables--concrete features of the school systems that can be changed for the better by state and local educators. Fair and credible reporting of state-by-state comparisons is essential if state cooperation is to be enlisted and maintained. As stated in Recommendation 10, the 1990 and 1992 pilot assessments should be used to explore several alternative schemes for making and reporting such comparisons. Following in part on work by the CCSSO, Burstein recommends several specific methods that should be explored for making and reporting state-level comparisons.

In Haertel's paper, he considers models used within states for school- or district-level comparisons, and considers their applicability to the problem of state-by-state comparisons. Haertel concludes that it will probably be necessary to place state achievement disparities in the context of broad differences in socioeconomic level, although in principle reporting of unadjusted means for subpopulations within states could suffice. He cautions, however, that patterns of actual achievement must not be obscured. Raw and contextualized reports of state-level achievement

differences serve different sets of purposes, both important. Adjustments must not be permitted to legitimate existing inequalities in educational outcomes for different groups of learners.

Evaluation

The investigations triggered by the 1986 reading anomaly have called attention to a serious need for more systematic, ongoing statistical evaluation and audit of NAEP procedures and results. The kind of evaluation referred to would not consider the value or utility of NAEP, but would examine closely the statistical quality of NAEP findings. Dr. Melnick's paper expresses several of the panel's concerns. First, there is a need for empirical studies of the error structure of the assessment. Expert judgments, including those of this panel, cannot resolve fundamentally empirical questions. Bridging studies need to mirror the procedures of the main data collections in every important respect. Second, studies of the accuracy and quality of reported NAEP results need to be conducted on a routine basis, not just in response to apparent anomalies. Third, to the extent possible, statistical evaluation of NAEP should address the full range of error sources that may compromise NAEP findings, including sampling, fair and consistent application of exclusion criteria, and compliance with other aspects of administration procedures. The need for this kind of ongoing audit function is clearly heightened by the expansion of NAEP to provide state-by-state comparisons.

As set forth in Recommendation 12, an ongoing evaluation function should be established, independent of the NAEP contractor, which would regularly examine the overall accuracy of the

assessment, assist in distinguishing real from artifactual patterns and changes in achievement, identify design problems, and if necessary, provide some basis for analytical adjustments to compensate for planned procedural changes as they are implemented, and not retrospectively. This statistical evaluation could also consider issues of subpopulation bias, possibly uneven student motivation, and other factors that might detract from the validity of NAEP findings. As part of these validation activities, Melnick recommends where feasible the linkage of NAEP to other sources of information on achievement.

SUMMARY OF INDIVIDUALLY AUTHORED PAPERS

This Report is based on separately authored papers by nearly all panel members. These papers represent the positions of their individual authors, but reflect the deliberations of the entire panel. They provide detail and arguments in support of the panel's findings.

Herbert J. Walberg's paper, National Assessment for Improving Education: Retrospect and Prospect, establishes the policy context for our examination of NAEP, and the importance of its continuation. It places the current assessment in its historical context, and sketches some bold ideas for the future.

Jeanne S. Chall's paper, Could the Decline Be Real? Recent Trends in Reading Instruction and Support in the U.S., places the results of the 1986 reading assessment in the context of long-term patterns and trends, and argues that at least part of the decline may be attributable to changes in methods of reading instruction, especially a too-early emphasis on higher cognitive processes, as

well as to less support for reading and remediation in the school, home, and community.

Larry V. Hedge's paper, The NAEP/ETS Report on the 1986 Reading Data Anomaly: A Technical Critique, reviews the technical report on the 1986 reading anomaly by Beaton, et al., and evaluates the evidence presented concerning various hypothesized explanations. He criticizes the strategy of asking whether each hypothesis in turn could explain the bulk of the decline at age 9 or age 17, and suggests that a combination of changes in administration procedures might account for a substantial proportion of the changes in reading performance.

Janet Baldwin's paper, Reading Trend Data from the National Assessment of Educational Progress (NAEP): An Evaluation, reviews the quality of NAEP reading trend data. She finds that changes in procedures and in test content have confounded the meaning and interpretability of NAEP trend data, especially in the 1984 and 1986 assessments. Dr. Baldwin finds it problematical to make direct comparisons between SAT or ACT trends, but recommends a more rational framework for NAEP objectives, and greater consistency over assessment cycles.

Tej Pandey's paper, Mathematics Trends in NAEP: A Comparison With Other Data Sources, compares NAEP mathematics trends over nearly two decades with findings from the SAT, ACT, ITBS, ITED, GED, NLS-72, and HS&B. He finds no evidence of inconsistencies in the directions of changes between NAEP and other data sources, although the magnitudes of changes are difficult to compare from one test to another where standard deviations are not reported. Dr. Pandey

recommends improvements in the taxonomy of content and process categories used in defining NAEP mathematics objectives, and cautions that if NAEP approaches the status of a "national test," then the choice of content for NAEP will influence school curricula.

William H. Schmidt's paper, Quality Control: The Custodian of Continuity in NAEP Trends, addresses the importance of procedural as well as statistical and sampling consistency in NAEP. Dr. Schmidt places the 1986 reading anomaly in the context of other difficulties caused by past procedural modifications, and calls for better quality control mechanisms based on systematic procedures for considering all changes from one exercise cycle to the next.

David E. Wiley's paper, Assessment of National Trends in Achievement: An Examination of Recent Changes in NAEP Estimates, pursues two lines of investigation of the 1986 reading anomaly. First, Dr. Wiley examines both levels and distributions of scores at all three age/grade levels, and finds that, as a consequence of increased variability of the score distributions from 1984 to 1986, at sufficiently low levels of performance, there were declines for all three groups, while at sufficiently high levels, there were improvements for all three groups. Second, Dr. Wiley compares the content of the age 17 NAEP exercises to that of the SAT. While noting important differences, he finds sufficient parallelism to support cautious comparisons of SAT reading performance and NAEP reading performance for high-ability students. He finds that SAT changes do parallel NAEP changes. Dr. Wiley concludes that the magnitude of the NAEP reading scale score changes between 1984 and 1986 together with the large increase in score distribution variability make

methodological changes between the two assessments the most likely primary cause of the decline.

Mark D. Musick's paper, Management and Administration of a State-NAEP Program, recommends that the State-NAEP program be established as a program unit within National-NAEP. Dr. Musick considers a range of issues in the administration and governance of such a unit, and in the articulation of State-NAEP with National-NAEP, including instrumentation, sampling, identification of students excluded from testing, local options for expanded assessments within states, test administration, reporting of findings, and other matters. He concludes that establishing and administering a nationwide student testing program that uses the NAEP to provide information on a state-by-state basis is a manageable task.

R. Darrell Bock's paper, Recommendations for a Biennial National Educational Assessment, Reporting by State, provides a comprehensive and detailed technical plan for a National Assessment permitting state-by-state comparisons, and permitting an orderly evolution as statistical methodological advances ("updateability"). Dr. Bock's design for an assessment addresses issues of sampling, assessment cycles, domain definitions, assessment instruments, background questionnaires, administration procedures, scoring, reporting and technical support. His design includes both objective questions and writing exercises, and allows for the linkage of existing state assessments to a national assessment.

John T. Guthrie's and Susan R. Hutchinson's paper, Objectives for State Assessment by NAEP, considers the content of State-NAEP assessments in the light of the purposes these assessments will

serve. In specifications for NAEP exercises, the authors argue that ancillary, or unintended, requirements of exercises must be considered, as well as intended objectives. For example, readability of exercises (other than those designed to test reading) should be at least two years below the age/grade level at which the exercises are to be used, and inference items should not depend on factual knowledge that cannot be presumed to be nearly universal among the group tested. The authors also consider whether content and processes should be assessed separately or in concert.

Joan Boycoff Baron's and Pascal D. Forgione, Jr.'s paper, Collecting and Profiling School/Instructional Variables as Part of the State-NAEP Results Reporting: Some Technical and Policy Issues, presents issues and recommendations relating to the collection of NAEP background data. The authors propose criteria for selecting background questions based on prior theory, research, and empirical findings, and propose a long term NAEP research agenda to improve and stabilize the NAEP background data collection.

Leigh Burstein's paper, Technical and Policy Issues Related to Reporting of State Level NAEP in a Fair and Credible Manner, addresses the problem of state-by-state reporting in 1990 and beyond. Dr. Burstein considers the purposes of state-by-state reporting, technical issues associated with possible reporting methods, and alternative bases of comparison. He recommends possible strategies to be used on a pilot basis in 1990. Dr. Burstein also contrasts the panel's recommendations with the corresponding recommendations of the CCSSO Consensus Planning Project concerning state-level achievement reporting, exclusion criteria for Limited English

Proficient (LEP) and special education students, and the administration of the 1990 assessment.

Edward Haertel's paper, Within-State Comparisons: Suitability of State Models for National Comparisons, first considers problems of equity or fairness that arise with any system for adjusting scores or setting different expectations for different schools, districts, or states. He then describes systems used in several states for reporting school- or district-level achievement, and considers the applicability of these methods for purposes of state-by-state comparisons. In conclusion, Dr. Haertel suggests three possible approaches, including the reporting of unadjusted means for demographic subgroups, comparisons of each state's performance to a predicted level derived from models for subunits (e.g., communities) within the state, and a method using "floating comparison groups," as recommended by the CCSSO Consensus Planning Project.

Dan Melnick's paper, Evaluation for National and State-Level NAEP, addresses the need for a statistical evaluation and validation of NAEP, especially as it is expanded to provide state achievement estimates, and in the light of the 1986 reading anomaly. Dr. Melnick calls for an ongoing statistical evaluation of the accuracy of NAEP findings, to be conducted on a regular basis, and independent of the conduct of the assessment itself.

In summary, both collectively and individually, we have given careful attention to the three issues we were charged to address. It is our hope that our report, recommendations, and conclusions will help to guide and improve the National Assessment of Educational Progress in years to come.

REFERENCES

Alexander, L., & James, H. T. (1987). The Nation's Report Card: Improving the assessment of student achievement (Report of the Study Group). Washington, DC: National Academy of Education.

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 (Report No. 17-R-01). Princeton, NJ: Educational Testing Service.

Beaton, A., Ferris, J. J., Johnson, E. G., Johnson, J. R., Mislevy, R. J., & Zwick, R. (1988). The NAEP 1985-86 reading anomaly: A technical report. Princeton, NJ: Educational Testing Service.

Chall, J. S. (1983). Literacy: Trends and explanations. Educational Researcher, 12(9), 3-8.

Chall, J. S. (1986a). School and teacher factors and the NAEP reading assessments (Paper commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report, The nation's report card.) (ERIC Document Reproduction Service No. ED 279 667.)

Chall, J. S. (1986b). New reading trends: The NAEP report card. Curriculum Review, 25(4), 42-44.

Koretz, D. (1986). Trends in educational achievement. Washington, DC: Congressional Budget Office.

Koretz, D. (1987). Educational achievement: Explanations and implications of recent trends. Washington, DC: Congressional Budget Office.

Murnane, R. J. (1987, April). Improving education indicators and economic indicators: The same problems? (Paper presented at the meeting of the American Educational Research Association, Washington, DC.)

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington, DC: The Commission.