



Audit
(312) 886-6503

UNITED STATES DEPARTMENT OF EDUCATION
OFFICE OF INSPECTOR GENERAL

REGION V
111 NORTH CANAL, SUITE 940
CHICAGO, ILLINOIS 60606

FAX: (312) 353-0244

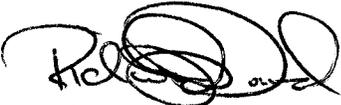


Investigation
(312) 353-7891

MEMORANDUM

DATE : June 17, 2003

TO: Grover J. Whitehurst
Director, Institute of Education Sciences


FROM: Richard J. Dowd
Regional Inspector General for Audit
Chicago, IL

SUBJECT : FINAL AUDIT REPORT
Review of Management Controls Over Scoring of the National Assessment of Educational Progress (NAEP) 2000
Control Number ED-OIG/A05-C0010

Attached is our subject final report that covers the results of our review of management controls over scoring of the National Assessment of Educational Progress 2000 assessment during October 1, 1999, through September 30, 2000. We received your comments concurring with the findings and recommendations in our draft audit report.

Please provide the Supervisor, Post Audit Group, Office of Chief Financial Officer and the Office of Inspector General with quarterly status reports on promised corrective actions until all such actions have been completed or continued follow-up is unnecessary.

In accordance with the Freedom of Information Act (5 U.S.C. § 552), reports issued by the Office of Inspector General are available, if requested, to members of the press and general public to the extent information contained therein is not subject to exemptions in the Act.

We appreciate the cooperation given us in the review. Should you have any questions concerning this report, please call me at 312-886-6503.

Attachment

**Review of Management Controls Over Scoring of the National
Assessment of Educational Progress 2000**



FINAL AUDIT REPORT
ED-OIG/A05-C0010
June 2003

Our mission is to promote the efficiency,
effectiveness, and integrity of the
Department's programs and operations.



U.S. Department of Education
Office of Inspector General
Chicago, Illinois

Notice

Statements that managerial practices need improvements, as well as other conclusions and recommendations in this report represent the opinions of the Office of Inspector General. Determinations of corrective action to be taken will be made by the appropriate Department of Education officials.

In accordance with Freedom of Information Act (5 U.S.C. §552), reports issued by the Office of Inspector General are available, if requested, to members of the press and general public to the extent information contained therein is not subject to exemptions in the Act.

Table of Contents

EXECUTIVE SUMMARY	1
BACKGROUND	2
NAEP MANAGEMENT CONTROLS OVER SCORING ARE ADEQUATE	8
Monitoring	8
Recommendations	10
Receipt and Control Process	10
Scoring	10
Data Quality	11
Analysis and Reporting	12
Other Issues	13
OTHER MATTERS	13
OBJECTIVE, SCOPE, AND METHODOLOGY	14
STATEMENT ON MANAGEMENT CONTROLS	17
ATTACHMENTS	
Attachment 1 – Additional Management Control Detail Not Presented in the Body of the Report	7 pages
Attachment 2 – Institute of Education Sciences’ Comments on the Draft Report	3 pages

EXECUTIVE SUMMARY

Our audit objectives were to determine whether management controls over scoring of the National Assessment of Educational Progress (NAEP) 2000 assessment were in place and adequate to provide reasonable assurance that the assessment results could be relied on during the period October 1, 1999, through September 30, 2000. Based on the work performed, we determined that the management controls over scoring of the NAEP 2000 assessment were adequate and generally working as intended. However, our audit work disclosed two nonmaterial weaknesses regarding the monitoring of mathematics qualification sets and scorer qualifications. State assessments required under the *No Child Left Behind* Act could also benefit from standards for management controls over scoring. We plan to report on this separately. The Institute of Education Sciences concurred with our recommendations and its written comments are included as Attachment 2 to this report.

To accomplish our objectives, we (1) obtained background materials and interviewed officials from the National Center for Education Statistics (NCES), National Assessment Governing Board (NAGB), Westat, Educational Testing Service (ETS), and NCS Pearson (NCS) to gain an understanding of their role in conducting the NAEP 2000 assessment; and (2) gained an understanding of current Administration and Congressional proposals that could have an affect on NAEP such as the Government Performance and Results Act (GPRA), the *No Child Left Behind* Act, the Elementary and Secondary Education Act, and other legislation affecting management controls. We also gained an understanding of state assessments through interviews with ETS and NCS officials. We reviewed and tested management controls over scoring to ensure the processes were adequate and working as intended. To review management controls over scoring, we interviewed officials to identify the management controls that were in place and reviewed various documents used in the process. To test the management controls over scoring, we examined the NCS mainframe final data for anomalies, identified the scorers for each subject, and interviewed judgmentally selected scorers. In addition, we reviewed the NCS and ETS computer processed data to ensure that it was reliable. To determine data reliability, we assessed data completeness, data authenticity, and the accuracy of computer processing. In addition, we gained an understanding of the ETS scoring analysis and reporting process. Further, we reviewed selected NCS employee payroll, personnel file, and position description records, and NCS' NAEP profit margin.

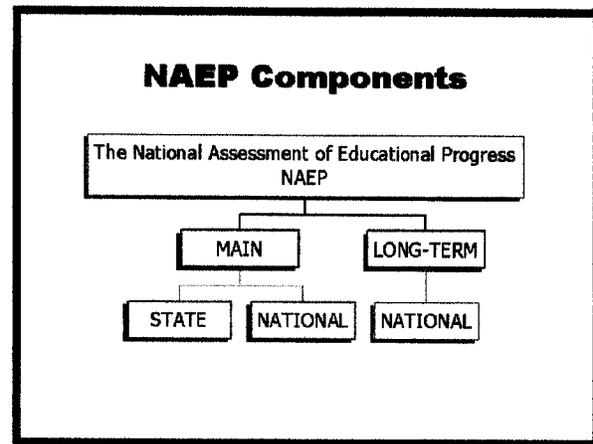
BACKGROUND

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

As mandated by Congress, NAEP surveys the educational accomplishments of U.S. students and monitors the changes in those accomplishments. NAEP, often called the “Nation’s Report Card,” is described as the only nationally representative and continuing assessment of what America’s students know and can do in various subjects. NAEP provides a comprehensive measure of students’ learning at critical junctures in their school experience. The assessment has been conducted regularly since 1969 and for over 30 years NAEP has been collecting data to provide educators and policymakers with accurate and useful information. Because NAEP makes objective information about student performance available to policymakers at national and state levels, it plays an integral role in evaluating the conditions and progress of the nation’s education.

Over the years, NAEP has evolved to address questions asked by policymakers, and NAEP now refers to a collection of national and state assessments. The collection of assessments includes main NAEP (state and national) and long-term trend NAEP (national).

The main assessments report results for grade samples of fourth, eighth, and twelfth grade students. They periodically measure students’ achievement in reading, mathematics, science, writing, U.S. history, civics, geography, and other subjects. In 1997, main NAEP returned to annual assessments. In 2000, the main NAEP assessed mathematics and science at grades four, eight, and twelve and reading at grade four.



The long-term trend assessments report results for age/grade samples (nine year-olds/fourth grade; thirteen year-olds/eighth grade; and seventeen year-olds/eleventh grade). They measure students’ achievement in mathematics, science, reading, and writing. Measuring trends of student achievement, or change over time, requires the precise replication of past procedures. Therefore, the long-term trend instrument does not evolve based on changes in curricula or in educational practices. In 1999, the long-term trend assessment began to be administered on a four-year schedule and in different years from the main national and state assessments in mathematics, science, reading, and writing. As a result, in 2000, this assessment was not administered.

Initiated in 1990, state assessments enable participating states to compare their results with those of the nation and other participating states. Because the national NAEP samples (main and long-term trend) were not designed to support the reporting of accurate and representative state level results, Congress authorized state assessments. State assessments have separate representative samples of students selected for each jurisdiction that participates, to provide these jurisdictions

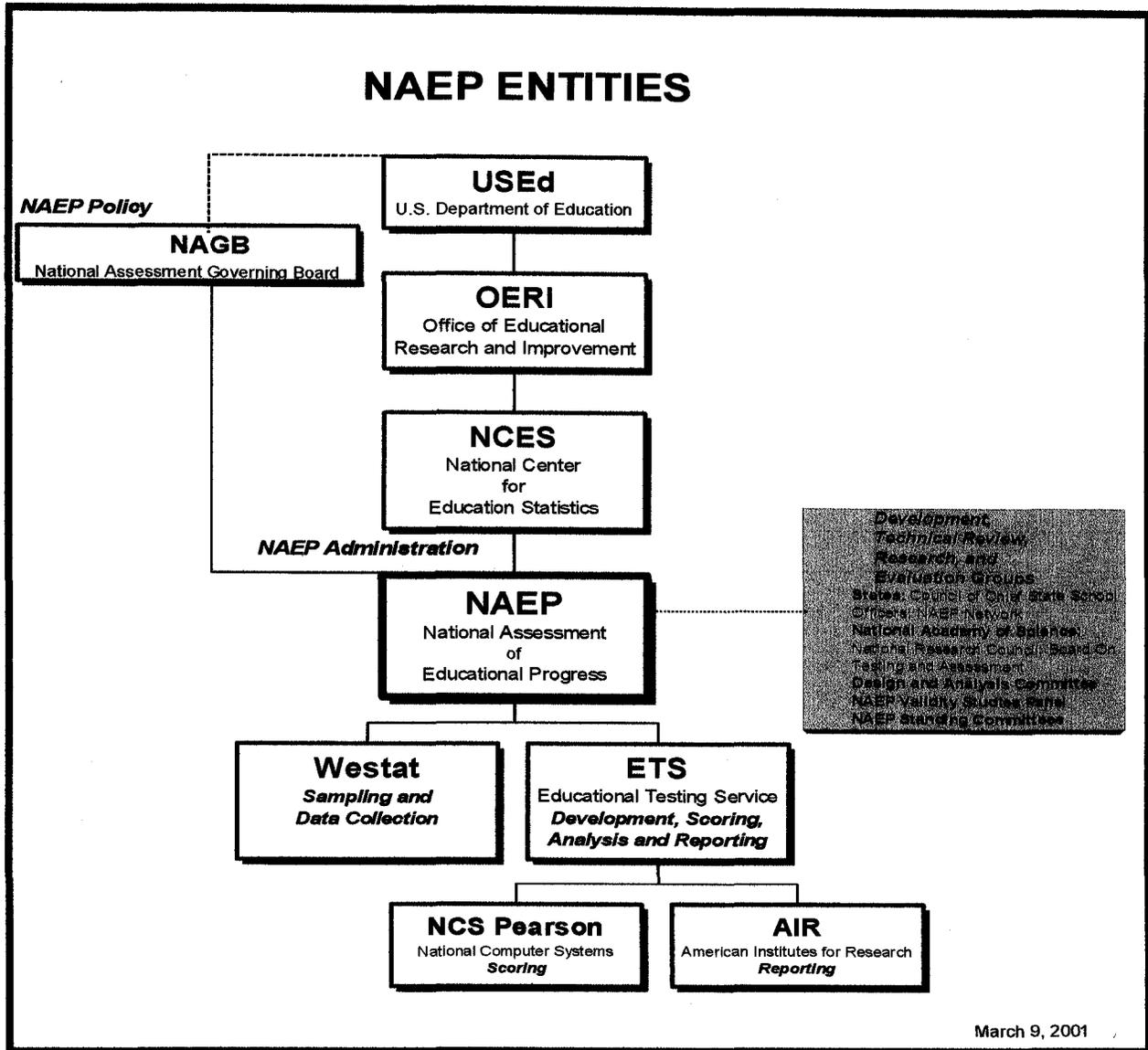
with reliable state level data concerning the achievement of their students. The main national and state assessments use the same assessment booklets. The state NAEP assessment is administered in every even year. In 2000, the state NAEP assessed mathematics and science in grades four and eight.

NAEP has two major goals: to reflect current educational and assessment practices and to measure change reliably over time. To meet these dual goals, NAEP selects nationally representative samples of students who participate in either the main assessments or the long-term trend assessments. These two assessments report information for the nation and for specific geographic regions of the country (Northeast, Southeast, Central, and West). These assessments use distinct data collection procedures, separate samples of students, and test instruments based on different frameworks. The results are also reported separately.

Participation in NAEP 2000 was voluntary for states, school districts, schools, teachers, and students. Some state legislatures mandated participation; others left the option to participate to their superintendents and other educational officials at the local level. Other states chose not to participate. Before any student selected to participate actually took the test, the student's parents decided whether or not their child would do so. Under the *No Child Left Behind* Act, NAEP participation is mandatory for all recipients of Title I funds.

NAEP assessments used a combination of multiple-choice and constructed response questions. The multiple-choice questions are electronically scanned and scored. Professional scorers evaluate the constructed response questions. The assessments are not designed to provide individual student scores. Each student receives only a small portion of the assessment. The assessment sessions last 45 to 90 minutes depending on the subject. The entire assessment process, from administering the assessments, to analyzing and reporting the results, can take anywhere from 9 to 18 months.

Since 1983, the Department of Education (Department) has conducted NAEP through a series of contracts, grants, and cooperative agreements with various entities. The following chart depicts the relationship of these entities for the audit period October 1, 1999, through September 30, 2000.



NATIONAL CENTER FOR EDUCATION STATISTICS

The Commissioner of Education Statistics, who heads the NCES in the Department, is responsible, by law, to carry out the NAEP project through competitive awards to qualified organizations. NCES establishes agreements with private companies for test development and administration services. NCES publishes the results of the NAEP assessments and releases them to the media and public. NCES strives to present this information in the most accurate and useful

manner possible, publishing reports designed for the general public and specific audiences and making the data available to researchers for secondary analyses.

NATIONAL ASSESSMENT GOVERNING BOARD

In 1988, Congress established the NAGB to formulate policy guidelines for NAEP. The NAGB, appointed by the Secretary of Education but independent of the Department, governs the program. It is authorized to set policy for the NAEP. NAGB selects the subject areas to be assessed, develops guidelines for reporting, and gives direction to NCES. It is required by law to approve all assessment questions and review the scoring guides. NAGB monitors the field-testing process and may suggest changes in assessment questions.

WESTAT

NCES has a cooperative agreement with Westat. Under this agreement, Westat selects the school and student samples, trains assessment administrators, and manages field operations (including assessment administration and data collection activities). For the national assessment, Westat administers the assessments and for the state assessment, the individual states administer the assessments. For the state assessments, Westat conducts quality control monitoring of the assessment administration by either sending staff to schools or calling the state administrators.

EDUCATIONAL TESTING SERVICE

NCES has an agreement with ETS. Since 1983, NCES has conducted the assessment through a series of contracts, grants, and cooperative agreements with ETS. Under these agreements, ETS is responsible for developing the assessment instruments, scoring student responses, analyzing the data, and reporting the results. ETS scores the multiple-choice questions and subcontracts with NCS to score the constructed response questions. ETS analyzes the scoring data and summarizes the results. ETS then drafts reports for NCES to review and approve.

NCS PEARSON

NCS, which serves as a subcontractor to ETS, is responsible for printing and distributing the assessment materials and for scanning and scoring constructed response questions. NCS handles all receipt control, data preparation and processing, scanning, and scoring activities. NCS performs optical scanning of multiple-choice selections, handwritten responses, and other data. This image based scoring system eliminates paper in the scoring process, which also permits on-line monitoring of scoring reliability and creation of recalibration sets.

AMERICAN INSTITUTES FOR RESEARCH

American Institutes for Research (AIR), which serves as a subcontractor to ETS, is responsible for development of the background questionnaires. Students, teachers, and principals complete these questionnaires to provide NAEP with data about students' school backgrounds and educational activities. Students answer questions about the courses they take, homework, and home factors related to instruction. Teachers answer questions about their professional qualifications and teaching activities, while principals answer questions about school level practices and policies. Relating student performance on the cognitive portions of the assessments to the information gathered on the background questionnaires increases the usefulness of NAEP

findings and provides the context for a better understanding of student achievement. AIR did not perform work related to our audit objectives; therefore, it was not included in our review.

NO CHILD LEFT BEHIND ACT

On January 8, 2002, President Bush signed into law the *No Child Left Behind* Act of 2001. This new law represents his education reform plan and contains the most sweeping changes to the Elementary and Secondary Education Act since it was enacted in 1965. It changes the federal government's role in kindergarten through grade 12 education by asking America's schools to describe their success in terms of what each student accomplishes. The act contains the President's four basic education reform principles: stronger accountability for results, increased flexibility and local control, expanded options for parents, and an emphasis on teaching methods that have been proven to work.

An "accountable" education system involves several critical steps:

- States create their own standards for what a child should know and learn for all grades. Standards must be developed in math and reading immediately. Standards must also be developed for science by the 2005-06 school year.
- With standards in place, states must test every student's progress toward those standards by using tests that are aligned with the standards. Beginning in the 2002-03 school year, schools must administer tests in each of three grade spans: grades 3 through 5, grades 6 through 9, and grades 10 through 12 in all schools. Beginning in the 2005-06 school year, tests must be administered every year in grades 3 through 8 in math and reading. Beginning in the 2007-08 school year, science achievement must also be tested.
- Each state, school district, and school will be expected to make adequate yearly progress toward meeting state standards. This progress will be measured for all students by sorting test results for students who are economically disadvantaged, are from racial or ethnic minority groups, have disabilities, or have limited English proficiency.
- School and district performance will be publicly reported in district and state report cards. Individual school results will be on the district report cards.
- If the district or school continually fails to make adequate progress toward the standards, then they will be held accountable.

The *No Child Left Behind* Act required changes in the NAEP assessment schedule. As a result, state participation in NAEP reading and mathematics biennial assessments in grades four and eight is required of states participating in Title I. Previously, state NAEP reading and mathematics was performed on a four-year cycle.

GOVERNMENT PERFORMANCE AND RESULTS ACT

This audit falls under the context of the GPRA, specifically data quality and reliability. To report the NAEP results, data needs to be accurate, complete, and timely, because the Department's programs rely on NAEP as a data source. The Department's 2000 Performance Report objectives identified Department goals and individual programs that relied on NAEP. Department goals 1 and 2 had objectives that relied on NAEP as a data source. There were six individual programs that contained objectives that relied on NAEP as a data source. The individual programs included: (1) Title I Grants for Schools Serving At-Risk Children, (2) Educational Technology State Grants, (3) State Assessments, (4) Indian Education, (5) Grants to States and Preschool Grants Programs - IDEA Part B, and (6) Perkins Vocational and Technology Education. While the Strategic Plan for 2002-2007 has changed significantly, NAEP is still used extensively as a data source.

NAEP MANAGEMENT CONTROLS OVER SCORING ARE ADEQUATE

The management controls over scoring of the NAEP 2000 assessment were adequate and generally working as intended for the period October 1, 1999, through September 30, 2000. Our audit work confirmed that the management controls provided reasonable assurance that the assessment results could be relied upon. However, our audit work did identify two nonmaterial weaknesses regarding mathematics qualification sets and scorer qualifications. We recommend that the Director of the Institute of Education Sciences (formerly Office of Educational Research and Improvement) instruct NCES to (1) improve its monitoring of ETS and NCS for adherence to the terms of its cooperative agreements and (2) require NCS to use a qualification set of papers for mathematics and document that the scorers passed a qualification set of papers. We also noted that state assessments required under the *No Child Left Behind* Act could benefit from standards for management controls over scoring. We plan to report on this separately. This report highlights the management controls. These controls are more comprehensive than presented here. For additional details regarding these management controls see Attachment 1.

Monitoring

For monitoring management controls, we considered NCES' monitoring of its NAEP Cooperative Agreements with ETS and Westat. We also considered ETS' monitoring of its sub-contract with NCS. NCES monitors its NAEP Cooperative Agreements with Westat and ETS through periodic meetings and reports. NCES officials informed us that its NCS monitoring is limited due to travel funds. ETS also monitored NCS through periodic meetings and reports. In addition, ETS monitored the scoring process through on-site assessment experts during the constructed response scoring at NCS. Our review of monitoring management controls disclosed they were adequate except for two nonmaterial weaknesses.

NCES needs to improve its monitoring to ensure adherence to the terms of the NAEP Technical Application. Our review of monitoring management controls disclosed two nonmaterial weaknesses where the terms of the NAEP 2000 Technical Application were not met. These weaknesses included mathematics qualification sets and scorer qualifications.

Mathematics Qualification Sets

NCS did not use and/or document mathematics qualifying sets for training on extended constructed response questions as required in the NAEP 2000 Technical Application. Extended constructed response questions are defined as questions worth four points or higher. According to the NAEP 2000 Technical Application, Chapter 14, page 10, before scoring live responses to extended constructed response questions, each scorer must pass a qualification set of papers to ensure that he or she was able to score with the acceptable level of reliability.

The audit disclosed an End of Project Report document that indicated, "NAEP Math did not use any qualifying sets for training so everyone that was trained scored. Only two people were

released due to poor performance." In addition, an NCS employee informed us that ETS made the decision that no qualifying sets would be used for mathematics.

ETS and NCS officials informed us that practice papers, rather than formal qualification sets, were used to ensure that scorers were able to score with an acceptable level of reliability. However, the use of practice papers for this purpose was not documented. While the quality of scoring was high, it may have been higher had NCS met the requirement for each scorer to pass a qualification set of papers to ensure that he or she was able to score with the acceptable level of reliability. ETS and NCS officials indicated that in the future only sets explicitly identified as qualification sets would be used for qualification and that a strict record of qualification performance would be kept.

Scorer Qualifications

Our audit work also disclosed that some scorers did not meet scorer qualification requirements. We interviewed 14 scorers of which 12 scored at the grade 12 level. Of these 12 scorers, 8 did not meet the scorer qualification requirements for assessments at the grade 12 level outlined in the NAEP 2000 Technical Application. According to the NAEP 2000 Technical Application, Chapter 14, pages 6 and 7, scorers had to have the following qualifications:

- a minimum of a bachelor's degree in an appropriate academic discipline, such as mathematics, science, English, or education, and
- demonstrable ability in performance assessment scoring, with
- teaching experience at the elementary or secondary level preferred.

For assessments at the grade 12 level, special academic experience in the subject being assessed was required. For example, to score the grade 12 science assessment, scorers needed to have high school science teaching experience, or a university or graduate degree in science or science education.

ETS and NCS officials informed us that the available work force at that time could not meet the qualification requirements for the grade 12 level. In the Spring 2000 marketplace, individuals with degrees in mathematics, science, and closely related fields, were in high demand and those interested in short-term positions scoring NAEP were difficult to find. The officials also indicated that a formal process for exceptions to the qualification requirements should have been implemented to allow for authorization by NCES. While the quality of scoring was high, it may have been higher had NCS met the qualification requirements for the grade 12 level. Changes to the new NAEP Cooperative Agreement removed the qualification requirements. However, NCES could improve its monitoring to ensure adherence to the terms of the Agreement.

Recommendations

We recommend that the Director of the Institute of Education Sciences (formerly Office of Educational Research and Improvement) instruct NCES to:

- 1.1 Improve its monitoring of ETS and NCS for adherence to the terms of the cooperative agreements.
- 1.2 Require NCS to use a qualification set of papers for mathematics and document that the scorers' passed a qualification set of papers.

Receipt and Control Process

Our review of the management controls related to the receipt and control process focused on the roles of Westat and NCS in ensuring that all assessment booklets sent to the participating schools were accounted for and returned to NCS for inclusion in its scoring database.¹ The receipt and control process used by Westat and NCS provided reasonable assurance that all assessment booklets sent to the selected schools were accounted for and returned to NCS.

Scoring

Our review of the scoring management controls considered the roles of ETS and NCS in ensuring that the (1) correct constructed response rubric and multiple-choice answer keys were used, (2) scorer qualification requirements were met, (3) scorers were trained, and (4) scorers were monitored for reliability to ensure the scoring of each question was consistent among the scorers and over time. NCS was responsible for scoring the constructed response questions and ETS was responsible for scoring the multiple-choice questions. ETS performed quality assurance steps before the assessments were conducted that are related to scoring. These steps included independent verification of multiple-choice question keys, review of constructed response questions and scoring rubrics, and review of all multiple-choice and constructed response questions by members of NAEP subject area committees. Before scoring live responses to extended constructed response questions, each scorer must pass a qualification set of papers to ensure that he or she is able to score with the acceptable level of reliability. In addition, ETS and NCS selected training materials for constructed response scoring, which included anchor, practice, calibration, and qualification papers for each response to be scored and final scoring rubrics.² NCS used these papers to provide scorer training prior to actual scoring of constructed response questions.

During scoring NCS used four methods to monitor reliability. These methods included calibration, backreading, interrater reliability, and trend scoring.³ We determined that NCS'

¹ For additional information on Receipt and Control Process see Attachment 1, page 1.

² For additional information on Scoring see Attachment 1, pages 1 and 2.

³ Ibid.

monitoring reliability methods provided reasonable assurance of scoring quality and that it met the minimum standards for NAEP 2000 regarding interrater reliability.

ETS performed on-site monitoring at NCS during constructed response scoring. This included monitoring interrater reliability reports, *t*-tests, frequency distributions of scores, and the rate of scoring.⁴ NCS also used these monitoring tools. The on-site monitoring kept NAEP management informed of scoring issues or problems.

Data Quality

For data quality management controls, we considered the roles of ETS and NCS. Our examination was based on interviews and review of documentation.⁵ ETS performed quality assurance before the assessments were conducted, on-site monitoring at NCS during constructed response scoring, database quality assurance on the scoring database during scoring and after scoring is completed, and quality assurance steps undertaken as part of statistical analysis of data. NCS performed quality assurance when scanning the assessment booklets into the database for image scoring, during scoring, and prior to data delivery to ETS. We also examined computer-processed data for reliability.

The quality assurance steps performed by ETS before the assessments were conducted related to pre-field testing the review process, field-testing the assessments, and preparing a thorough scoring planning memorandum ensured that meaningful data would be obtained. If multiple-choice questions lack single correct answers, or if constructed response questions do not have solid scoring rubrics, then no scoring or analysis process, no matter how carefully planned and executed, will yield meaningful data.

ETS performed quality assurance steps before the assessments were conducted that were related to data quality. These steps were designed to ensure multiple-choice questions have a single correct answer and constructed response questions have a solid scoring rubric in order to yield meaningful data.

The on-site monitoring performed by ETS was instrumental in ensuring the quality of the scoring data as constructed response scoring was being performed. The various reports monitored while on-site would identify problems with data quality before scoring was completed and the scoring data sent to ETS.

ETS database quality assurance involved steps taken once the assessment data was sent to ETS. Many of these steps were designed to ensure that the data has expected characteristics and meets the basic quality standards before analysis work is completed.

The NCS data quality assurance steps included scanning, scoring, and data delivery. The NCS scanning process provided reasonable assurance that the data entered into the database was

⁴ Ibid.

⁵ For additional information on Data Quality see Attachment 1, pages 2 through 6.

accurate. The NCS data quality assurance steps for scoring and data delivery ensured that the data was accurate.

For data quality we examined computer-processed data for reliability. Our testing for data reliability, focused on assessing the competency of the data. To determine data reliability we assessed data completeness, data authenticity, and the accuracy of computer processing. For details on our testing see the Objective, Scope, and Methodology section.

As part of our data completeness work, we tested management controls over scoring by examining the NCS mainframe final data for anomalies, identifying the scorers for each subject, and judgmentally selected scorers to interview. Our examination of the data for anomalies considered many issues, such as (1) the number of scorers by question, subject, scoring date, and scorer identification number; (2) various scoring scenarios; and (3) scorer consistency by question and identification number. We identified and reconciled the number of NAEP 2000 constructed response and multiple-choice questions, number of scorers, and time period for scoring the constructed response questions to various documents provided by ETS and NCS.

Our examination of the NCS data for anomalies disclosed no issues of concern. Our reconciliation of the above information to various documents provided by ETS and NCS disclosed that they generally reconciled.

The review of data quality management controls disclosed no concerns regarding its reliability. The quality assurance steps performed by ETS and NCS disclosed no concerns and provided reasonable assurance that the data was reliable. Our testing for data reliability regarding data completeness, data authenticity, and the accuracy of computer processing disclosed no concerns. We compared 100 percent of the scoring data from the NCS mainframe final data to the ETS Secondary User data. Our testing confirmed that ETS processed all NAEP 2000 scoring data properly once it received the data from NCS. In addition, we found no anomalies in the NCS data that caused concern. Our testing disclosed that the ETS Secondary User database accurately reflected the source records. We determined that the number of assessments received by ETS and available for use in the Nation's Report Card generally met the Westat sample requirements.

Analysis and Reporting

For analysis and reporting management controls, we considered the role of ETS. Our examination was based on interview and review of documentation.⁶ Quality assurance steps undertaken as part of statistical analysis of data and preparation of reports included three distinct sets of quality assurance processes. These included a system of formal procedural and statistical checks on the data analysis process, a thorough series of plausibility checks, and quality assurance of NAEP reports. However, the reporting process was outside the scope of this audit so we did not perform work in this area.

⁶ For additional information on Analysis and Reporting see Attachment 1, page 6.

Our review of the quality control steps undertaken as part of statistical analysis of the NAEP 2000 data disclosed that the steps were adequate. The procedural and statistical checks on the data analysis process should provide reasonable assurance that any data abnormalities were caught and resolved prior to reporting on the data. The quality controls were augmented with computerized checking that should reduce the likelihood of human error in the process. The plausibility checks, which compare data to expectations, historical precedent, and data obtained through other analysis methods, were designed to make sure the data "makes sense", and thereby further increase the reliability of the data. The statistical analysis process used by ETS provides reasonable assurance that the data accurately reflects the NAEP 2000 scoring results.

Other Issues

We considered other issues that might affect the management controls such as incentive payments for scorers and NCS' NAEP profit margin. To determine whether these issues were of concern and whether management controls were working as intended we (1) interviewed 14 NCS scorers; (2) reviewed NCS position descriptions for a scoring director, a scoring supervisor, a trainer, and a scorer; (3) reviewed 3 NCS scorer personnel files; (4) reviewed 4 NCS scorer payroll records; and (5) examined NCS' December 2000 accounting records regarding its NAEP profit margin. Our work disclosed no concerns regarding incentive payments for scorers or NCS' NAEP profit margin.

OTHER MATTERS

The state assessments required as a result of the *No Child Left Behind* Act might benefit from the NAEP management controls. In addition, to the biennial assessments required under NAEP, the *No Child Left Behind* Act requires schools receiving Title I funds to have annual state assessments in mathematics and reading in three grade spans beginning in the 2002-03 school year. Beginning in the 2005-06 school year, assessments must be administered every year in grades three through eight in mathematics and reading. States create their own standards for each subject and must assess every student's progress toward those standards. We believe that each state's design of this assessment should include some minimum level of management controls over scoring for uniformity. The Department should consider whether the types of management controls over scoring used for NAEP are appropriate for state assessments. We plan to report on this separately.

OBJECTIVE, SCOPE, AND METHODOLOGY

Our audit objectives were to determine whether management controls over scoring of the NAEP 2000 assessment were in place and adequate to provide reasonable assurance that the assessment results can be relied on during the period October 1, 1999, through September 30, 2000. To accomplish our audit objectives we

1. interviewed officials from NCES, NAGB, Westat, ETS, and NCS to gain an understanding of their role in conducting NAEP;
2. reviewed and tested management controls over scoring to ensure the processes were working as intended;
3. reviewed and tested the ETS and NCS computer processed data to ensure that it was reliable;
4. reviewed background materials related to NCES, NAGB, Westat, ETS, and NCS, such as:
 - a. The NAEP Guide, 1999 Edition
 - b. ETS Standards for Quality and Fairness 2000
 - c. ETS NAEP 2000 Technical Application
 - d. NAEP 1998 Technical Report
 - e. Special Provisions Cooperative Agreement
 - f. ETS Subcontract with NCS Pearson
 - g. NCES Handbook of Survey Methods, September 2001
 - h. NCES Statistical Standards, June 1992 and draft May 2002
 - i. *No Child Left Behind Act*
 - j. Department's 2000 Performance Report
 - k. NCES Statistics and Assessment
 - l. Government Performance and Results Act of 1993
 - m. Federal Managers Financial Integrity Act of 1982
 - n. Chief Financial Officers Act of 1990
 - o. Government Management and Reform Act 1994
5. gained an understanding of current Administration and Congressional proposals that could have an affect on NAEP, such as GPRA, the *No Child Left Behind Act*, the Elementary and Secondary Education Act, and other legislation effecting management controls;
6. gained an understanding of the ETS scoring analysis and reporting process;
7. reviewed selected NCS employee payroll, personnel file, and position description records and NCS' NAEP profit margin; and
8. gained an understanding of state assessments through interviews with ETS and NCS officials.

To review management controls over scoring, we interviewed officials to identify the management controls in place and reviewed various documents used in the process. To test the management controls, we examined the NCS mainframe final database for anomalies, identified the scorers for each subject, and interviewed judgmentally selected scorers. We also examined the NCS data to determine if NCS met the minimum interrater reliability standards and second scoring

requirements. To determine whether management controls were working as intended and whether there were other issues of concern, we judgmentally selected 14 scorers to interview. We selected the sample from a universe of 46 reading, 211 mathematics, and 273 science scorers.⁷

Reliability of Computer-Processed Data

To accomplish our objectives, we relied on computer-processed data. To determine the reliability of that data, we assessed data completeness, data authenticity, and the accuracy of computer processing. We tested data completeness to confirm that the universe contained all scoring data elements relevant to our audit objectives and that the data transfer from NCS to ETS was accurate. To test for data completeness, we compared 100 percent of the scoring data from the NCS mainframe final data to the ETS Secondary User data. We also compared the number of national and state NAEP sample assessment booklets requested by Westat for each academic area and grade level to the number of assessment booklets NCS printed and distributed, and to the number of assessment booklets received by ETS as assessed. See table below for details.

SESSION	SAMPLE SIZE NATIONAL/STATE	PRINTED/ISSUED	ASSESSED NATIONAL/STATE
Grade 4 Reading	8,000/0	24,000/12,000	8,504/0
Grade 4 Mathematics	13,750/112,500	208,000/189,375	14,396/101,764
Grade 4 Science	15,750/112,500	222,000/192,376	16,749/96,935
Grade 8 Mathematics	15,750/112,500	208,000/192,375	16,846/97,509
Grade 8 Science	15,750/112,500	222,000/192,376	16,837/94,055
Grade 12 Mathematics	13,750/0	39,000/20,625	14,130/0
Grade 12 Science	15,750/0	55,500/23,626	15,879/0

⁷ For our judgmental sample selection, we selected scorers from each academic area (reading, mathematics, and science) and scorer position description (trainer, scoring director, supervisor, scorer). We also selected scorers that tended to score a higher number of questions than other scorers and/or scored in more than one academic area.

The results of our testing confirmed that ETS processed all NAEP scoring data properly once it received the data from NCS and that we are reasonably certain that the data is complete. We determined that the number of assessments received by ETS and available for use in the Nation's Report Card met the Westat sample requirements for the national assessments and for states that participated. The ETS Secondary User database was used for analysis and reporting of NAEP results.

Our testing of data authenticity determined if the computer data accurately reflected the source records. To test data authenticity, we randomly selected a sample of 35 assessment booklet records from the ETS Secondary User database and compared various scoring data to the actual assessment booklets. We randomly selected 5 assessment booklets from each subject grade level in the national and state NAEP 2000. The sample universe, subject, and grade levels included 8,504 reading – 4th; 116,160 mathematics – 4th; 114,355 mathematics – 8th; 14,130 mathematics – 12th; 99,570 science – 4th; 104,928 science – 8th; 15,009 science – 12th. The actual universe for science grades 4, 8, and 12, respectively, were 113,684, 110,892, and 15,879. Our testing disclosed no errors in the scoring data and that the correct constructed response rubrics were used with each question. For multiple-choice questions the scores in the ETS database accurately reflected the assessment booklet bubble answer. A bubble answer is the question answer circle that the student must fill in. In addition, the range for bubble answers in the ETS database accurately reflected the range for bubble answers in the assessment booklets. For example, the assessment booklet question may provide answer selections that ranged from A through D, therefore the ETS database should also provide for selections that ranged from A through D. Also, the correct answer in the ETS database accurately reflected the rubric correct answer key. For constructed response questions we determined that the score point given for the question response fell within the acceptable rubric range for that question. We did not test to see if the student was given the correct score for the constructed response questions because scoring is subjective and may vary depending on the scorer. In addition, we did not test the cluster type questions because they are a variation of multiple-choice and constructed response questions that would have the same subjective nature as the scoring for the pure constructed response questions. Our testing of data authenticity also included tracing the data from the ETS Secondary User database back to the NCS data to ensure there were not any extra ETS Secondary User data records that were not supported by NCS data. Our testing disclosed the exact same number of assessment booklet records in the ETS Secondary User database as there were in the NCS data. Based on our testing we believe that the ETS database is reliable and accurate.

The steps aimed at the accuracy of computer processing were designed to verify that all relevant records were completely processed and that computer processing met the intended objectives. To verify that all relevant records were completely processed, we performed a 100 percent test of data elements, as discussed above, and verified the conversion of items from the NCS database to the ETS Secondary User database was performed accurately. ETS converted some NCS data elements, such as scoring labels and question names. Our testing disclosed that all relevant records were completely processed and accurately converted to meet the intended objectives.

Organizations and Locations

We conducted our audit at (1) NCES' offices in Washington, DC, on October 10, 2001, December 18, 2001, and November 13, 2002; (2) NAGB's offices in Washington, DC, on December 19, 2001; (3) Westat in the NCES offices in Washington, DC, on December 20, 2001; (4) ETS' offices in Princeton, NJ, from February 26, 2002, through March 6, 2002; (5) NCS' offices in Iowa City, IA, from April 2, 2002, through June 6, 2002; and October 22, 2002. We held an exit conference with officials from NCES, ETS, and NCS on November 13, 2002. We performed our audit work in accordance with generally accepted government auditing standards appropriate to the scope of review described above.

STATEMENT ON MANAGEMENT CONTROLS

We have made a study and evaluation of the management control structure over scoring of the NAEP 2000 assessment for the period October 1, 1999, through September 30, 2000. Our study and evaluation was conducted in accordance with generally accepted government auditing standards. For the purpose of this report, we assessed and classified the significant management control structure into the following categories:

- Monitoring
- Receipt and Control Process
- Scoring
- Data Quality
- Analysis and Reporting

The scoring of NAEP is a collaborative effort by several entities, which include NCES, Westat, ETS, and NCS. The above listed categories of significant management control structures are a combined effort of these entities. So, the management of these entities is responsible for establishing and maintaining the scoring management control structure. In fulfilling this responsibility, estimates and judgments by the entities' management are required to assess the expected benefits and related costs of control procedures. The objectives of the system are to provide management with reasonable, but not absolute, assurance that assets are safeguarded against loss from unauthorized use or disposition and that the transactions are executed in accordance with management's authorization and recorded properly, so as to permit effective and efficient operations.

Because of inherent limitations in any management control structure, errors or irregularities may occur and not be detected. Also, projection of any evaluation of the system to future periods is subject to the risk that procedures may become inadequate because of changes in conditions, or that the degree of compliance with the procedures may deteriorate.

In our opinion, the management control structure over scoring of the NAEP 2000 assessment for the period October 1, 1999, through September 30, 2000, taken as a whole, was sufficient to meet the objectives stated above insofar as those objectives pertain to the prevention or detection of errors, irregularities or inefficiencies that would be material in relation to the reliability of the assessment results.

Nonmaterial weaknesses, which in the auditors' judgment are reportable conditions, are included under the NAEP MANAGEMENT CONTROLS OVER SCORING ARE ADEQUATE section.

ADDITIONAL MANAGEMENT CONTROL DETAIL NOT PRESENTED IN THE BODY OF THE REPORT

Receipt and Control Process

Westat used an Administration Schedule as a control document for the assessments. The Administration Schedule is used to select the schools, students, and assessments given for testing. NCS bar coded the assessment booklets, which allowed it to identify which booklets were sent to each school and assigned to which students. The assessment administrators verify they have received all materials from NCS. After the assessment is conducted, the assessment administrator accounts for all assessment booklets and updates the Administration Schedule using the appropriate administration codes. The Administration Schedule and the assessment booklets are returned to NCS for scoring. All booklets are returned to NCS. NCS has a schedule of all assessments. If it does not receive the assessment booklets in a timely manner, it contacts Westat. Westat then uses the FedEx tracking system to locate the booklets. All boxes of assessment booklets received by NCS are scanned using pre-printed shipping labels NCS provided for the return of the assessment materials. NCS opens the boxes and verifies the contents. NCS compares the distribution file to the receipt file in order to determine if all assessment booklets were returned.

Scoring

ETS and NCS selected training materials for constructed response scoring, which included anchor, practice, calibration, and qualification papers to provide scorer training prior to actual scoring of constructed response questions. An anchor set of papers is a collection of questions from prior years with the score reported to illustrate the scoring for that question. A practice set of papers is a collection of questions from prior years without the score reported. The scorer will score each question then the trainer reviews and indicates the correct score along with an explanation for the score. A qualification set of papers is a collection of questions from prior years without the score reported, which the scorer will score and the trainer will grade.

During scoring NCS used four methods to monitor reliability. These methods included calibration, backreading, interrater reliability, and trend scoring. Scorers performed periodic calibration scoring to make sure that similar answers to the same question were scored consistently. To prevent drift, whenever the scorers took a break longer than 15 minutes they scored a set of calibration papers to refresh their training and reinforce the scoring criteria. During backreading, scoring supervisors reviewed each scorer's work to ensure that the scorer applied the scoring criteria consistently across a large number of responses and over time. NCS officials indicated that scoring supervisors evaluated about 10 percent of each scorer's work in progress. NCS also used reliability scoring, often referred to as interrater reliability, to maintain uniformity of scoring and to ensure that scorer agreement rates met minimum standards. For interrater reliability, a second

rater scores a sample of questions and the agreement between the first and second scores is compared. If the interrater reliability does not meet minimum standards, that entire question set is re-scored. An ETS official stated that the minimum standards for NAEP 2000 were 75 percent for a four point or more question, 80 percent for a three point question, and 90 percent for a two point question. Six percent of grades four and eight mathematics and science constructed responses and 25 percent of grade four reading and grade 12 mathematics and science constructed responses were required to be scored by a second scorer to obtain statistics on interrater reliability. NCS used trend scoring to ensure scoring was consistent across years. Trend scoring included steps and checks to ensure that scoring decisions were consistent with those made in earlier years. For each trend question used in a previous NAEP cycle, a minimum number of responses in the base year were scored along with the NAEP 2000 responses. The scoring system compared the scores assigned in the original cycle with those assigned in NAEP 2000 to determine comparability of scoring across years. We determined that NCS' methods for monitoring reliability provided reasonable assurance of scoring quality and that it met the minimum standards for NAEP 2000 regarding interrater reliability.

ETS performed on-site monitoring at NCS during constructed response scoring. This included monitoring interrater reliability reports, *t*-tests, frequency distributions of scores, and the rate of scoring. NCS also used these monitoring tools. Interrater reliability reports were reviewed daily to provide immediate feedback to the scorers and correct any scoring difficulties. During the scoring of trend questions, a *t*-test was performed. If the *t*-test was outside the acceptable range of +/- 1.5 of zero, scoring was stopped in order to determine a plan of action. Generally, the *t*-test compares the mean score this time with the mean score from a previous time. If the scorer did not pass, the scorer would be retrained. For each question, a report could be run that showed the frequency distribution of the scores. This report indicated the separate frequencies for first and second scores. The rate of scoring could be monitored using a status tool that displayed the number of responses scored, the number of responses first scored that still needed to be second scored, the number of responses remaining to be first scored, and the total number of responses remaining to be scored. This allowed for accurate monitoring of the rate of scoring and to estimate the time needed for completing the various phases of scoring. The on-site monitoring kept NAEP management informed of scoring issues or problems.

Data Quality

ETS performed quality assurance steps before the assessments were conducted that are related to data quality. These included:

- Pre-field testing the review process that includes independent verification of multiple-choice answer keys; review of constructed response questions and scoring rubrics; and review of all multiple-choice and constructed response questions by members of NAEP subject area committees and measurement specialists, the Instrument Development Committees, NCES, and NAGB.

- Field-testing of all assessments prior to selection for operational use which includes administering all potential NAEP assessments to a sample of 500 students, evaluating the functioning of constructed response rubrics, and statistical checks to identify problems in keying of multiple-choice assessments.
- Preparing a scoring planning memorandum that details for NCS the overall structure of the scoring process, ETS statistical and data requirements, and a summary of scoring completion and data delivery dates.

The ETS database quality assurance involved steps taken once the assessment data was sent to ETS. Many of these steps were designed to ensure that the data had expected characteristics and met the basic quality standards before analysis work was completed. The database quality assurance procedures included:

- Test runs of the database using preliminary data received from NCS and Westat.
- Review of (a) sampling weights received from Westat, (b) scoring data sent by NCS, (c) sampled booklets to check accuracy of the optical mark reading system, and (d) special control files to check the accuracy of score assignments made in the NCS image-based constructed response scoring system.
- Resolution of any database issues or problems.
- Calculation of final scoring reliability figures for technical reporting.

The NCS data quality assurance steps included scanning, scoring, and data delivery. During the scanning process, assessment booklets are batched, scanned, and bar code read. An NCS official stated that NCS performed diagnostic tests on the scanning machines prior to each new production run. Each production run also included three quality assurance check sheets, which are documents placed in the batch and scanned along with the pages from the assessment booklets. NCS used Optical Mark Readers for scanning that also included intelligent character recognition. The scanning machine numbered each page scanned in case a page needed to be located later. During the scanning process, infrared was used to capture only needed information, such as students' handwriting, into the data file. The scanning process had two edit phases that included machine edits and image editing. Machine edits verified that each page of each assessment booklet was present and that each field had an appropriate value. The edit program checked each assessment booklet number, school code, and other data on the booklet cover for valid value ranges. The edit program then checked each block of the assessment booklet for validity and continued through each question within the block. Each piece of input data was checked to verify that it was of an acceptable type, that values fell within a specified range, and that it was consistent with other data values. Each scanning machine has built in recovery methods. Attached to the scanning machine was a portable computer that reported scanning errors.

NCS used image editing to scan pages that the main scanning process was unable to scan. If a document could not be scanned, the information was entered into the system manually. The image editing department also reviewed suspect errors on-line. A suspect error is an indication that an error may exist. Two individuals separately reviewed the suspect error and made a determination regarding its resolution. The two resolutions were then compared to determine if the individuals came to the same conclusion. The Administration Header Schedule (front cover of assessment booklet) has 100 percent verification of keyed items. The scanning system incorporated a program called the generalized batch editing system. This program generates reports of suspect errors. The error correction continues until all errors are corrected. The NCS scanning process provided reasonable assurance that the data entered into the database was complete and accurate.

NCS uses a scoring tool called Image Capture Environment (ICE). The ICE includes significant controls to ensure accurate scoring. To ensure that scanned images are matched with the appropriate scoring prompts, the system loads the scanned images into the database with control information, such as the type of booklet, question number, and book number. When the image is captured, it is tied to the control data. The scanned images, which are the responses from the assessment booklets, are merged with another file that contains the question and prompts used for scoring. The scanned image is called the "clip" and the merged file is called the "overlay." The clip is placed in the center and the overlay surrounds the clip. The scoring prompts would consist of the defined scoring system and the labels "B," "X," "IL," "?," and "OT." The defined scoring system could consist of correct or incorrect or some type of number point value, such as 1 for incorrect and 4 for correct. The labels are the special coding categories for unscorable responses.

To ensure that the appropriate overlay is matched with the right clip, NCS used control information contained in its databases. The ICE used four databases: scoring, application repository, workflow, and operational. The scoring database contained statistical information, the application repository database contained information about the overlay, the workflow database contained the scoring information, and the operational database contained information about the other databases. An NCS official identified the question to score and the clips for that question were loaded into the workflow database along with control information that identifies the individual clips. The ICE corresponds with the application repository (definition database) to determine the correct overlay to merge with the clip. The application repository defines the scoring and the labels to be used for the specific question. This includes the scoring rubric, which is used to set up the scorer shell dialog box. The ICE software tool gets five clips from the workflow database, attaches the overlays from the application repository, and sends the information to a scorer. The clips and overlays are loaded into their respective databases based on a scoring schedule.

The ICE tracks scoring, limits access to scoring batches, and runs edit checks. During scoring information was saved to a table in the workflow database. This information included the score, scorer's identification number, and a time stamp.

NCS had controls in place that limited access to make scoring changes, which enhanced data quality. During scoring a question may get scored by one to three different scorers. Some questions are scored a second time for interrater reliability and/or by a scoring supervisor. The scores in the workflow database are referred to as the reported 1st score, 2nd score, and original (supervisor) score. The 1st scorers could change their own scores and were allowed to go back five questions and make changes. A review queue held up to five questions. When a new question was added to the review queue, the oldest question moved out of the queue and the score was updated to the database. Once the score was updated to the database, the scoring supervisor was allowed access to backread the score. The scoring supervisor was allowed access to the scores up to four hours after the completion of the scoring batch, when the batch was closed. The database allowed only one 1st score to be recorded, zero or one 2nd score, and multiple supervisor scores. While the NCS database maintains the multiple supervisor scores, only the final supervisor score is included in the data files sent to ETS.

While NCS used interrater reliability to ensure scoring quality, it also had steps to ensure the quality and timeliness of the data. A table in the workflow database tracked the interrater reliability for questions so that the scoring supervisor could calculate the interrater reliability percentage. The interrater reliability table was constantly updated during the scoring process so that the scoring supervisor could calculate the interrater reliability any time. The calculation only included the 1st and 2nd scores. The five questions in the review queue were not included in the interrater reliability calculation. The interrater reliability percentage was calculated based on individual questions. An individual question may require more than one scoring batch.

NCS had data quality assurance steps for the batches to ensure that the data was accurate and complete. A batch identification number identified the scoring batch. The scoring batch remained in the database until the batch was completely scored. The completion of scoring was signified by a prescribed number of scores being entered. The scoring information was extracted from the database, and quality assurance edit checks were performed to ensure data was accurate and complete. The NCS data quality assurance steps for scoring ensured that the data was accurate and complete.

The NCS process for data delivery to ETS included steps to ensure ETS had all the needed data and that ETS knew which score to use for analysis and reporting. These steps included merging the data from the scoring batches into a file and determining which score was the official score. The scanned images were not included in the file. The optically read bubbles for the multiple-choice questions were combined with the score given by the scorer for constructed response questions in the file. NCS had previously scanned the assessment booklet multiple-choice bubble answers and converted them into number values for ETS to use for scoring. The files sent to ETS were separated by national, state, grade level, and subject. Separate files were created for scorer identification, question name, date question scored, and assessment booklet identification number. The ICE tracked all scorers for each question and identified the question scored. The scorer identification file contained the official score. When NCS created the scorer identification file, it determined which scores became the official scores. To make this

determination, NCS examined the reported 1st score and supervisor score. If the reported 1st score and the supervisor score were different, NCS made the supervisor score the official score. If there was no supervisor score, the reported 1st score became the official score. The scorer identification file sent to ETS included the reported 1st, 2nd, and original scores. Determining the official score ensured that ETS would always know which score to use in its analysis and reporting. The NCS data quality assurance steps for data delivery ensured ETS received all the needed data and that ETS knew which score to use for analysis and reporting.

As part of our data completeness work, we tested management controls over scoring by examining the NCS mainframe final data for anomalies, identifying the scorers for each subject, and judgmentally selected scorers to interview. We used the information below as part of our examination of the database for anomalies and reconciliation to various documents provided by ETS and NCS. Our review of the NCS database disclosed that there were:

SESSION	CONSTRUCTED RESPONSE QUESTIONS	MULTIPLE-CHOICE QUESTIONS	CONSTRUCTED RESPONSE SCORERS	CONSTRUCTED RESPONSE SCORING DATES
Grade 4 Reading	46	35	46	March 31, 2000–April 19, 2000
Grade 4 Mathematics	60	86	211	March 11, 2000 – May 28, 2000
Grade 8 Mathematics	62	98		
Grade 12 Mathematics	64	100		
Grade 4 Science	80	70	273	March 13, 2000 – June 8, 2000
Grade 8 Science	110	95		
Grade 12 Science	105	91		

Analysis and Reporting

The ETS system of formal procedural and statistical checks was designed to ensure that the data analysis followed the right steps in the right order and that data abnormalities were caught and resolved. These checks included item analysis, scorer reliability programs, item calibration, item plots, condition variable processing, and scale score estimation. ETS used a variety of automated programs to assist in performing these checks. The plausibility checks are a system of comparing data to expectations, historical

precedent, and data obtained through other analysis methods to make sure the results make sense. When NAEP reports are written, statistics, figures, web tools, and other materials are subjected to quality assurance.



UNITED STATES DEPARTMENT OF EDUCATION
INSTITUTE OF EDUCATION SCIENCES

June 2, 2003

MEMORANDUM

To: Richard J. Dowd
Regional Inspector General for Audit
Chicago, IL

From: *for* Grover J. Whitehurst *G. Whitehurst*
Director, Institute of Education Sciences

Subject: Response to Draft Audit Report
Review of Management Controls Over Scoring of the National Assessment of
Educational Progress (NAEP) 2000
Control Number ED-OIG/A05-C0010

Thank you for the opportunity to respond to your draft report. We are pleased that the study resulted in a determination that management controls over scoring of the NAEP 2000 assessment were adequate.

We concur with the findings and recommendations and have taken steps to ensure that contractor monitoring is improved and that the non-material weaknesses have been addressed. Attached is a response from the National Center for Education Statistics (NCES) that documents the changes and improvements that have been implemented to address OIG findings and recommendations.

Attachment

cc: Richard Rasa, Director, Advisory & Assistance for State & Local Programs, OIG
Valerie Plisko, Director, NCES

NCES Response to OIG Draft Audit Report ED-OIG/A05-C0010

The draft OIG Audit Report, *Review of Management Controls Over Scoring of the National Assessment of Educational Progress 2000*, is a thorough review of the adequacy of the quality of management controls over scoring of the 2000 National Assessment of Educational Progress. The audit comments on two nonmaterial weaknesses in the monitoring of the management controls over scoring of the 2000 NAEP in Mathematics and Science. Both are being addressed through improvements NCES has made to directing and monitoring contractor work and are described in this memo.

1) The first weakness is that 'NCS [now NCS Pearson, the NAEP contractor for scoring] did not use and/or document mathematics qualifying sets for training on extended constructed response questions as required in the NAEP 2000 Technical Application.' To remedy this weakness for the 2003 NAEP, NCES has ensured that NCS Pearson and Educational Testing Service (ETS), the NAEP contractor for design, analysis, and reporting, are working together to reorganize the training sets for all the extended constructed response items. These activities are being fully documented in contractor monthly reports to NCES. NCS Pearson is keeping complete records regarding qualifying sets of items for scoring.

In more detail, between 10 and 20 practice papers were pulled from the Practice Set and placed in Qualification Sets of 10. If there were less than 10 papers remaining in the Practice Set, the Practice Set was supplemented with responses from 2003. The new Training Sets for extended constructed response items were implemented for the scoring of the 2003 NAEP mathematics assessment. Training sets for new extended constructed response items are to contain:

- Anchor Papers Approximately 10 papers that definitively show the score points. These papers have scores printed on them.
- Practice Papers Usually 2 sets of 10 papers each that show more of the 'gray' areas. There are no scores printed on these papers. Scorers have an opportunity to practice scoring and also ask more questions to flesh out their understanding of the rubric.
- Qualification Papers Usually 2 sets of 10 papers each. There are no scores printed on these papers. The scorers must attain an 80% correct score to begin scoring the item.

All of these papers are part of the training set; for each paper, the trainer explains to the scorers why a response was given a specific score. As recommended in further quality control studies, some of the training sets will be further expanded.

2) The audit report also disclosed that some scorers did not meet scorer qualification requirements in 2000. Under the current scoring contract, NCS Pearson is allowed to substitute scoring experience for some academic qualifications. This is due to the difficulty of hiring enough scorers with the previously required academic credentials, and

the determination by NCES that prior successful experience in scoring and effective training are as critical a prerequisite for consistent, high-quality scoring as possession of an advanced degree or even classroom teaching experience in a specific content area. NCES is monitoring the qualifications of scorers through the contract process. In addition, NCES has added a new level of external evaluation of NAEP scoring quality to the program through the award of a separate Quality Assurance contract to the Human Resources Research Organization (HumRRO).

We appreciate the opportunity to respond to the findings of the auditors and document the changes and improvements that have been implemented to address these findings.