THE NATIONAL CENTER FOR EDUCATION STATISTICS

THE NATION'S REPORT CARD

FOCUS on NAEP

Vol. 4 No. 2 November 2000

How Does NAEP Ensure Consistency in Scoring?

Abstract: *Each NAEP assessment requires the scoring of thousands, and often millions, of written responses to open-ended questions. NCES and its contractors have devised a variety of techniques to ensure that these heterogeneous responses are scored consistently.*

The National Center for Educational Statistics (NCES) has been conducting the National Assessment of Educational Progress (NAEP) since 1969. In addition to regular assessments in reading, mathematics, science, and writing, NCES also conducts assessments in such subjects as geography, U.S. history, civics, and the arts.

All of these assessments include constructed-response questions in addition to multiple-choice items. Many include "short constructed-response" questions, which require students to provide a numerical response or write a few words or sentences, as well as "extended constructed-response" questions, which may require the student to write a paragraph or more, perform a science experiment and write a description of what was done, or solve a word problem in mathematics, providing a written explanation of the answer. Writing assessments require students to produce two extensive writing samples, while the arts assessments require students to create and perform art.

Extended constructed-response questions for NAEP assessments such as reading, U.S. history, geography, and civics are scored according to four-level scoring guides. Four-point answers are typically scored as incorrect, partial, essential, and fully correct. However, some assessments, such as the arts, mathematics and writing assessments, have questions that recognize five or even six levels of performance.

Each national assessment generates thousands of student responses that must be scored individually, and combined state/national assessments can generate almost five million responses.[1] NCES and its contractors have developed a large number of special techniques to ensure that these constructed-response questions can be scored consistently. This *Focus on NAEP* will discuss the techniques used to score written assessments such as reading, mathematics, writing, and science. A separate *Focus on NAEP* will cover the special problems encountered in assessing the arts.

Selecting Scorers

In the year 2000, NCES will conduct two national/state assessments, in mathematics and science, at grades 4, 8, and 12 at the national level and at grades 4 and 8 at the state level. In addition, there will be a national reading assessment for grade 4 only. The three assessments will generate close to 10 million constructed responses. The scoring will be done, as it has been done for previous assessments, by National Computer Systems (NCS). Educational Testing Service (ETS) develops the scoring guides for the questions and provides training in their use.

Scoring will be done at two on-line Professional Scoring Centers, one in Iowa City and the other in Tucson, Arizona. The contractors will hire about 150 scorers for the mathematics assessment, about 175 for the science, and about 50 for the reading.

Scorers selected for the assessment will have the following qualifications:

- A minimum of a bachelor's degree in the appropriate academic discipline (mathematics, science, or English), or in education;

- Scoring experience in NAEP or non-NAEP assessments preferred;
- Teaching experience at the elementary or secondary level preferred.

The 2000 Mathematics Assessment will have bilingual booklets for the 4th and 8th grades. Scorers fluent in Spanish will be hired for the scoring of booklets answered in that language.

Training Scorers

Training scorers to score short and extended constructed-response questions consistently is one of the most important parts of the entire scoring procedure. There is separate training for each constructed-response question.[2]

Training involves the following:

- Presenting and discussing the question to be scored and the question's rationale;
- Explaining the scoring guide to the team and discussing the "Anchor Packet," which contains the scoring guide, the question, its scoring rationale, and the "Anchor Set" of student responses that represent the various score points in the guide;
- Discussing the rationale behind the guide, focusing on the criteria that differentiate the levels in the guide;
- Practicing scoring on a "Practice Set" of students' answers;
- Continuing to practice until a consensus is reached on how to apply the scoring guide.

Trainers and participating experts in the field begin by selecting from 150 to 300 student answers to an extended constructed-response question. They score them all, for training purposes, and use the answers to create three different training sets, the Anchor Set, the Practice Set, and the Qualification Set.

Answers in the Anchor Set have the scores written on them. An Anchor Set contains at least three answers for every score point in a question. The Anchor Set for a three-point question will usually have 10 answers, and the Anchor Set for a four-point question will have about 15. The trainers also score a Practice Set of about 10 to 20 answers, and a Qualification Set of similar size, but do not put the scores on the answers.

Scorers, divided into training teams, will first study the scoring guide developed for a given question. Then they receive the Anchor Set of answers, which they review in conjunction with the scoring guide. Then they are given the Practice Set. Scorers score each of the answers, and then are given the "true" score, arrived at earlier by the trainers, for comparison and discussion.

Once the scorers are familiar with the scoring of a question, they are given a Qualification Set of answers to score. At least 80 percent of their scores must match the scores given by the trainers. Scorers who fail to get 80 percent discuss the scoring of the Qualification Set with their trainer and then are given a second Qualification Set. If they fail to get at least an 80 percent match on this set, they cannot score the question.

Image Scoring and Monitoring

Scoring of constructed-response questions is done by an "Image" process. While student answers are written in traditional answer booklets, for scoring purposes they are converted into computer images. This allows all the answers for a given question to be grouped together and scored at the same time. Scorers are trained to score the answers to a question, and then work exclusively on answers to that question until each one has been scored.

When scorers begin scoring answers to a question, they first take turns scoring the same question, comparing answers, or score in pairs as a final quality check before scoring on their own. They receive retraining at the beginning of each day and after any break that exceeds 15 minutes.

Scorers will be monitored by supervisors (known as "table leaders") in a variety of ways. A certain percentage of answers for constructed-response questions will be scored twice.[3] The second scorer will not know the score assigned by the first scorer. Because all scoring is done on a linked computer network, table leaders will have data on the scoring agreement rates for all scorers while the scoring is in progress. Figure 1 provides a "reliability summary" used to keep track of scoring consistency.

A minimum standard agreement rate will be set for each question, which will take into account both the number of score points for a question and the subject being assessed. For example, a higher agreement rate is set for a three-point question than a four-point question; and agreement rates will be higher for a subject such as mathematics, where the "correct" answer can usually be defined with greater precision, than for a subject such as reading. In 1998, the average standard agreement rate for questions on the reading assessment was 91 percent for grade 4, 90 percent for grade 8, and 89 percent for grade 12. For the 1996 mathematics assessment, it was 96 percent for all three grades.

If the minimum agreement rate is not met for a question, a number of different remedial actions may be necessary. If all or most members of a scoring team appear to be below the average, retraining may be appropriate. If there seems to be a problem with one scorer, the scorer may be reassigned.

The answers that were scored with insufficient agreement rates need to be rescored. This may be done by a group of supervisors, or all the scores for a question may be erased, and the team starts over again. Sometimes, the question is assigned to a different scoring team.

Occasionally, the scoring trainer may decide that the scoring guide needs to be refined, although this rarely happens during an assessment. Scoring guides are more likely to be refined during preliminary testing of assessment questions.

Table leaders will have methods to review an individual scorer's consistency as well as the consistency of a scoring team. A table leader will typically review 10 percent of the answers scored by a scorer, and will discuss with the scorer any score that appears inappropriate. A table leader has the authority to rescore any answer, although this does not affect the inter-rater reliability data. To check on scoring consistency across individual scorers, a table leader can also review all the answers that were given a particular score by a scoring team or the committee that developed the assessment questions.

The NAEP assessments that NCES will be conducting in 2000 are periodically redesigned to keep them responsive to changes in curricula and also to reflect improvements in assessment techniques. However, because NCES uses the same assessment instrument several times before making changes, these assessments usually offer some

Figure 1.—Reader Reliability Summary

First Scorer	**Blank**		**1**		**2**		**3**		**4**		**Illegible**		**Off Task**	
Second Scorer	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Blank	115	100%												
1			18	90%	2	10%								
2			2	1%	330	95%	17	5%						
3					**7**	**4%**	**156**	**92%**	**6**	**4%**				
4							3	13%	21	88%				
Illegible														
Off Task													1	100%

Total Times 2nd Read: 678

Percent Agreement: 94.5%

This sample "Reader Reliability Summary" shows how table leaders at National Computer Systems keep track of the scoring consistency of the second scoring of a single NAEP extended constructed-response question.

The sample summary is for a four-point question, whose answers are scored as either "incorrect," "partial," "essential," or "fully correct"—"fully correct" answers receiving the full four points. (The rows and columns marked "Blank," "Illegible," and "Off Task" are for answers that are unscorable due to omission, completely illegible handwriting, and unresponsiveness to task.)

This summary shows the cumulative agreement rate for all second scoring of students' answers to a single four-point question. Scoring decisions by the first scorer head the double columns at the top of the chart, while those for the second scorer, appearing in the far left-hand column, govern the rows. The chart should be read row by row. (The "3" row has been bolded for illustration.)

The cells created by the intersection of the "3" row and the double columns labeled "2", "3," and "4"give information on answers that received a "3" score from the second scorer. The first "n" or "number" cell shows that 7 answers scored as "3" by the second scorer received a score of "2" from the first scorer. The first "%" cell indicates that these 7 answers constitute 4% of the answers scored as "3" by the second scorer.

The next two cells to the right indicate that 156 answers, or 92% of all the answers receiving a "3" score from the second scorer, received a "3" from the first scorer as well. The next two cells indicate that 6 answers (4%) received a "3" from the second scorer and a "4" from the first scorer.

Ideally, all numbers and percentages would be in the shaded cells, and all percentages would be 100%. In fact, however, this only occurs for the "Blank" and "Off Task" answers. The "Percent Agreement" of 94.5% seen in the lower right-hand corner is obtained by dividing the total number of "agreed" scores (641) by the total number of scores (678).

trend data. For this reason, decisions by scorers working on the current assessments will be compared with decisions by past scorers when appropriate. A similar procedure is used for the NAEP long-term trend assessments, whose primary function is to track student performance over time.

Conclusion

Achieving consistency in the scoring of constructed-response questions begins with the selection of individuals who have a background in education and experience in scoring. These individuals are trained carefully in the scoring of each question, so that all the scorers, working independently, give the same number of points to any answer to that question. Regular second scoring of answers to every question ensures that this consistency is maintained throughout the scoring process.

Endnotes

[1] The NAEP 1997 arts assessment (in music, theatre, and the visual arts) covered the 8th grade only, and involved a total of about 6,500 students. The arts assessment involved relatively few questions, because students devoted much of their time to a single creating or performing task. A national/state assessment in a subject such as science will involve about 7,500 students at each of three grades (4th, 8th, and 12th), plus about 2,500 per state per grade. In the past, more than 40 states and other jurisdictions have participated in each NAEP state assessment.

[2] The training procedures described are for extended constructed-response questions. The procedures for short constructed-response questions are similar but less elaborate.

[3] Six percent of the answers for the constructed-response questions of the mathematics and science assessments for grades 4 and 8 will be scored twice. This will include both the national and state assessments for these subjects and grades. In addition, 25 percent of the answers for the grade 12 assessments in science and mathematics will be scored twice, a procedure that will also be followed for the reading assessment (grade 4 only). A larger percentage will be scored for these assessments because they are national assessments only, and thus will involve substantially fewer answers.

For Further Information

The NAEP 1996 Technical Report, NCES 1999–452, discusses all technical aspects of the 1996 Mathematics and Science Assessments and the 1996 Long-Term Trend Assessments.

Technical Report: NAEP 1996 State Assessment Program in Science, NCES 1998–480, covers the technical aspects of that assessment in detail.

Single copies of both reports are available free from ED Pubs, P.O. Box 1398, Jessup, Md. 20794–1398. Copies may also be downloaded from the World Wide Web ***http://nces.ed.gov/pubsearch/***

The *Focus on NAEP* series briefly summarizes information about the ongoing development and implementation of the National Assessment of Educational Progress (NAEP). The series is a product of the National Center for Education Statistics (NCES), Gary Phillips, Acting Commissioner, and Peggy Carr, Associate Commissioner for Education Assessment. This *Focus on NAEP* issue was written by **Sheida White** of NCES, **Connie Smith** of National Computer Systems, and **Alan Vanneman** of the Education Statistics Services Institute.

To order other NAEP publications, call toll free 1–877–4ED–Pubs (1–877–433–7827), TTY/TDD 1–877–576–7734;
E–mail: edpubs@inet.ed.gov;
Internet: ***http://www.ed.gov/pubs/edpubs.html***

The NCES World Wide Web Home Page is:
http://nces.ed.gov/