Handbook of Research and Policy in Art Education

Edited by

Elliot W. Eisner

Michael D. Day

A Project of the National Art Education Association
When designing a large-scale assessment, it is important to view the assessment as a dynamic, evolving system—not as a static system that is going to look the same in 20 years as it did when it was first implemented. The construct that is being measured may not change, but our knowledge of how best to carry out the measurement of that construct may. Those who design an assessment and those who use the assessment results should expect that the assessment will change over time and that modifications will be made. An assessment system needs to have the capacity to respond to sound research and evaluation studies that inform assessment practice. Advances in technology and in psychometrics, as well as advances in our practical knowledge of how to carry out the critical tasks of assessment design, administration, scoring, and reporting, should necessarily lead to improvements in the assessment over time.

As with many things in life, implementing change has its up side and its down side. When changes are made in an assessment with the aim of improving it, establishing score comparability for different forms of the assessment can become a thorny challenge. Suppose that students take an assessment, and then substantial changes are made in the assessment specifications the following year. A different blueprint is used to construct a new form of the assessment. Can we assume that students’ scores on the two forms will be comparable? Will the inferences we make about a student’s performance (and the decisions we might make about the student based on those inferences) be the same, regardless of which form of the assessment the student took? Are the alternate forms of the assessment truly interchangeable? Just how much change can we make in an assessment before lack of comparability of forms becomes an issue?

The Standards for Educational and Psychological Testing, a joint publication of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), provide a set of criteria for evaluating assessments, assessment practices, and the effects of assessment use. Several standards speak to the issue of establishing the comparability of assessments that evolve over time, acknowledging the difficulties encountered when the assessments “measure different constructs, . . . differ materially in reliability, time limits, or other conditions of administration, or . . . are designed to
different specifications” (pp. 51–52).1 Clearly, balancing the need for maintaining the comparability of assessment forms with the need to change an assessment to improve its measurement capabilities can be a tense struggle.

In this paper, we focus on two large-scale art assessments that have evolved over time and have borne witness to that struggle: the National Assessment of Educational Progress (NAEP) visual arts assessment and the Advanced Placement (AP) Studio Art portfolio assessment. Both have been in existence for over 25 years and have undergone significant changes while, at the same time, maintaining their integrity as psychometrically sound assessment programs. As we shall see, the two programs present an intriguing story of contrasts. They have different goals, and the assessment information they provide is used for very different purposes. Each has weathered its own set of “growing pains” over the years, and both have wrestled with concerns over the necessity of maintaining score comparability across different forms of the assessment.

In the next section of the paper, we describe what each assessment looks like, explain how each assessment has evolved since its inception, and discuss research and evaluation studies that have made use of these assessments. We then compare and contrast how each of the assessments has dealt with the issue of comparability. Finally, we close the paper with a discussion of the implications of these large-scale assessments for future assessments.

THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS VISUAL ARTS ASSESSMENTS

What Does the Assessment Presently Look Like?

The National Assessment of Educational Progress (NAEP) is an ongoing assessment of what America’s students know and can do in a number of different subject areas. Assessments have been carried out since 1969 to provide information about students’ knowledge and skills at ages 9, 13, and 17 (and in more recent years, for students in grades 4, 8, and 12). Policymakers, educators, and the general public use the assessment information to gauge the condition and progress of education at the state and national level. NAEP reports levels of student achievement and student- and school-reported background variables that are associated with those levels of achievement. For individual students, the NAEP assessment is a low-stakes assessment, because no individual scores are given. The major function of the assessment is to provide valid and reliable information about the knowledge, skills, and abilities of the nation’s students—not to report on the performance of individual students, schools, or districts.

Design of the Assessment. The design of the NAEP 1997 visual arts assessment was guided by an arts education assessment framework that laid out the general parameters of the assessment, and by an assessment and exercise specifications document that provided the detailed information needed for devising the assessment instruments (National Assessment Governing Board, 1994a, 1994b). Additionally, at the time that the assessment framework was being developed, the Consortium of National Arts Education Associations (1994) was beginning its work to define national standards for education in the arts, specifying what

---

1 Several standards are included that address the issue of establishing score comparability for alternate forms of an assessment (Standards 4.10 and 13.8). Other standards address comparability concerns that arise when the administration of an assessment has been modified to provide accommodations for individuals with disabilities (Standards 10.1, 10.4, 10.5, 10.7, and 10.11) or to provide linguistic modifications in an assessment adapted in a secondary language (Standards 9.5, 9.6, 9.7, and 9.9)
students should know and be able to do in dance, music, theater, and visual arts. The leadership of the two projects worked collaboratively toward establishing a common vision of the goals of arts education. The members of the visual arts group of the NAEP Planning Committee took the national achievement standards in the visual arts that had been developed by the Consortium project, identified those that were appropriate for a large-scale assessment, and then adapted those achievement standards so that they could be included in the content outlines that would be used to guide the development of the NAEP art assessment.

**Structure of the Assessment.** The NAEP 1997 visual arts assessment consisted of a series of seven “blocks,” or sets of multiple related exercises that were administered separately, each block to be completed within a particular time frame. A block included one or more stimuli and sets of multiple-choice exercises, short or extended constructed-response exercises, and/or creating exercises. The assessment was composed of three Creating blocks and four Responding blocks. All students who participated in the visual arts assessment took either one Responding block and one Creating block, or two Responding blocks.² (The NAEP assessments employ a matrix sampling approach to test design, which does not require that all students take the exact same form of the assessment. Rather, exercises are spiraled across forms so that more of the content domain can be covered in a single assessment.)

**Administration of the Assessment.** The visual arts assessment was administered to a nationally representative random sample of public and nonpublic students in grade 8 (\(N = 2,999\) students from 128 schools).³ As part of the assessment, students and school administrators completed questionnaires to provide demographic and background information that would be used when reporting assessment results.

How Has the Assessment Evolved Since Its Inception? What Are the Major Changes It Has Undergone?

The 1997 assessment was the third NAEP assessment in the visual arts to be carried out. Earlier assessments were conducted in 1975 and in 1978. As the assessment has evolved over the years, there have been significant changes in its design, administration, scoring, and reporting.

**Differences in Design.** The overarching design frameworks that have guided the development of the assessments have radically changed over the years. The assessment frameworks used in the 1970s defined the content domain of visual arts education in terms of a series of behavioral objectives (as was also the case in other disciplines that NAEP assessed during that time period). By contrast, in the more recent framework, the visual arts assessment was defined in terms of a design matrix that delineated the processes and content of arts education and a series of detailed content outlines that were based on that matrix.

Wilson (1970, 1971) explained the process of laying out the content domain for the 1975 NAEP visual arts assessment. A set of 5 fairly broad objectives that art educators agreed reflected major outcomes of art education were devised, and a set of specific behaviors supporting each of the broad objectives were prepared. There were 153 major and subobjectives that together defined the content domain for the assessment (Wilson, 1975) and provided the

---

²Even though some students took two Responding blocks in the visual arts assessment, all but one of the Responding blocks included some Creating exercises.

³The visual arts assessment was originally designed to be administered at grades 4, 8, and 12. However, due to budgetary constraints, the assessment could only be administered to one grade level.
### TABLE 28.1
The “Produce Works of Art” Portion of the Content Outline for the 1974 NAEP Visual Arts Assessment

#### III. Produce Works of Art

<table>
<thead>
<tr>
<th></th>
<th>Produce works of art with a particular composition, subject matter, expressive character, or expressive content.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age 9</td>
</tr>
<tr>
<td></td>
<td>1. Produce a work of art that fulfills the intrinsic demands of a space or shape.</td>
</tr>
<tr>
<td></td>
<td>2. Produce a work of art containing specified subject matter.</td>
</tr>
<tr>
<td></td>
<td>3. Produce a work of art with a particular mood, feeling, or expressive character.</td>
</tr>
<tr>
<td></td>
<td>Age 13 (in addition to Age 9)</td>
</tr>
<tr>
<td></td>
<td>1. Produce a work of art with a particular mood, feeling, or expressive character.</td>
</tr>
<tr>
<td></td>
<td>a. Produce a work that fits the mood of a poem or piece of music.</td>
</tr>
<tr>
<td></td>
<td>b. Produce a work that shows a mood such as calmness, excitement, gaiety, or sadness.</td>
</tr>
<tr>
<td></td>
<td>c. Produce a work (landscape, city, or town) that has a particular feeling such as coolness, loneliness,</td>
</tr>
<tr>
<td></td>
<td>warmth, wetness, or spookiness.</td>
</tr>
<tr>
<td></td>
<td>2. Produce a work of art with meaning based on the use of established symbols.</td>
</tr>
<tr>
<td></td>
<td>3. Produce a work of art with meaning based on the use of new symbols.</td>
</tr>
<tr>
<td></td>
<td>4. Design a poster that advertises an event, product, etc.</td>
</tr>
<tr>
<td></td>
<td>5. Produce a work that has a particular type of order or variety.</td>
</tr>
<tr>
<td></td>
<td>6. Modify the form of an object to improve its aesthetic quality or functional character.</td>
</tr>
<tr>
<td></td>
<td>Age 17, A (in addition to Age 13)</td>
</tr>
<tr>
<td></td>
<td>Produce a work of art that has a particular composition such as vertical, horizontal, diagonal, concentric,</td>
</tr>
<tr>
<td></td>
<td>symmetrical, and asymmetrical; that uses deep or shallow space; or that has an open or closed composition.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Produce works of art that contain various visual conceptions.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age 9</td>
</tr>
<tr>
<td></td>
<td>1. Demonstrate the ability to represent spatial conceptions (one person standing in front of another,</td>
</tr>
<tr>
<td></td>
<td>something close and something far, a street and a building, etc.).</td>
</tr>
<tr>
<td></td>
<td>2. Demonstrate the ability to represent accurately (depict the essential attitude and position of a model</td>
</tr>
<tr>
<td></td>
<td>and indicate such things as clothing patterns).</td>
</tr>
<tr>
<td></td>
<td>3. Produce an accurate reportage drawing.</td>
</tr>
<tr>
<td></td>
<td>4. Produce works in which the subject matter aspects indicate expressions and emotions (running, walking,</td>
</tr>
<tr>
<td></td>
<td>falling, laughing, crying, anger, fright, happiness, etc.).</td>
</tr>
<tr>
<td></td>
<td>Ages 13, 17, A (in addition to Age 9)</td>
</tr>
<tr>
<td></td>
<td>Demonstrate the ability to represent an object from different viewpoints under different light conditions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Demonstrate knowledge and application of media, tools, techniques, and forming processes.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age 9 (None)</td>
</tr>
<tr>
<td></td>
<td>Ages 13, 17, A</td>
</tr>
<tr>
<td></td>
<td>1. Perform processes such as coiling a pot, cutting and printing a linoleum block, mixing specific colors,</td>
</tr>
<tr>
<td></td>
<td>etc.</td>
</tr>
<tr>
<td></td>
<td>2. Select the appropriate tools to accomplish certain tasks such as printmaking, clay modeling, etc.</td>
</tr>
</tbody>
</table>

**Note:** This portion of the content outline was extracted from the full set of objectives found in chapter 3 of the 1971 National Assessment of Educational Progress publication, *Art Objectives*, edited by Eleanor L. Norris and Barbara Goodwin.
basis for exercise development. The 5 major objectives for the 1975 visual arts assessment were: (1) perceive and respond to aspects of art, (2) value art as an important realm of human experience,4 (3) produce works of art, (4) know about art, and (5) make and justify judgments about the aesthetic merit and quality of works of art. Table 28.1 shows the content outline that was developed for the third major objective, “produce works of art” for the 1975 NAEP visual arts assessments for ages 9, 13, and 17.

The process used to delineate the content domain for the 1978 NAEP visual arts assessment was very similar (National Assessment of Educational Progress, 1981b). A sample of arts educators reviewed the objectives that had been prepared for the 1975 assessment to determine their appropriateness for the 1978 assessment. None of the objectives were changed, but changes were made in the recommended amount of assessment time to be allotted to certain objectives.

The assessment framework developed for the 1997 NAEP art assessment depicted visual arts education in terms of a design matrix of processes and content. Along the vertical axis of the matrix are two central processes that exemplary teaching in the visual arts seeks to foster: (1) creating works of art, defined as “expressing ideas and feelings in the form of an original work of art”; and (2) responding to existing works of art, defined as “observing, describing, analyzing, and evaluating works of art” (Persky, Sandene, & Askew, 1998, p. i). Along the horizontal axis of the matrix are two components of learning (knowledge and skills) that comprise the content of visual arts. Students studying in the visual arts work toward attaining “knowledge and understanding about the arts, including the personal, social, cultural, and historical contexts for works”; as well as “perceptual, technical, expressive, and intellectual/reflective skills” (National Assessment Governing Board, 1994a, p. 8). The matrix is shown in Figure 28.1.

The purpose of the 1997 assessment, then, was to measure students’ knowledge and skills in creating and responding to works of art. The NAEP Visual Arts Planning Committee developed a series of content outlines to guide the design of exercises for the assessment, using the design matrix as the scaffolding for preparing their outlines. Table 28.2 shows the grade 8 visual arts content outline for the “creating” portion of the content domain.

A comparison of Tables 28.1 and 28.2 makes readily evident key differences in the two approaches to defining the production portion of the visual arts content domain. Clearly, the two committees charged with designing these assessments had differing conceptions of how this particular portion of the content domain should be defined. Although there is some obvious overlap in the two content outlines, there are important aspects unique to each. Similar conceptual differences are also evident when one compares the 1975 and 1997 content outlines that lay out the “responding” portion of the visual arts content domain. (For the interested reader, the assessment framework can be downloaded at the following Web site: http://www.nagb.org/pubs/artsed.pdf.)

The Nature of Innovation in Assessment Design Changes Over Time. All three NAEP arts assessments could be characterized as breaking ground for their time, pushing the envelope of assessment methodology in new directions, because there were few models for large-scale arts assessment upon which to draw when the NAEP assessments were developed. The arts assessments administered in the 1970s were certainly on the cutting edge of assessment design:

---
4A unique aspect of the assessments conducted in the 1970s was their inclusion of a number of exercises (i.e., about 20% of the assessment) to evaluate students’ attitudes toward art to cover this second art objective. Some critics of these assessments (Gardner & Grunbaum, 1986) have voiced concern about allocating valuable assessment time to ask questions about students’ attitudes toward art. The 1997 visual arts assessment did not tap attitudes, only knowledge and skills. Attitudes were not considered to be part of the content domain to be assessed.
The item writers of both the art and music assessments were told to be innovative in developing test procedures and to ignore constraints of cost, time, administration, and scoring. Most important, they were not asked to restrict tasks to those that could be tested in multiple-choice formats scorable by machine. Although many innovative testing procedures were developed, most were set aside before the art and music assessments were undertaken because they were considered too difficult to administer and too expensive and time-consuming to analyze. Nonetheless, the National Assessment did result in an examination that went beyond a machine-scored format. (National Endowment for the Arts, 1988, p. 93)

Although all three assessments have included a blend of multiple-choice, constructed-response, and production exercises, the relative numbers of each exercise type (and amount of assessment time allocated to each) have changed over time. The authors of the Design and Drawing Skills report from the first NAEP art assessment (National Assessment of Educational Progress, 1978d) acknowledged an important design challenge faced by those who want to include art production exercises in large-scale assessments: obtaining adequate content coverage. Assessment time is quite limited, and the content domain of possible art production exercises is large. Given that such exercises are time consuming to administer and score, does it make sense to include any in large-scale assessments? The assessment developers struggled with this dilemma:

In developing the art objectives, National Assessment was aware of the fact that resources would not permit a comprehensive survey of student abilities to produce art. Nevertheless, NAEP considered it important to acknowledge as full a range of production goals as possible in the objectives. Hopefully, some of the goals that have not been measured will be measured in future assessments. Their exclusion from the first assessment does not suggest that they cannot be assessed. (p. 4)

The design strategies adopted for the 1975 and 1978 assessments placed a premium on assessing breadth of coverage within the content domain. To help accomplish that goal, the
TABLE 28.2
The “Creating” Portion of the Grade 8 Visual Arts Content Outline for the 1997 NAEP
Visual Arts Assessment

Visual Arts, Grade 8

I. Creating

A. Generate subjects, themes, problems, and ideas for works of art and design in ways that reflect knowledge and understanding of values (personal, social, cultural, historical), aesthetics, and context.
   1. Speculate and discriminate among various ideas, making the most appropriate choices for specific artistic or design purposes.
   2. Interpret and speculate on the ways that others have used subject matter, symbols, and ideas in visual, spatial, or temporal expressions, and how these are used to produce meaning or function that is appropriate in their own works.
   3. Analyze the characteristics of art and design works of various eras and cultures to discover possible expressions or solutions to problems.
   4. Speculate on how factors of time and place (such as climate, resources, ideas, and technology) influence the visual, spatial, or temporal characteristics that give meaning or function to a work of art or design.

B. Invent and use ways of generating visual, spatial, and temporal concepts in planning works of art and design.
   1. Demonstrate the development of ideas across time.
   2. Analyze and consider form, media, techniques, and process, and analyze what makes them effective or ineffective in communicating specific ideas.
   3. Demonstrate knowledge of how sensory qualities, expressive features, and the functions of the visual arts evoke intended responses and uses for works of art and design.
   4. Speculate about the effects of visual structures (elements and principles of design) and reflect on their influence on students’ ideas.
   5. Evaluate and discriminate among various ideas, making the most effective choices for specific artistic purposes or design uses.

C. Select and use form, media, techniques, and processes to achieve goodness of fit with the intended meaning or function of works of art and design.
   1. Experiment, select, and employ form, media, techniques, and processes and analyze what makes them effective or ineffective in communicating ideas.
   2. Utilize knowledge of characteristics of materials and visual, spatial, and temporal structures to solve specific visual arts and design problems.
   3. Interpret the way that others have used form, media, techniques, and processes and speculate how these produce meaning or function.

D. Experiment with ideas (sketches, models, etc.) before final execution as a method of evaluation.
   1. Evaluate, discriminate, and articulate differences among various ideas and forms, making the most effective choices for specific artistic purposes or design uses.
   2. Simulate and articulate new insights and changes in direction that result from representation or simulation of ideas.
   3. Employ organizational structures and analyze what makes them effective or ineffective in the communication of ideas.

E. Create a product that reflects ongoing thoughts, actions, and new directions.
   1. Use media, techniques, and processes and analyze what makes them effective or ineffective in communicating ideas.
   2. Integrate visual, spatial, and temporal concepts with content to communicate intended meaning in their art works.
   3. Use subjects, themes, and symbols that demonstrate knowledge of contexts, values, and aesthetics that communicate intended meaning in art works.
   4. Evaluate ideas and artwork throughout the creating process, making the most effective choices for specific artistic purposes or design uses.

(Continued)
TABLE 28.2
(Continued)

Visual Arts, Grade 8

F. Reflect upon and evaluate their own works of art and design (i.e., students judge the relationship between process and product; the redefinition of current ideas or problems; and the definition of new ideas, problems, and personal directions).

1. Evaluate final compositions for use of compositional and expressive features.
2. Demonstrate knowledge of the various purposes and reasons for works of visual art and design based on people’s experiences (cultural backgrounds, human needs, etc.).
3. Propose how works in the visual arts and design affect the way people perceive their experiences.
4. Compare and evaluate the characteristics of works in two or more art forms that share similar subject matter, historical period, or cultural context.
5. Describe new insights that have emerged from process and products of art and design that are meaningful to daily life.

Note: This portion of the grade 8 visual arts content outline was extracted from the full content outline found in the 1994 National Assessment Governing Board’s publication, *1997 Arts Education Assessment Framework*.

assessments made relatively heavy use of multiple-choice items (i.e., at least half the exercises in each assessment were multiple choice), but the assessments still managed to include at least some production exercises. (The 1975 art assessment incorporated eight design and drawing exercises to evaluate students’ production skills and abilities, whereas the 1978 assessment included four production exercises.) As Gardner and Grunbaum (1986) noted in their critique of these two assessments, about three fourths of the items required written responses.

By contrast, in the 1997 assessment, the assessment and exercise specifications called for less emphasis on breadth of coverage and more emphasis on tapping depth of knowledge and skills within the domain (National Assessment Governing Board, 1994b). Consequently, many fewer multiple-choice items were used, and many more constructed-response exercises and production exercises were included (i.e., 12 multiple-choice items, 24 constructed-response exercises, and 10 Creating exercises). In the 1997 assessment, students spent 70% of their time working on Creating exercises and the remaining 30% engaged in Responding exercises (Persky et al., 1998).

Unlike the developers of the earlier NAEP arts assessments, the developers of the 1997 assessment created exercise “blocks.” All the exercises included in a block made use of the same stimulus material and were organized around a single theme, issue, or problem. The exercises were designed to assess both creating and responding processes, integrating the two processes within the same block. (In this volume, chap. 27, Persky describes how exercise blocks were developed, administered, and scored. She provides an illustrative example of an exercise block from the visual arts assessment and presents its design rationale.) The adoption of this “exercise block” design greatly facilitated the probing of the depth of students’ knowledge and skills within the content domain.

The 1997 visual arts assessment also extended the range of art production tasks beyond those included in the 1974 and 1978 assessments, resulting in a widening of the definition of the visual arts. The production tasks employed in the 1974 and 1978 assessments were

---

5As Gardner and Grunbaum (1986) noted in their critique of these two assessments, about three fourths of the items required written responses.
limited to those appropriate for the assessment of design and drawing skills (i.e., draw a design for a piece of jewelry, a series of sketches representing ideas for a work of art, four people seated at a table from a particular perspective, three people on a playground, a person running; design a bedroom wall, a package for a specified product). By contrast, the 1997 assessment included an assortment of both two- and three-dimensional production tasks (i.e., create a self-portrait, a collage, a design for a package to carry a fish on a bicycle, a design for a package to hold a sound, a chipboard and plasticine monument, a plasticine and wire kitchen sculpture, a plasticine sculpture showing the metamorphosis of a man to a fish). In effect, the 1997 assessment became a national testing ground for learning how to design, administer, score, and report on a much wider variety of production tasks than had previously been attempted in any other large-scale visual arts assessment.

**Differences in Administration.** In both the 1975 and 1978 art assessments, national stratified random samples of 9-, 13- and 17-year olds in and out of school took part in the 1975 assessment ($N = 27,500$ for 1975, and $N = 32,000$ for 1978). For the 1975 assessment, each student spent about 130 min in assessment-related activities, completing on average about 110 exercises. The 1978 assessment was a combined assessment that included art, writing, and music. Each student participating in the assessment responded to exercises in two of the subject areas. The assessment time allocated to art was significantly reduced in the 1978 assessment; each student spent about 45 min responding to art exercises and another 45 min responding to exercises in the other subject area. A much smaller nationally representative random sample of students that included only grade 8 students participated in the 1997 art assessment ($N = 2,999$). Because the 1997 assessment incorporated many more open-ended exercises than the previous NAEP art assessments, it was more time consuming and expensive to administer. Parallel assessments in music and theater that contained many open-ended exercises were also administered as part of this assessment. Therefore, it became imperative that smaller samples for each of the arts disciplines be used (and that students be assessed at only one grade level) in order for the assessment to be administratively feasible and stay within budget.\(^6\)

**Differences in Scoring.** In all three NAEP art assessments, the preparation of scoring guides to evaluate students’ responses to the open-ended exercises has been a daunting enterprise. In the early 1970s, few guidelines for writing open-ended exercises and devising scoring guides were available. In short, the technology of designing and conducting large-scale performance assessment in the arts was in its infancy. Though arts educators had developed methods for informally evaluating students’ works of art, a process for designing scoring guides that could be reliably applied by multiple judges in a cost-effective manner in a large-scale assessment context had not yet been developed. As Wilson (1975) noted, several of the scoring guides that were used in the 1975 assessment to score students art works were nearly 200 pages in length, reflecting the inherent complexity of the judgment task. Training scorers to use the scoring guides reliably was a time-consuming and costly task. Keeping track of all the relevant details in the scoring guides required much skill and concentration on the part of the scorer. The scoring became a major challenge for the assessment. Due to budget cutbacks, only half of the

\(^6\)Jo Ann Pottorff, a member of the National Assessment Governing Board, served as spokesperson for NAGB when the results from the 1997 NAEP Arts Assessment Report Card were released. In her remarks, she noted that the NAEP arts assessment was an expensive effort, costing about $10 million dollars over a five-year period to develop, administer, score, and report. That figure did not include the development of the assessment framework and the assessment and exercise specifications. Those costs ($1.2 million) were borne by the National Endowment for the Arts and the Getty Center for Education in the Arts.
exercises requiring human scoring that had been administered were actually scored (National Assessment of Educational Progress, 1978d). Furthermore, only four of the eight production exercises that had been administered were scored and reported. (The unscored exercises were held back to be re-administered in the 1978 assessment.)

The 1978 assessment was designed to include many fewer open-ended exercises than the 1975 assessment, resulting in fewer responses to be scored. (The only open-ended exercises included in 1978 were ones that had been previously administered as part of the 1975 assessment.) Scoring guides used to evaluate examinees’ responses to the six open-ended exercises administered in 1978 ranged from 15 pages in length for the two short constructed-response exercises that required written responses, to over 100 pages for each for the four production exercises that involved the evaluation of students’ original works of art (National Assessment of Educational Progress, 1981b). High scoring costs again plagued the assessment, and students’ responses to the six open-ended exercises could not be scored until 1980 when additional funding became available.

In the following 20 years, NAEP assessment staff gained considerable knowledge and experience in constructing and scoring complex performance assessments in disciplines outside the arts. NAEP assessment frameworks for these disciplines increasingly began to call for the inclusion of more open-ended exercises and fewer multiple-choice items in the assessments. Consequently, the development of a technology for designing, administering, and scoring large-scale performance assessments became a high priority for NAEP (and for statewide assessments as well). During this time period, NAEP assessment staff developed approaches to devising scoring guides that could be implemented in a cost-effective manner without sacrificing reliability or validity. The 1997 NAEP arts assessment directly benefited from these technological advances. (In this volume, chap. 27, see Persky for examples of scoring guides that were employed in the 1997 visual arts assessment.) Though there were many more open-ended exercises included in the 1997 art assessment than in the earlier art assessments, many fewer students took part in the 1997 assessment. Consequently, the scoring task was somewhat more manageable, because there were considerably fewer responses to score than in the previous two art assessments.

**Differences in Reporting.** The initial reporting plan for the 1975 assessment called for the results to be reported by the five major objectives. However, just prior to the administration of the assessment, funding for the assessment was reduced, and the assessment plan had to be rather drastically scaled back. As a result, some of the major objectives had an insufficient number of exercises as measures of those objectives to justify reporting results by objectives (National Assessment of Educational Progress, 1978b). The remaining exercises comprising the assessment were categorized into three topics (i.e., knowledge about art, attitudes toward art, design and drawing skills). For both the 1975 and 1978 assessments, summary results were reported by topic rather than by major objective.

A unique aspect of the report on the 1978 assessment was the inclusion of value judgments that appeared alongside selected results. Several art consultants who had been hired by the National Assessment reviewed the findings of the assessment and provided commentary, interpreting various results for the art education community and discussing the implications of the findings for the field. For example, the report includes the following commentary by Brent Wilson regarding students’ performance on drawing tasks included in the assessment:

Wilson also felt that the drawing skills results were too low. “Elementary school curriculum guides often contain drawing units relating to the depiction of actions and expressions,” he said. “But the assessment results for the ‘draw an angry person’ exercise show that students don’t put much
action into their figures and that they do not draw expressively. These are not trivial things. One of the primary reasons young people draw is to produce visual symbolic models of themselves and their worlds so that they can anticipate and test future results. If children can’t move the characters they create into action, if they can’t show emotions, or if they can’t draw expressively, then they are deprived of an extremely important way of developing and comprehending reality.” (National Assessment of Educational Progress, 1981a, p. 15)


True to form, the 1997 NAEP arts assessment pushed the envelope of assessment reporting, just as it had pushed the envelope of assessment design, administration, and scoring. In the past, NAEP had produced report cards describing the results of the assessments conducted in the various disciplines. These reports were intended to be used to inform the decision making of policymakers, educators, and the general public.

When plans were being made for the 1997 arts assessment, a new approach was envisioned for reporting results and disseminating information about the assessment. Unlike previous NAEP arts assessments, that new approach was to involve heavy use of the internet as a dissemination tool. A NAEP arts assessment Web site was created (http://nces.ed.gov/nationsreportcard/arts/), and many of the assessment documents were posted there to make it easy for persons to access and download information about the design, administration, scoring, and reporting of the assessment.

The 1997 arts assessment report card had quite a different look and feel than did previous NAEP assessment reports in the arts (and in other disciplines, as well). For the first time, NAEP produced a full color report card (Persky et al., 1998) and an accompanying CD-ROM (National Center for Education Statistics, 1998b) to be used in conjunction with the report card. The CD-ROM included sample questions and student responses from the various exercise blocks, as well as some of the actual sights and sounds captured in the assessment. (The report card can be downloaded at the NAEP arts assessment Web site. The Web site also contains information for ordering a copy of the CD-ROM.) Additionally, the National Center for Education Statistics (1998a) published a full color brief report, “Arts Education: Highlights of the NAEP 1997 Arts Assessment Report Card,” that provided a thumbnail sketch of the arts assessment and presented selected sample results in a brochure format intended to be accessible to a wide lay audience. (This document can also be downloaded from the Web site.) A series of short Focus on NAEP publications were prepared to provide brief nontechnical descriptions of the design of the arts assessment (Vanneman & Goodwin, 1998; Vanneman, Morton., & Allen, 1998; Vanneman, Shuler, & Sandene, 1998; White & Vanneman, 1998a, 1998b; White et al., 1998). Finally, a series of short NAEP Facts publications were prepared to present selected findings from the arts assessment (White & Vanneman, 1998a, 1998b, 1998c). (These reports can all be downloaded from the following Web site: http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031#020.)

The approach taken to reporting results of the 1997 NAEP visual arts assessment was also quite revolutionary compared to the approach used in the 1970s. In 1997, results were presented in terms of the two key processes identified in the design matrix, Creating and Responding, rather than in terms of “topics” (as has been done in the 1970s). The results for Creating and Responding in the visual arts were presented separately. The report card included overall
Upper Level
(294) Support plausible interpretation of Kendall mother/child portrait with observations about pose, background, contrast, and color.
(281) Label and specifically describe three or more features of Bearden collage that the artist wants you to notice.
(273) Identify a technical similarity between Schiele and Kollwitz self-portraits.
(237) Support plausible interpretation of Schiele self-portrait with observations about style, line, and color.
(234) Describe three ways one mother/child portrait is less realistic than another.
(232) Explain materials and design of own package idea for transporting fish on bicycle.
(222) Describe three aspects of own self-portrait that communicate something about you.
(218) Explain how own package for transporting fish on bicycle attaches to bike.
(209) Explain how Bearden creates contrast between interior and exterior areas in collage.
(203) Identify genre of Bearden collage.
(202) Analyze and describe narrative conveyed by Bearden collage.
(201) Identify an example of 20th century Western art.
(199) Identify an example of Renaissance art.
(195) Explain three ways an artist made some things look near and some far in mother/child portrait.
(189) Support plausible interpretation of Schiele self-portrait with observations about style, line, or color.
(189) Identify a stylistic emphasis in Bearden collage.
(181) Describe two aspects of own self-portrait that communicate something about you.
(179) Describe two aspects of own self-portrait that communicate something about you.

Middle Level
(173) Identify purpose of insulating packaging material.
(169) Infer from photograph advantages of shrink-wrap packaging.
(162) Describe one aspect of own self-portrait that communicates something about you.
(157) Identify a difference between interior and exterior areas in collage.
(152) Give general idea of how own package idea for transporting fish on bicycle attaches to bike.
(145) Analyze subject of five mother/child portraits of different genres.
(144) Describe interesting aspects of three of four elements (colors, lines, shapes, and composition) in a mother/child portrait.
(141) Describe one characteristic of charcoal in Kollwitz self-portrait.
(140) Describe one or two ways one mother/child portrait is less realistic than another.
(138) Describe three specific and accurate first impressions of Bearden collage.

Lower Level
(129) Explain one or two ways an artist made some things look near and some far in mother/child portrait.
(120) Describe aspect(s) of Bearden collage.
(107) Describe what Schiele self-portrait is communicating or describe an aspect of style, line, or color.
(104) Describe aspect(s) of Kendall mother/child portrait.
(102) Infer from photograph one advantage of mailing a poster in a mailing tube.
(50) Describe interesting aspects of one or two of four elements (colors, lines, shapes, and composition) in mother/child portrait.

NOTE: Italic type indicates a multiple-choice question. Regular type indicates a constructed-response question.

FIG. 28.2. Map of selected questions on the NAEP visual arts responding scale: Grade 8.
summaries of Creating and Responding results and displayed the relationship of those results to student- and school-reported variables. Reporting results by artistic process was not in and of itself particularly revolutionary, however.

What was revolutionary was NAEP’s exploration of the use of item response theory (IRT) to scale data from an art assessment. (Scaling methods had been routinely used in NAEP assessments in other disciplines for some time, but not in the arts assessments.) The 1997 visual arts assessment included a sufficient number of Responding exercises to make it possible to scale that portion of the assessment. The Responding results were displayed on a 0 to 300 scale.

The “item map” that is produced when scaling is implemented is particularly useful as a means of visually representing what it means for students to be more (or less) proficient in terms of their performance on a set of exercises. The item map for the NAEP visual arts assessment orders the Responding exercises along a continuum, from those that were easier for students to perform well on (at the bottom of the map) to those that were harder for students to perform well on (at the top of the map). Figure 28.2 displays the item map. By reviewing the item map, one can readily identify the types of exercises that students at any given level along the scale had a high probability of having completed successfully, as well as those that they had a lower probability of having completed successfully.

Such reporting has potentially important instructional implications, because an art educator can review the item map and gain an understanding of how the Responding portion of the visual arts content domain is organized—In short, the art educator can see what it means to “get better” at Responding. If art educators would like to design instruction to help students become more proficient in responding to visual arts stimuli, the item map helps them determine what types of exercises to focus on in the initial phases of instruction (i.e., are the easiest for students to master). As students become more adept at responding, the item map helps art educators identify the types of exercises that could then be introduced that would likely be more challenging for students (i.e., those further up the scale that are more “difficult”). With this information in hand, art educators are in a far better position to be able to plan instruction that will foster deep learning of this critical aspect of the content domain. The inclusion of item maps in the NAEP Arts Assessment Report Card represents an important evolutionary step, because it moves us beyond our reliance on basic descriptive statistics for reporting results.

Have There Been Research/Evaluation Studies of This Assessment and Reports of Its Findings?

Following the publication of the reports from the two NAEP arts assessments administered in the 1970s, surprisingly only a few publications appeared that discussed the assessments. Chapman (1982), Wilson (1986), and Ward (1982) presented selected findings from the assessments and drew implications from those results for art education, whereas Gardner and Grunbaum (1986) critiqued the assessments and offered recommendations for future NAEP arts assessments. Several other publications included brief discussions of the assessments, their findings, and their implications for art teachers and administrators (National Endowment for the Arts, 1988; Wolf & Wolf, 1984; Zimmerman, 1984). However, despite all the time and effort spent in reporting on the assessments, the impact on the field was negligible. As Wilson (1986) lamented:

Because the Creating exercises were much more involved and time consuming for students to complete, there were many few Creating exercises included in the assessment. Each student participating in the visual arts assessment took only one Creating exercise. Consequently, there were too few Creating exercises (and too few students taking a set of Creating exercises) for those exercises to be scaled. Rather, Creating results are reported in terms of average percent of the maximum possible score.
There has been virtually no response to the Assessment findings. In fact, within the field of art education, the results of the Art Assessment have been treated with indifference. The findings from the Art Assessment are unknown to most art teachers and supervisors. The findings have seldom been discussed in the art educational literature. (p. 7)

The first two NAEP arts assessments spawned few published research studies. One exception is Knight (1979), who analyzed data from the 1974 NAEP art assessment, looking for patterns in male–female performance on the assessment. She compared mean differences in performance between males and females on the NAEP art assessment to mean differences in performance between males and females on NAEP assessments in other content areas (i.e., music, literature, science, citizenship, social studies, and math). She found that patterns of male–female performance in art were different from those in the other content areas.

Following the release of the 1997 NAEP Arts Education Report Card, several articles about the arts assessment appeared in a symposium issue of the electronic journal Arts Education Policy Review. The contributing authors (Eisner, 1999; Lehman, 1999; Stankiewicz, 1999) provided reactions to the assessment, commented on its strengths and weaknesses from their individual perspectives, and discussed implications of the assessment results for arts education. Creating and responding were appropriate domains for the visual arts assessment, Eisner noted; but the assessment report included little information about how the content of the items was chosen or how the performance expectations were determined. He pointed to the lack of information provided about the validity of the assessment, questioned whether the motivational tools used to engage students in the creating tasks were sufficient and whether the assessment may have depended too heavily on students’ writing ability. Lehman applauded the “extraordinarily rich and diverse mix of artworks” that were included as stimuli and the important methodological contributions the assessment made “through its creative approaches to designing tasks and scoring rubrics for the assessment of skills in performing and creating” (p. 2). However, he questioned whether the framework specifications regarding the allocation of time to the different artistic processes were truly met and pointed to the lack of information on performance included in the Arts Report Card. Further, Lehman voiced concern over the small sample sizes and the consequent limitations that sampling imposed on the assessment’s overall utility. Stankiewicz appreciated the comprehensiveness of the assessment framework and noted that the assessment provided art educators with “complex models of knowing and doing, and specific examples of performance assessments” (p. 4) that could be useful for designing state and local arts assessments. She too voiced concern about the use of small samples of only grade 8 students, which was not the intention of those who prepared the assessment framework. All three authors lamented the inability to directly compare students’ results from the most recent assessment to the assessments conducted in the 1970s.

Beattie (1999) critiqued the 1997 NAEP visual arts assessment, examining the degree of correspondence between 15 specifications included in the assessment exercise and specifications document and the actual assessment that was designed from those specifications. She found that in a number of cases the specifications were only partially met (or, in some cases, not met at all).

Several researchers obtained a grant from the National Center for Education Statistics (NCES) to conduct secondary analyses of data from the 1997 NAEP visual arts assessment (Diket, Burton, McCollister, & Sabol, 2000; Diket, Sabol, & Burton, 2002). Diket (2001) sought to identify sets of variables that were related to variation in performance on the NAEP visual arts assessment. She conducted confirmatory factor analyses to evaluate a hypothesized structural model composed of sets of demographic variables, resources variables, and opportunity-to-learn variables. Sabol (2001) used multiple regression analysis to investigate the relationships between student and school variable sets and student performance on the NAEP visual arts
assessments. The study’s purpose was to determine whether there were variables that were associated with differences in regional performance on the Creating and Responding exercises. Burton (2001) employed quartile analysis to examine the performance of those students scoring in the highest and lowest quartiles on the NAEP visual arts assessment. His goal was to pinpoint variables that seemed to contribute most to differences in the performance of these two groups.

In all three NAEP arts assessments, the scoring of complex responses to constructed-response exercises has proven to be a challenge. The 1997 NAEP visual arts assessment has stimulated interest in research on scoring guides. Researchers have investigated varied approaches to evaluating student performance on Creating exercises included as part of the NAEP visual arts assessment. During the NAEP visual arts field test, raters experimented with descriptive graphic rating scales for evaluating students’ performance on four of the production exercise blocks (Myford, 2002). Descriptive graphic rating scales are continuous score scales with two defined endpoints. A line connects these points, and descriptive phrases identify different points along the continuum. When a rater uses the scale, the rater makes a slash along the line to indicate where along the continuum the work lies. Descriptive graphic rating scales can incorporate different design features (i.e., presence or absence of a defined midpoint, presence or absence of hatchmarks along the line that connects the endpoints). Myford varied the design features of the descriptive graphic rating scales to learn about how raters used those features and to determine which features, if any, affected reliability. Siegesmund, Diket and McCulloch (2001) worked with middle school teachers to adapt the Collage exercise block for use as a classroom assessment tool. In the NAEP assessment, students’ responses to the exercises included in this block were evaluated using a series of holistic scoring guides. The teachers worked with a consultant to devise a series of five analytic scoring guides to use instead. Their goal was to develop a scoring process that would enable them to derive information from the scoring of students’ responses that would be more instructionally relevant than the limited information available from holistic scores.

THE ADVANCED PLACEMENT STUDIO ART PORTFOLIO ASSESSMENT

What Does the Assessment Presently Look Like?

The Advanced Placement Studio Art courses are three of the 34 subjects offered by the Advanced Placement (AP) program of the College Board. Like all AP subjects, the overriding mission is to document achievement in its field by high school students. Each AP course is intended to serve as an analog for the corresponding first-year or introductory college course. The annual assessments measure the degree to which students have met that challenge. In all subjects except studio art, students take exams that combine multiple-choice questions with substantial essay sections. The essays are scored by teachers of the relevant discipline in centralized, carefully monitored scoring sessions. AP grades in all subjects are reported on a scale of 1 to 5.8 The AP Studio Art program is unique: It is not a written test, but rather a portfolio assessment. As radical as this may seem in the context of standardized testing, it is obviously not so in the world of art or art education; rather, its design makes it an unusually valid or “authentic” assessment. Students submit portfolios each May; the corresponding body of work is usually the product of at least a year’s effort.

8The AP program provides the following description of the grades to both students and colleges: 5—Extremely Well Qualified, 4—Well Qualified, 3—Qualified, 2—Possibly Qualified, 1—No Recommendation. Some colleges will grant credit or advanced placement to students having grades of 3 or higher. Other colleges require grades as high as 4 or, in some cases, 5.
Currently, students interested in AP Studio Art may choose among three portfolios: Drawing, 2-D Design, and 3-D Design. Each consists of three sections, which are consistent across the portfolios. Section I, Quality, asks the students to submit five works that they feel best represent their accomplishment. In the cases of Drawing and 2-D Design, actual work is submitted; its only stipulation is that it must fit into the 18 in. × 24 in. portfolio that is sent to the schools for the exam. Students in 3-D Design do not submit any actual work; instead, they send slides (two views) of each of their five strongest works. Section II, Concentration, asks the students to submit 12 slides that document the investigation of a visual idea or problem of particular interest to the student, as well as a short written commentary. For Section III, Breadth, 12 slides in Drawing and 2-D Design, or 16 slides for 3-D Design (two views of each of eight works) are submitted to demonstrate a wide range of experiences germane to the particular portfolio.

Portfolios are most commonly submitted by high school seniors, although many juniors and even some sophomores also participate. The work in the portfolio, in many cases, has been developed over a longer period of time than 1 school year. Although many students are enrolled in classes specified as AP Studio Art classes, others work alongside somewhat less advanced or less motivated peers, and still others prepare portfolios outside of their high schools. Work may come from out-of-school as well as from in-school experiences. Thus, students are free to pursue this work in art from a wide range of vantage points; a small number, for example, prepare portfolios by working out informal mentor/apprentice arrangements with artists who are not teachers in the students’ high schools.

The requirements for each section are intentionally broad and flexible; each portfolio should ideally serve as an equivalent to the college-level course that it parallels, but few, if any, assumptions are made about how a particular teacher in a particular school with a particular group of students will best achieve that goal. It is perhaps worth noting that the program strives to avoid the uneasy sense that teachers may be “teaching to the test” in the sense of providing students with what approaches canned assignments that seem to produce high scores. Although it is impossible to eliminate that as a possibility, “teaching to the test” in this case should result in students working with commitment and passion at visual issues that stretch their knowledge, strengthen their conceptual abilities, and, of course, help them to develop the framework of technical skills that are necessary to realize their ideas. At its best, the open-ended nature of the assessment serves these goals effectively.

Each year the College Board makes available a document called the Advanced Placement Program Course Description: Studio Art. In this document, the requirements for the portfolio are delineated. All of the requirements for the portfolios are determined by the AP Studio Art Development Committee, a group of college-level and AP teachers who meet regularly and who are also involved in the annual evaluation of the portfolios. The Committee does not believe that a single AP Studio Art course can or should exist but rather encourages art educators to exercise their creativity in designing courses that will enable students to produce portfolios meeting the stated guidelines. As Askin (1985) noted, the AP Studio Art course “does not consist of a fixed body of ideas, but is affected by ongoing reevaluations of both current and past art” (p. 7). Likewise, there is no one approved course outline and/or method of teaching. Teachers have a great deal of flexibility to create AP courses in studio art that will prepare their students for the portfolio assessment.

The College Board also publishes a full-color poster each year that students and teachers receive. The poster features exemplary artworks from portfolios submitted to the AP Studio Art program in the previous year and provides a condensed version of the portfolio requirements. The three sections of the portfolio are defined, and guidelines are included for submitting works appropriate for each section.

The evaluation of the portfolios is inextricable from the curricular requirements. Because of AP’s history as a premier program rooted in holistic scoring, the Studio Art program was able
to draw on the knowledge base of sample-picking, standard-setting (training evaluators), and evaluation procedures in general. In general AP terminology, those who evaluate the portfolios are known as Readers; they work under the Chief Reader and his or her assistants, Exam Leaders and Table Leaders, during the evaluation process, known as the Reading. Everyone who scores portfolios in any capacity is either an AP Studio Art teacher with at least 3 years of experience or a college faculty member who teaches comparable first-year classes. Each section of the portfolios is scored using a 6-point scale. Section I (Quality; usually actual work) is scored by three Readers working independently of one another; Sections II (Concentration) and III (Breadth) are each scored by two Readers.

Before the Reading actually begins, the Chief Reader, Exam Leaders, and Table Leaders spend at least 3 days choosing samples of each section of each portfolio to use as training samples for the Readers. Typically, these will include very clear samples (i.e., a solid example of each score point) to start off the training for each. As the training progresses, the sample sets become more complex; they may include borderline examples, bodies of work in which the achievement demonstrated varies widely (e.g., for a Quality section, two strong pieces, two middle-level pieces, and another that is weak), an array of work from several students that varies in achievement but is all unusually strong, another array focusing specifically on photography and digital art, and so on. Working in small groups, the Readers “grade” the samples and then engage in intensive discussions with their peers and the leaders. Not until the Chief Reader is convinced that all Readers understand and can apply the scoring guides does any actual scoring take place.

Once the Reading starts, quality control procedures are an inherent part of the process. If there is a discrepancy of 3 or more points between two scores given to the same section, the portfolio section is immediately pulled and forwarded to a team of Table Leaders for resolution. In the event that an individual Reader—or the whole group—appears to be drifting away from the standards, the individual may be engaged in conversation about samples; or in more unusual cases, the Chief Reader may stop the reading for supplemental training. With these procedures and the relatively large number of independent judgments, the score reliability of each portfolio as a whole is typically about .90. The current scoring guides for the portfolios can be found on the AP Web site, AP Central, at http://apcentral.collegeboard.com/repository/sg_studio_art_02_11395.pdf. Additional information on standard-setting and the portfolio evaluation process in AP Studio Art can be found in Askin (1985), Mitchell and Stempel (1991), Myford and Mislevy (1995), and Ott (1994).

Relative success on an AP exam is intended to carry with it the reward of placement out of the corresponding course, or credit towards a major, or simply toward graduation. Thus, the exams do not really fit the idea of “high-stakes” testing. On the other hand, despite the College Board’s repeated statements that AP is not an admissions testing program, de facto policies at many institutions of higher education—particularly the most competitive ones—have informally made participation in and successful completion of AP courses highly desirable. Actual policies about how the AP grades are used vary tremendously across colleges and universities.

How Has the Assessment Evolved Since Its Inception?
What Are the Major Changes It Has Undergone?

In contrast to the NAEP visual arts assessment, AP Studio Art has a continuous history, with many modifications taking place over the last 30 years. When it began in 1972, AP Studio Art consisted of a single portfolio. Beginning with the 1979 to 1980 school year, at the behest of the Development Committee, the studio art portfolio was split into two exams, the General portfolio and the more focused Drawing portfolio. The General portfolio, which remained closer to the
original design, was seen as being more tightly related to the way art is typically taught in most high schools; it was very open with respect to media, and it required a modest amount of three-dimensional work (which was persistently the section with the lowest average scores). The Drawing portfolio was conceived as more closely analogous to a basic college drawing course. Despite expectations that it would grow rapidly in volume, the Drawing portfolio remained at roughly one third of total submissions from 1980 through 2001. The two-portfolio format provided the basic outline of the AP Studio Art program throughout this period.

Sources of Change. Over the course of the last 22 years, revisions ranging from slight to major have been made in the portfolios. The impetus for such changes may include a combination of factors such as:

- Data from statistical analyses of the portfolio scores that are produced annually
- Data from special studies conducted irregularly, such as college comparability studies
- Annual reports from the Chief Reader, which incorporate the group wisdom of all those who have evaluated the portfolios
- Pedagogical expertise of the members of the AP Studio Art Development Committee.

The role of research in AP Studio Art has largely taken the guise of these ongoing program analyses, supplemented by occasional special studies. From the beginning, ETS statisticians have produced annual analyses that track, among other statistics, the reliability of the evaluation of each section as well as of each portfolio as a whole, the correlations among the various sections, frequency distributions of scores, mean scores for every section of every portfolio, and historical distributions of the final AP grades. The analyses are produced annually, typically before meetings of the AP Studio Art Development Committee, which always reviews them. The Committee also has input into any special studies and, of course, reviews the results, which in many cases influence their subsequent decisions. This group, like all such AP committees, consists of college and university faculty members who teach the equivalent first-year courses, and experienced AP teachers. The Chief Reader, whose annual report is another mainstay of the program, attends and participates in the meetings, as do relevant ETS staff. It is the Development Committee, with the participation of the Chief Reader, which sets—and revises—the specifications for the portfolios.

Example 1: Statistical Information and Professional Expertise Intersect to Illuminate a Structural Problem. In 1983, the Breadth section (then known as Section C) of the General portfolio consisted of several subparts that had existed since the inception of AP Studio Art. C1 (sometimes referred to as CI) consisted of 12 slides of drawings and was rated on a 9-point scale; C2 (similarly referred to by some as CII) had six different specifications, each of which was graded on a scale of 0 to 2. Each Reader summed his or her six individual scores for the parts of C2. Reliability and correlations of scores were calculated on the basis of each Reader’s total score for C2, which logically ranged from 0 to 12. The six specified areas of C2 were defined as Technique, Color, Design, Spatial, Content, and Three-Dimensional. Students were required to submit at least one work in which each category was demonstrated, with a total of 14 slides among the six areas. The May 1980 *Advanced Placement Course Description: Art* (pp. 9–11) described them, in part, as follows:

Technique: a work in which the main thrust is a sensitive and personal use of specific materials and techniques.

Color: a work in which color is the major component and which clearly demonstrates the principle of visual operation of the color methodology used.
28. THE EVOLUTION OF LARGE-SCALE ASSESSMENT PROGRAMS

Design: a work in which classic design components (elements and principles) are the major concept being presented.

Spatial: a work emphasizing one or more approaches to creating the illusion of space or working with aspects of three-dimensional space.

Content: a work in which the major consideration is his or her subjective or intuitive reaction to a specific subject or set of circumstances such as a reinterpretation of nature, an observed situation, or a personal experience.

Three-dimensional: a work utilizing the partial and/or full potential of three dimensions through traditional, found or created, materials

Students received a poster that detailed the portfolio requirements and provided instructions for preparing their portfolios for evaluation. The instructions included on the portfolio reminded students that, “Although it is possible that a single work might have qualities that make it eligible for more than one of the categories, you may NOT submit the same work in more than one category of Part II.”

In his report on the 1980 Reading, the Chief Reader, William A. Lewis (1980), commented on difficulties encountered in scoring Section CII of the portfolio:

Section CII of the General Art portfolio is more complex [than CI], in the past it has been the source of difficulty for students and readers alike, and in the current reading still presents some unsatisfactory aspects. CII consists of six characteristics of art, all common to many forms of work, and generally regarded as important fundamental elements. These six categories are scored on a 00 to 02 scale, indicating the quality is not present by a 00 score, it is present in minimal, perhaps chance, fashion, or it is demonstrated deliberately, with intelligence and some distinction. (p. 3)

Lewis (1980) went on to summarize the idea behind each of the six categories and registered some of his concerns with those categories:

For example, the terms Design and Color should elicit responses stating common principles learned in class. Too seldom was there evidence of such instruction. The term 3-Dimensional describes specific 3D objects which the slides should show, such as sculpture. This category was sometimes represented by slides of paintings, flat designs, other items of a 2-dimensional nature. . . . The term Content. . . continues to be ill understood by many students. When the idea of concept is understood, the works are frequently glorious to behold, presenting thought or concept far in excess of what a young person is usually deemed capable. (p. 4)

Perhaps not surprisingly, Lewis’s concerns were reflected in statistical analysis of the portfolio evaluation. An internal ETS memo (Hecht & Bleistein, personal communication, February 8, 1980) reported that, “For the past two years, Section C-2 of the Studio Art examination has had a lower rating reliability than the other three sections of the evaluation” (p. 1). The memo then delineated the results of a study undertaken “to investigate possible explanations for the apparently lower rating reliability” (p. 1). In addition to noting a small percentage of addition errors by the readers, the study examined the rating reliability of each category, as well as the percentage of zero scores assigned by both readers for each category.

The discussion noted initially that the Three-dimensional category had the highest reliability of the categories; but it pointed out that 11% of the candidates in the study received two scores of zero in this category, and that “There is a near-perfect positive relationship, for the six criteria, between the reliability of the rating and the number of zero scores assigned” (p. 3). Almost
certainly, many of the students who received zero scores for the Three-dimensional category were among those who Lewis pointed out misunderstood the category and actually submitted two-dimensional work. In the cases of Technique and Design, probably the most accessible to teachers and students and the most certain in their demonstration, the opposite case applied: Almost no students received zero scores, so that the rating scale was again compressed to two effective points, 01 and 02.

At the time of the 1980 study, the Development Committee met less frequently than once a year; because decisions involving the specifications for the portfolios reside with that group, consideration of the study and consequent action were relatively slow. In his report of the following year, Lewis (1981) noted that, “Though the Course Description has been rewritten for clarity’s sake [an interim measure to try to improve the situation] and description of representative examples included, there remains more evidence of misunderstanding of the section than any other” (p. 3). In the fall of 1983, the Development Committee for AP Studio Art was convened. The results of the Hecht and Bleistein study were presented. Walter Askin, the Chief Reader who had succeeded Lewis, had continued to press the Readers’ related concerns. In his report, Askin (1983) made the following points:

... sections CI and CII of the general portfolio should be combined for a more accurate and efficient evaluation. A mode has been provided for discussion combining eight slides of drawings (half figure studies and half spatial drawing), four color studies (excluding atmospheric and monochromatic color schemes), four design slides emphasizing compositional or formal organization (pure design, brochures, poster, calligraphy, album cover, etc.), and four slides of sculpture or three dimensional works describing two aspects of two works,... This will eliminate the more problematic and less apt areas of “technique” and “content.” It will also, hopefully, make possible the use of [a] four point grading scale for each segment. (p. 1)

In fact, the Development Committee adopted this proposal almost exactly as presented, and went further to adopt a 4-point scale, not only for this section but for all sections of both portfolios. The one exception to the proposal was the Committee’s decision not to make any stipulations as to the nature of the eight drawings, beyond saying that they must show breadth of experience. Reactions to the change were uniformly positive from all constituents—teachers and students had a clearer outline to follow, readers were therefore able to rate the work more confidently, and statisticians documented the result: reading reliability that was, finally, in keeping with the reliabilities for the other sections of both portfolios. This arrangement stood for several years; at about the same time that readers were beginning to chafe at the difficulty of distinguishing between Color and Design, the 1992 analysis revealed an estimated true-score correlation of .97 between the ratings for those two categories—high enough to suggest strongly that the two were not measuring distinct areas of accomplishment. Because the Development Committee was by then meeting annually, the decision to create a combined set of eight slides called “Color/Design” was made and implemented quickly. That structure, which allowed students to submit the eight slides that they felt best reflected their accomplishments in the combined area, remained in place through June 2001, after which the General portfolio was retired.

In 1996, a change was made to a 6-point rating scale, primarily as an effort to distribute scores more effectively. Other changes were more pedagogical than psychometric, for example, increasing the size of the standard portfolio (and the work it could contain) from 16 in. × 20 in. to 18 in. × 24 in., and the gradually increased inclusion of photography and digital art into more sections of the General portfolio over time.

Example 2: The Program Evolves for a Better Fit With the Field. In 1998, in accordance with AP policy, a curriculum survey of colleges, universities, and art schools was conducted (Sims-Gunzenhauser, 1999, unpublished). Typically, surveys of this sort focus almost
exclusively on questions about the content of the higher education course that the AP course is intended to parallel—for example, an art history survey might ask what percentage of the course is focused on the art of various time periods and/or parts of the world. For AP Studio Art, the needs were more complicated. Anecdotal evidence for years had suggested that students with the highest scores on the AP Studio Art exam were less likely than their colleagues in other subject areas to be granted credit or advanced placement by the colleges they attended. The Development Committee decided that the survey should step back a level and inquire not only into the content of particular courses but also, more generally, into the composition of first-year or foundation programs at the responding schools. In addition, the survey included questions that essentially focused on the respondent institutions’ values with respect to what first-year students should be learning and on the kinds of work in a portfolio that would be most likely to be granted credit or advanced placement. In other words, the survey was conceived as a way of determining the best fit between the AP program and the institutions that the AP candidates might well attend.

The survey was mailed to the 300 colleges, universities, and art schools to which the greatest number of AP Studio Art candidates had sent their AP grades. Faculty members from 81 institutions responded. Table 28.3 summarizes the responses to the question about the content and organization of the first-year or foundation program.

The evidence from this set of schools, as is obvious, demonstrated a preponderance of programs that include drawing, two-dimensional design, and three-dimensional design. When the data were sorted to separate out art schools, the proportion reporting this constellation of courses was similar. Indeed, the data from this question provided a clear answer to the question of what the overall structure of the AP Studio Art offerings should be. With the support of the College Board, the decision was made early in 1999 to retire the General portfolio and introduce two new portfolios: 2-D Design and 3-D Design. These two, along with a slightly restructured Drawing portfolio, became active in the 2001 to 2002 school year. (The delay was considered critical in order for schools and teachers to have sufficient time to prepare new courses or to reorganize their existing courses.)

In order to make slightly finer decisions regarding the content of the new portfolios, the Development Committee referred to the results of a survey question that asked how likely the institutions would be to grant credit or placement for portfolios that were focused on any of a variety of art media. In practice, it has actually been unusual for a student’s portfolio

### Table 28.3

<table>
<thead>
<tr>
<th>One-Year Program</th>
<th>Two-Year Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>A sequence of two-dimensional and three-dimensional design</td>
<td>9</td>
</tr>
<tr>
<td>A sequence composed of drawing, two-dimensional design, and three-dimensional design</td>
<td>32</td>
</tr>
<tr>
<td>A sequence that includes four-dimensional design as well as drawing and two- and three-dimensional design</td>
<td>2</td>
</tr>
<tr>
<td>A sequence that is organized around conceptual themes rather than more traditional courses</td>
<td>2</td>
</tr>
<tr>
<td>A sequence that is differentiated according to students’ interests</td>
<td>0</td>
</tr>
<tr>
<td>A single, comprehensive course</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>13</td>
</tr>
</tbody>
</table>
to consist of work in a single medium; even the Drawing portfolio has always been open
with respect to media, and most students tended to include work in a range of media, if only
because the Breadth section made this a requirement for the General portfolio and a strong
suggestion for the Drawing portfolio. On the other hand, students whose primary interest was
in three-dimensional art forms, photography, or digital art were considerably more constrained
than their colleagues. In the case of students primarily interested in three-dimensional art, the
requirement to submit only flat work for Quality obviously limited their ability to present the
strongest case for their achievement; in addition, eight drawings were required in the Breadth
Section. Photographers and digital artists were able to submit prints of their work for Quality
and slides for the Concentration, but were required to submit nontechnological drawings, as
well as some sculpture, in the Breadth section.

Responses to the survey question about media showed that the level of acceptance ranged
from a high of 91% positive for drawing to 48% for digital art, to the low of 38% responding
positively to the possibility of a portfolio in film, video, or animation. On the basis of these
results, the Development Committee decided to remove the restrictions on photography and
digital art. Within the bounds of the 2-D Design portfolio, students may submit as much of
the portfolio as they like—up to and including an entire portfolio—in photography or digital
art. Obviously, the institution of the 3-D Design portfolio (for which Quality is evaluated
through slides) accommodates the students working in that area. In all of these cases, indeed
for any student who chooses to work in a single medium, there is still the need to demonstrate
breadth of experience in that section of the portfolio. The same set of survey data, though,
also constitute the rationale for eliminating film and video from inclusion in any of the AP
portfolios, although stills from a film or videotape may be submitted for 2-D Design.

Though the information presented in Table 28.3 confirmed the course outlines for the
portfolios, the Development Committee was also interested in knowing whether and to what
extent the pedagogical thrust of the portfolios and the primary values inherent in the evaluation
process were shared by the responding institutions. Still another survey question focused on the
underlying values of the responding institutions. Over the years, the Development Committee
and Chief Reader had spent a great deal of time and thought teasing out what they considered
important in student art work; the simplest example would be the tension between the group’s
desire to reward students who took risks successfully and the much easier task of discerning
(and rewarding) a high level of technical skill in a student’s work. In concert with this effort, the
scoring guides for the portfolios had evolved to address more overtly the aspects of a body of
work that underlay judgments about its level of accomplishment and to articulate those aspects
at each score point without becoming restrictive. Thus, another purpose of the curriculum
survey was to try to see whether and to what extent these broader values were shared by the
responding institutions. Table 28.4 summarizes the results from a survey question focused on
these issues.

In this case, the results were the cause of some satisfaction. Had the results been primarily
negative for the various categories listed, they would have given rise to a reexamination of
the basic conceptual underpinnings of the program. In fact, the first two items in the left-hand
column of the table, “broad experience” and “in-depth exploration,” describe, respectively,
Section III, Breadth, and Section II, Concentration, of the portfolios. The bottom three items
identify the three basic strands that are woven throughout all of the scoring guides for the
portfolios, as well as the program publications for teachers and students. Thus, these survey
data suggest that both the overall direction of the portfolio specifications and the portfolio
evaluation place the portfolios squarely in the accepted constructs of the domain, to the extent
that those constructs are represented by the responding institutions. The program is optimistic
about the scope and design of these changes; in June 2002, approximately 19,000 students
submitted portfolios for assessment, an increase of nearly 25% from 2001.
28. THE EVOLUTION OF LARGE-SCALE ASSESSMENT PROGRAMS

TABLE 28.4
If You Were Considering What Kinds of Student Work Would Persuade You to Grant Credit or Advanced Placement to a Student, How Important Would It Be to You to See Evidence of:

<table>
<thead>
<tr>
<th></th>
<th>Very</th>
<th>Moderately</th>
<th>(% Positive)</th>
<th>Neutral</th>
<th>Unimportant</th>
<th>Detrimental</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad experience</td>
<td>28</td>
<td>23</td>
<td>(39%)</td>
<td>13</td>
<td>4</td>
<td>3</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>(84%)</td>
<td>(18%)</td>
<td>(6%)</td>
<td>(4%)</td>
<td>(4%)</td>
<td>(4%)</td>
<td></td>
</tr>
<tr>
<td>In-depth exploration</td>
<td>25</td>
<td>30</td>
<td>(35%)</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>(77%)</td>
<td>(11%)</td>
<td>(3%)</td>
<td>(8%)</td>
<td>(8%)</td>
<td>(8%)</td>
<td></td>
</tr>
<tr>
<td>Concept/ideation</td>
<td>45</td>
<td>19</td>
<td>(63%)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>(90%)</td>
<td>(10%)</td>
<td>(10%)</td>
<td>(10%)</td>
<td>(10%)</td>
<td>(10%)</td>
<td></td>
</tr>
<tr>
<td>Originality/creative exp</td>
<td>42</td>
<td>24</td>
<td>(59%)</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>(34%)</td>
<td>(93%)</td>
<td>(6%)</td>
<td>(1%)</td>
<td>(1%)</td>
<td>(1%)</td>
<td></td>
</tr>
<tr>
<td>Technical skill/craftsmanship</td>
<td>43</td>
<td>23</td>
<td>(61%)</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>(32%)</td>
<td>(93%)</td>
<td>(7%)</td>
<td>(7%)</td>
<td>(7%)</td>
<td>(7%)</td>
<td></td>
</tr>
</tbody>
</table>

What Are the Challenges That This Assessment Faces?

By definition, the AP Studio Art portfolios involve relatively unstructured tasks. Within the very general constraints of its three major divisions—Quality, Concentration, and Breadth—the program’s philosophy is to allow teachers and students to work toward fulfillment of the requirements through the curriculum and teaching strategies that suit them best. Thus, there are literally an infinite number of ways of “solving” each section. A natural outcome of this fundamental stance is that the scoring guides for each section are holistic and quite general. Rather than trying to define in a restrictive or constrictive way what one should find in each section (barring the most basic guidelines, such as “12 slides”), the guides outline a variety of traits of work that one might find at each score level. For example, some students’ work will be more technically competent and less interesting in terms of ideation, whereas other students may demonstrate relatively strong conceptual skills or be highly expressive but less accomplished in terms of skill in manipulating the medium. Similarly, not all of the works submitted for a single section will always demonstrate the same level of accomplishment. The scoring guides cover these situations. Great effort over many years has gone into making the scoring guides value free with respect to media, style, and content (e.g., middle-level, tightly rendered drawings should receive the same scores as highly expressive drawings that demonstrate an equivalent level of achievement). Reader training has evolved over time to focus carefully on these issues. To the extent that it is possible to discern such variables from artwork, care is also taken to keep the scoring value free with respect to the gender of the students, their ethnicity, or any other similar variables. In addition, special efforts have been made in recent years to augment reader training with respect to photography and digital art. These remain areas of concern, and efforts in these directions are ongoing.

Scoring guides are viewed as evolving documents and are reviewed each year. The introduction of the two new design portfolios logically necessitated the development of new scoring guides for 2-D Design and 3-D Design. These were initially rooted in the old guides for Color/Design and Sculpture/Three-Dimensional Design within Section III of the old General portfolio. They have, however, been worked through systematically and with considerably more depth. As the new portfolios become more familiar, it is likely that the guides will be revised further. Similarly, experience from the portfolio Reading will inform efforts to communicate
more effectively with teachers about the nature of the two design portfolios and expectations for student work.

As with all AP subjects, the student’s own teacher is not part of the assessment process. Many AP teachers serve as Readers, but if during the portfolio Reading, they see work by a student whom they know or whose work they recognize, they are required to ask another Reader or a Table Leader to evaluate that work. Not allowing the classroom teacher to participate in the evaluation (or contriving a systematic method of collecting input from classroom teachers) means that information regarding the context in which the work was produced is largely absent. Beyond what they may read in the student’s written commentary on the Concentration, Readers do not know anything about the student—whether she or he has overcome unusual obstacles, or been a leader in the class, or any of a host of other factors that can affect performance. Because Readers are not only human, but are all teachers, probably every one of them has often wished for the chance to talk with students about their portfolios. On the other hand, one can argue that the body of work should stand on its own, because its qualities are the basis for the judgments that colleges and universities have to make about whether or not to grant credit or placement. There is also the inevitable argument that this form of evaluation is particularly objective in the sense that all students’ work is evaluated on the basis of the same carefully worked-out scoring guides.

To date, only the single AP grade is reported for Studio Art. The high score reliability of the portfolios relies on pooling all of the judgments on all of the sections. It is the sense of the College Board and Educational Testing Service that only scores with high reliability should be reported. Each section by itself would not meet that statistical challenge. In the interest of supporting both student learning and teaching, it is hoped that this dilemma can be at least partially resolved in the future.

Have There Been Research/Evaluation Studies of This Assessment and Reports of Its Findings?

Several recent investigations have compared AP Studio Art to other approaches to teaching studio art. In Tomhave’s (1999) study, students in two visual arts classrooms prepared portfolios that were then evaluated by AP-trained examiners. Students in one classroom were taught using the International Baccalaureate Art/Design curriculum, whereas students in the other classroom were taught using the AP Studio Art curriculum. Portfolio scores for the two classrooms were compared. Tomhave found no significant differences between the groups in the scores they received on their portfolios. The researcher also examined the relationship between the portfolio scores and the end-of-year course grades the students’ teachers gave and reported that the correlation between the two was .01.

Willis (1999) conducted a qualitative study to compare and contrast the assessment procedures employed in the AP Studio Art program and in the International Baccalaureate Art/Design program. Through surveys and interviews, Willis studied the training of the teachers and the examiners who conduct the assessments, and the examiners’ interpretations of various descriptors included in the scoring guides. He also compared the moderation processes employed in the two programs and the procedures used to handle discrepancies when examiners’ scores of the same portfolio disagreed.

Blaikie (1994) investigated the assessment processes employed in three visual arts programs: Arts PROPEL, the International Baccalaureate Art/Design program, and the AP Studio Art program. Using content analysis, she reviewed curriculum materials prepared by each program and literature describing each program in order to identify a set of assessment criteria that she then employed to compare their assessment processes. Through her analysis, she identified common themes that crossed programs, as well as themes that were unique to individual programs.
Other researchers have conducted case studies of the AP Studio Art program. As part of his case study of the Virginia Beach art program, Wilson (1984, 1986) observed AP Studio Art classes in Kempsville High School and conducted interviews with the teacher and students to learn about the role that AP played within the Virginia Beach program. The Virginia Beach site was one of seven sites chosen for study by a team of investigators because it provided a well-rounded arts education program that included instruction in art history, art criticism, and art production.

Mitchell and Stempel (1991) prepared six case studies of performance assessments for the U.S. Office of Technology Assessment to illustrate what performance assessment is capable of accomplishing and to demonstrate its potential for assisting in educational reform efforts. One of their case studies focused on the AP Studio Art program as an example of a large-scale national portfolio assessment. To prepare their case study, they observed the 1990 portfolio reading, reviewed various program documents, and interviewed key program personnel as well as AP teachers, students, and college professors who teach introductory art courses.

Myford and Mislevy (1995) used two complementary approaches—quantitative and qualitative—to gather, analyze, and interpret information about critical aspects of the AP Studio Art assessment process. Their study provided an illustration of the use of the two approaches in tandem for the quality control monitoring of large-scale assessment systems. For the naturalistic portion of their study, they analyzed rating data from the 1992 AP Studio Art portfolio Reading using a many-faceted Rasch measurement model. They then engaged raters in discussions about hard-to-rate portfolios to discover why certain portfolios were more difficult to rate than others. Through their interviews, they gained an understanding of the cognitive processes underlying the unusual or discrepant patterns that the statistical analyses picked up in the ratings of these portfolios.

CONCLUSION

As is evident from this discussion, the NAEP visual arts assessment and the AP Studio Art assessment have evolved in very different ways since their inception. That evolution has created some concerns regarding the comparability of the scores provided by each of the two assessments over time.

In the case of the NAEP visual arts assessment, over 20 years had elapsed between the two most recent assessments. In that time, the field of visual arts education had changed rather dramatically, as Stankiewicz (1999) explained:

More is expected of art education today. As part of the political construct arts education, the visual arts are expected to prepare students for work, to focus efforts for interdisciplinary learning, to provide challenging subject matter drawn from four foundational disciplines, and more. Expectations for student performance and learning are more complex as well. Memorizing art vocabulary, drawing figures in motion, designing a room of one’s own, and other exercises that tend to call forth specific behaviors have given way to more complex, global ability areas [e.g., see Erickson et al., 1999]. We look for the ability to synthesize knowledge and skills, the integration of making and response. Those are positive advances. The NAEP arts framework was designed to encompass the complex nature of creating, performing/interpreting, and responding in the arts, to address specific content—knowledge and skills—in the arts disciplines. (p. 3)

As we have seen, the design framework used to construct the 1997 assessment differed radically from the frameworks used in 1975 and in 1978. Consequently, the assessment and exercise specifications arising out of those frameworks were also very different. Indeed, a close
comparison of the content outlines driving those assessments might lead one to conclude that the construct of learning in the visual arts had been defined in very different ways. A convincing case could be made to argue that the constructs themselves may not have been comparable. As noted in the 1999 *Standards for Educational and Psychological Testing*, tests that measure different constructs or are built from different sets of specifications should not be considered comparable.

Several authors (Eisner, 1999; Lehman, 1999; Stankiewicz, 1999) have lamented the inability to directly compare students’ results from the most recent assessment to the assessments conducted in the 1970s. As Lehman (1999) explains,

> Throughout its history, NAEP has emphasized the importance of identifying trends in student achievement and has compared the results of each assessment with those of previous assessments in that discipline. Indeed, measuring progress is so fundamental to NAEP that it is reflected in its name. Nevertheless, although there have been two previous NAEP assessments in music and two in visual arts, no information whatever has been provided concerning how student achievement in 1997 compares with that of previous years. That omission represents another huge lost opportunity. (p. 5)

NAEP has been able to maintain the comparability of assessments in other disciplines by implementing the statistical process of equating test scores from one assessment to the next. In theory, equating should be able to provide accurate score conversions for any set of students that are drawn from the population for which a given test has been developed and should then allow for comparisons to be drawn between student performances over time. Additionally, the same score conversion should be applicable irrespective of score interpretation or intended use. However, in the case of the NAEP arts assessment, equating was not feasible, nor would it have been a psychometrically acceptable practice in which to engage, given the warnings provided in the 1999 *Standards for Educational and Psychological Testing*:

> It is not possible to construct conversions with these ideal properties between scores on tests that measure different constructs; that differ materially in difficulty, reliability, time limits, or other conditions of administration; or that are designed to different specifications. (pp. 51–52)

It is important to note that comparisons were drawn between students’ performances on selected exercises from the 1975 and 1978 assessments and reported (National Assessment of Educational Progress, 1981a). A set of common exercises were administered in both those assessments to allow for direct comparisons to be made. Great care was taken to ensure that the conditions of administration were comparable across the two assessments and that similar time limits were observed. For the six common open-ended tasks, the same scoring guides were employed to score students’ responses in both assessments, and one set of raters scored all those responses (National Assessment of Educational Progress, 1981b). Having all those factors in place made it psychometrically feasible to draw justifiable comparisons and to identify trends in performance. However, none of these factors were in place for the 1997 assessment. There were no common exercises from earlier NAEP visual arts assessments, the conditions for assessment administration were very different from the conditions in effect in the 1970s, and different time limits were imposed. All these factors mitigated against being able to compare results from past NAEP visual arts assessments to results from the 1997 assessment.

In the case of the AP Studio Art portfolios, change over the last 30 years has, for the most part, been incremental. To maintain comparability of student scores from one year to the next, the AP Studio Art program relied in the beginning on the consistency of the group of people who served as Readers, and their passing on of accrued wisdom about levels of
accomplishment. As the group of Readers has grown (there were 82 Readers in 2003), more formal and more specific scoring guides have been developed. The year 1996 can be seen as a watershed; it was then that section-specific rubrics were instituted. These have been refined with input primarily from the Chief Reader and Table Leaders but also, to some extent, from all of the Readers. Because there are no “questions” that change each year, the scoring guides work well as enduring, though, mutable, documents. Throughout the Reading, quality control procedures such as reviewing every instance of widely divergent scores given to the same body of work keep the group as a whole, or individual Readers, from drifting away from the standards.

In addition, the membership of the Reading group is relatively constant by design. Ideally, about 15% of Readers in any year are serving in that capacity for the first time. Not everyone who evaluates portfolios necessarily returns; the group as a whole, though, represents a strong and broad pool of expertise. Readers may serve a 6-year term, after which they retire if they do not move on to become Table Leaders. A critical aspect of the standard-setting discussions that serve as training is that all Readers, from the most to the least experienced, take part in them. Those with experience have obvious breadth of knowledge and familiarity with the assessment process to contribute; but those who are new often bring new viewpoints, or even new challenges, to the discussions. It is in part working through these divergences that keeps the program vital.

The final AP grades that are reported to candidates are determined through what is known as “grade-setting.” All of the raw scores for each student are converted by a computer program into a single number, the composite raw score, which ranges from 0 to 72 points. In an intensive meeting with statisticians and program staff from both Educational Testing Service and the College Board, the Chief Reader decides what the lowest composite score will be for each AP grade on the 1 to 5 scale that is used for reporting. He or she is aided by historical data; at least as important, though, are detailed tables that show the average score on each section of the portfolio for students at each composite score point. Thus, the Chief Reader is able to make clear connections between actual judgments made by the Readers and the final AP grades.

Implications of These Large-Scale Assessments for Future Assessments. At a time when accountability in education is again becoming a high-profile national issue, these two large-scale art assessments bear consideration. Their differences are obvious: NAEP relies on group data and reports no individual scores; nor does it have a direct impact on the future course of students’ education. It required no particular curriculum in visual arts, but rather measured what students had learned from existing curricula. AP Studio Art, though its curricular requirements are broad and flexible, does demand support for some form of advanced study in art. It not only allows but also encourages an individual approach to curriculum and, therefore, individually distinctive results. Those results are then assessed for each student and do have some impact on the student’s future work in the field.

Nonetheless, an overwhelming link between the two is, of course, the subject matter that they share. It is perhaps not coincidental that these two ground-breaking assessments exist in a field of endeavor that is frequently devalued by schools and is seen by the public as nonessential or less rigorous than the other core subjects. It is true that art is “different”: To demonstrate mastery by definition means to demonstrate independent thinking and a willingness to move away from proven solutions. So, too, the two assessments profiled here have moved away from the traditional models that we have come to know as large-scale assessment. One might ask whether it is because of the perceived lesser importance of the arts that they have been allowed to do so. But one may also ask whether the traditional academic subjects might not benefit from this work. Imagine, for example, a large-scale history, chemistry, or Spanish assessment that included a requirement for students to define, pursue, and document an individual interest. Or,
at a more basic level, common acceptance of classroom-based English composition assessment for which scoring guides actually acknowledged the writer’s creativity as well as command of syntax and organization.

Were this to become an educational goal, a raft of difficulties would immediately arise: If work is accomplished outside of class, how can evaluators know that it is the work of the individual who turns it in? If creativity is to be assessed, what value does it hold relative to the mechanics of language? Will two or more evaluators ever be able to agree? The answers to these and other thorny questions are not simple. Designing assessment of this sort, whether for a national program or a district-based experiment, is arduous and intellectually demanding. NAEP and AP stand as evidence that it can be done, and done well. Perhaps expanded efforts of this sort are worth undertaking: If “teaching to the test” comes to mean “teaching students to work rigorously in a discipline, to think independently, and to tap their potential,” then assessment will truly serve education.

ACKNOWLEDGMENTS

We would like to acknowledge the helpful comments and suggestions from Hilary Persky, Steve Lazer, Rick Morgan, and Dennie Palmer Wolf in the preparation of this paper. The material contained herein is based on work supported by the Educational Testing Service and the University of Illinois at Chicago. Any opinions, findings, conclusions, and recommendations expressed are those of the authors and do not necessarily reflect the views of the Educational Testing Service, the College Board, the University of Illinois at Chicago, the National Center for Education Statistics, or the National Assessment Governing Board.

REFERENCES

28. THE EVOLUTION OF LARGE-SCALE ASSESSMENT PROGRAMS


Ward, B. J. (1982). A look at students’ art achievements: Results from the National Assessment of Educational Progress. Visual Arts Research, 8(16), 12–18.


