Appendix A

Appendix A

Procedures

Overview of Procedures Used for the NAEP 1997 Arts Assessment

Introduction

Conducting a large-scale assessment such as the National Assessment of Educational Progress (NAEP) entails the successful coordination of numerous projects, committees, procedures, and tasks. This appendix provides an overview of the NAEP 1997 arts assessment's primary components: *Framework,* instrument development, administration, scoring, and analysis. A more extensive review of the procedures and methods used in the arts assessment will be included in the forthcoming *NAEP 1997 Arts Analysis Technical Report.*[1]

The NAEP 1997 Arts Assessment

The *Framework* for the 1997 arts assessment was produced under the auspices of the National Assessment Governing Board through a consensus process managed by the Council of Chief State School Officers, which worked with the College Board and the Council for Basic Education. The framework was developed over an eighteen-month period between September of 1992 and March of 1994. The following factors guided the process for developing consensus on the arts *Framework* and its exercise specifications:[2]

- The assessment should affirm and articulate the arts as ways of knowing and forms of knowledge with a unique capacity to integrate the intellect, the emotions, and physical skills in the construction of meaning.
- The assessment should assess students' knowledge, attitudes, and performance in the modalities and forms of expression characteristic of the arts (music, dance, painting or drawing, acting) as well as verbal or written linguistic modes, i.e., writing or talking about the arts.
- Assessment should go beyond quantification to include critical judgment. An effort should be made to ensure that reporting includes descriptive information on student performance as well as numerical data.
- The feasibility of producing appropriate subscales by artistic process (Creating, Performing, Responding) within each arts discipline should be investigated. If feasible, separate subscores for artistic process within each arts discipline should be reported by grade level.
- The assessment should use a common list of background variables to recognize differences and inequities in school resources and the conditions related to achievement, such as teacher qualifications, instructional time in the arts, school structure, cultural and social background of the school community, and incentives. This recognition must be evident in

1 in press

2 National Assessment Governing Board. (1994). *Arts education assessment framework* (pre-publication ed.) Washington, DC: Author; National Assessment Governing Board. (1994). *Arts education assessment and exercise specifications* (pre-publication ed.) Washington, DC: Myford, C. and the College Board.

Figure A.1 **Descriptions of the Three Processes**

Creating	Refers to generating original art. This may include, but should not be limited to, the expression of a student's unique and personal ideas, feelings, and responses in the form of a visual image, a character, a written or improvised dramatic work, or the composition or improvisation of a piece of music or a dance.
Performing/ Interpreting	Means performing an existing work, a process that calls upon the interpretive or re-creative skills of the student. Typically, "performing" an existing work does not apply to the visual arts, where reproducing an artist's existing work is not central. However, it does suggest the engagement and motivation involved in creating a work of art.
Responding	Varies from that of an audience member to the interactive response between a student and a particular medium. The response is usually a combination of affective, cognitive, and physical behavior. Responding involves a level of perceptual or observational skill; a description, analysis or interpretation on the part of the respondent; and sometimes a judgment or evaluation based on some criteria which may be self-constructed or commonly held by a group or culture. Responding calls on higher order thinking and is central to the creative process. Although a response is usually thought of as verbal (oral or written), responses can and should also be conveyed non-verbally or in the art forms themselves. Major works of art in all traditions engage artists in a dialogue that crosses generations.

SOURCE: *Arts education assessment framework.* National Assessment Governing Board. (1994). Washington, DC: Author.

reporting the data. Results have meaning only in terms of the availability and continuity of arts education.

- The assessment should address both processes and products, and expand the public's information about the importance of each.
- The assessment should be based on a comprehensive vision of arts education and should communicate that vision clearly. The assessment should focus on what ought to be in arts education rather than what is, but idealism should be tempered with reality. Hence, exercises should model multifaceted and thoughtful activities, without making unreasonable demands on time, materials, and human resources.
- To stimulate support for arts education, the assessment should produce information useful to a variety of audiences—students, artists, teachers, and administrators; local, state and national policymakers; and community members such as parents, business persons, etc.—and be disseminated in a variety of ways for different audiences.
- The assessment should reflect a pluralistic view of arts education, both in terms of individual products and the cultural bases of the arts. It should be oriented toward the demonstration of student learning, be sensitive to a variety of instructional approaches, include the range of contemporary theories evident in arts education, and include examples of appropriate exercises addressing universal themes.

The Framework Design

The *Framework* for the 1997 arts assessment is represented as a matrix with two dimensions represented by the four arts disciplines (dance, music, theatre, and visual arts) and the three processes (Creating, Performing, and Responding). Figure A.1 describes the Creating, Performing, and Responding processes in the four arts disciplines.

Distribution of Assessment by Process Areas

Table A.1 summarizes the percentage of assessment time devoted to each process in the arts assessment, by arts discipline. Care was taken to ensure the congruence between the percentages used in the assessment (actual) and those as recommended by the discipline-specific subcommittees of the framework committee (specified).

Table A.1 **Specified and Actual Distribution of Assessment Time Across Arts Processes**

	Music		Theatre		Visual Arts	
	Actual	Specified*	Actual	Specified*	Actual	Specified*
Creating	30%	20-30%	–	–	70%	50-70%
Performing	35%	35-45%	–	–	–	–
Creating/Performing	–	–	67%	60%	–	–
Responding	35%	30-40%	33%	40%	30%	30-50%

* *1996 NAEP Arts Education Assessment Framework* (pre-publication ed.)
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1997 Arts Assessment.

The Assessment Design

The assessments in music, theatre, and visual arts included "blocks"[3] or sets of questions, of approximately 25 or 50 minutes. Each block consisted of one or more stimuli and sets of multiple-choice, constructed-response, or Creating/Performing items to assess students' mastery of material. Table A.2 summarizes the number of blocks by arts discipline. The grade 8 music assessment included five Creating/Performing blocks and four Responding blocks. The five Creating/Performing blocks were divided into three Creating/Performing blocks for students in the general population, and two Creating/Performing blocks for students currently active in some type of music activity. All students sampled for the music assessment completed one of the Creating/Performing blocks for the general student population and two Responding blocks. In addition, a small sample of students who indicated current involvement in a music activity completed one of the two additional Creating and Performing blocks as a fourth block. The grade 8 theatre assessment included three Creating/Performing blocks and four Responding blocks. All theatre students completed one Creating/Performing block and two Responding blocks. The grade 8 visual arts assessment included three Creating blocks and four Responding blocks. All students sampled for visual arts completed either one Responding block and one Creating block or two Responding blocks.[4]

Table A.2 **Total Number of Blocks by Arts Discipline**

	Music		Theatre		Visual Arts	
	Total	Number Completed by an Individual Student	Total	Number Completed by an Individual Student	Total	Number Completed by an Individual Student
Creating/Performing	5	1*	3	1	3	0 or 1
Responding	4	2	4	2	4	1 or 2

NOTES: The five Creating/Performing blocks in the music assessment included 3 blocks for the general student population and 2 blocks for a subsample of students who were currently enrolled in a music activity.

* Students who were selected as part of the subsample of music students completed 2 blocks: 1 block of activities for students from the general population and 1 block of activities for students in the music subsample.

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1997 Arts Assessment.

3 "Blocks" are collections of questions grouped, in part, according to the amount of time required to answer them.

4 In visual arts, three of the four Responding blocks featured Creating tasks.

Item Types and Scoring Guides

The data presented in Table A.3 reflect the number of questions by item type for the 1997 arts assessment. The assessment pool for the three arts disciplines assessed contained a total of 168 unique questions — 41 multiple-choice, 97 constructed-response, and 30 Creating/Performing tasks. Most of the Creating/Performing tasks were scored with multiple scoring guides. The total number of scoring guides for such tasks was 106, with 63 in music, 22 in theatre, and 21 in visual arts.

Using information gathered from the field test, the booklets were carefully constructed to balance time requirements for the questions in each block. More information on the design of the assessment is presented in the forthcoming *NAEP 1997 Arts Analysis Technical Report.*

Student Questionnaires

Each booklet in the assessment also included several sets of background questionnaires. Students sampled for the 1997 arts assessment completed one 5-minute set of student demographic background questions and one 10-minute set of subject-specific background questions. The subject-specific background questionnaires were designed to gather contextual information about students, their instructional and out-of-school arts experiences, and their attitudes toward the art domain in which they were being assessed. The specific content of each of the background questionnaire sections is provided below:

- The student demographic background questionnaire included a common set of questions about students' race/ethnicity, language spoken in the home, parents' level of education, number of reading materials in the home, amount of television watched by students, the amount of homework assigned, and which parents live at home.
- The student subject-specific questionnaires covered three general categories of information: students' interest in the subject; students' in-school experiences in the subject; and students' out-of-school experiences in the subject. These 10-minute sections contained 37 questions for music, 45 questions for theatre, and 33 questions for visual arts.

The types of questions asked in each of the three categories of the subject-specific questionnaires are described as follows: students' interest in the subject included students' ratings of their interest and ability in the subject. For example, in music, three of the statements to which students were asked to respond "Agree," "Not Sure," or "Disagree," included: "I like to listen to music," "I think I have talent for music," and "People tell me I am a good musician." (Analysis of the relationship between students' reported interest in an arts discipline and assessment performance is beyond the scope of this report.) Students' in-school experiences were characterized by the frequency with which their teachers provided various subject-related instructional activities during class and by student participation in various arts-related activities during school. Students' out-of school experiences were characterized by the frequency with which students were involved in various arts-related activities outside of school, not in connection with school work.

Table A.3 **Distribution of Questions by Item Type**

	Music	Theatre	Visual Arts
Multiple-Choice	21	8	12
Constructed-Response	35	38	24
Performing/Creating	14	6	10

SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1997 Arts Assessment.

Teacher Questionnaire

To supplement the information on instruction reported by students, the theatre teachers of the targeted students participating in the NAEP theatre assessment were asked to complete a questionnaire about instructional practices, teaching backgrounds, and characteristics. The results of the field tests in music and visual arts showed high percentages of missing data for the music and visual arts teachers' questionnaires. Because of this, teacher questionnaires were not administered in the operational music and visual arts assessments. The theatre teacher questionnaire was divided into three parts. The first part contained 85 questions about teachers' general educational background and training. The second part contained 25 questions pertaining to teachers' background, activities, and preparation in theatre. The third section contained 50 questions on specific instructional practices.

School Questionnaire

Principals of students sampled for the assessment were asked to complete a questionnaire about the school's characteristics and students' access to instruction in the arts. The school questionnaire covered three broad areas. The first part pertained to the availability of courses in the arts and students' access to computers. The second part asked questions about the status of staff members teaching in the arts, the facilities and available resources for the arts, and the existence of special programs in the arts, such as artists-in-residence and summer arts programs. The final part of the school questionnaire pertained to demographics at the school, such as school enrollment. It also included variables used to describe the general climate of the school, such as attendance rates of students and staff, and the frequency of various problems in the school.

SD/LEP Questionnaire

An additional questionnaire was designed to gather information about students with disabilities (SD) and students who have limited English proficient (LEP) skills.

The SD/LEP questionnaire was completed by a school staff member knowledgeable about those students who were selected to participate in the assessment and who were identified as (1) having an Individual Education Plan (IEP) or equivalent plan (for reasons other than being gifted or talented) or (2) having limited English skills. A questionnaire was completed for sampled students who were disabled or had limited English skills regardless of whether the student participated in the assessment. Each questionnaire took approximately five minutes to complete and asked about the student's background and the special programs in which he or she may have been enrolled.

National School and Student Samples

The *Framework* called for assessments in all four arts disciplines at grades 4, 8, and 12. In 1995, field tests were conducted in all of the arts disciplines at grades 4 and 8. Field tests for all of the arts disciplines at grade 12 were conducted in 1997. Because of financial limitations, it was not possible to conduct final assessments in all arts disciplines at grades 4, 8, and 12. Members of the Arts Standing Committee therefore recommended that final assessment activities be conducted at grade 8 so that the *Framework* could be fully assessed at one grade.

Results of the field test in dance indicated that too few schools offered dance instruction at grade 8 to obtain a nationally representative sample of dance-educated students. Based on these findings, members of the National Center for Education Statistics (NCES), the National Assessment Governing Board, and representatives from the dance community recommended that a targeted sample of grade 8 students complete the dance assessment. However, a student sample of sufficient size and distribution could not be located.

Based on the results of the 1995 field test in theatre, members of the Theatre Test Development Committee recommended that theatre items be administered only to students who had undergone theatre instruction. Therefore, the theatre assessment at grade 8 was administered to a "targeted" sample of students selected only from schools with theatre programs.

The visual arts assessment was administered to a nationally representative random sample of students. All four of the Responding blocks and three of the Creating/Performing blocks in the music assessment were administered to a

nationally representative random sample of students. In addition, two additional Creating/Performing blocks of exercises were administered to a subsample of students who were currently enrolled in some type of in-school or out-of-school musical activity.

For the music and visual arts assessments, the national and regional results presented in this report are based on nationally representative probability samples of all eighth-grade students. The results for theatre are based on a representative sample of eighth-grade theatre students representing those schools in the nation with an extensive theatre curriculum. The samples were selected using a complex multistage sampling design that involved sampling students from selected schools within selected geographic areas across the country. The sample design had the following stages:

1. Selection of geographic areas (a county, group of counties, or metropolitan statistical area)
2. Selection of schools (public and nonpublic) within the selected areas
3. Within each sampled school, intact classrooms of students were selected for the music and visual arts assessments. In order to ensure random sampling of students with training in each arts discipline, criteria for sampling classrooms specified that the subject taught in each classroom selected should not be the subject being assessed (e.g., classes sampled for the visual arts assessment could not be visual arts classrooms; classes sampled for the music assessment could not be music classrooms). For the theatre assessment, students were sampled randomly from within eligible schools.

Each selected school that participated in the assessment and each student assessed represents a portion of the population of interest. Sampling weights are needed to make valid inferences between the student samples and the respective populations from which they were drawn. In addition, NAEP oversamples nonpublic schools and schools in which more than 15 percent of the student population is non-White. Sampling weights adjust for disproportionate representation due

Table A.4 **1997 School, Student, and Teacher Questionnaire Participation Rates**

	Weighted School Participation Rate Percentages	Total Number of Schools Participating	Weighted Student Participation Rate Percentages	Total Number of Students Assessed	Weighted Percentage of Students Matched to a School Questionnaire	Total Number of Students Matched to a School Questionnaire
Music	**80**	**98**	**91**	**2,275**	**94**	**2,114**
Public	79	84	91	1,999		
Nonpublic	83	14	94	276		
Theatre	**67**	**42**	**82**	**1,386**	**92**	**1,193**
Public	69	40	79	1,335		
Nonpublic	40	2	93	51		
Visual Arts	**84**	**128**	**91**	**2,999**	**93**	**2,799**
Public	84	116	90	2,756		
Nonpublic	85	12	95	243		

	Teacher Questionnaire Participation Rate Percentage	Total Number of Teachers Completing a Questionnaire	Weighted Percentage of Students Matched to a Teacher Questionnaire	Total Number Students Matched to a Teacher Questionnaire
Theatre	**95**	**57**	**98**	**1,320**

NOTE: A total of 567 students from the reported sample for music completed an additional supplemental music Creating or Performing block.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1997 Arts Assessment.

to such oversampling. In the analysis of student data and reporting of results, nonresponse weighting adjustments have been made at both the school and student level, with the aim of making the sample of participating students as representative as possible of the entire eligible eighth-grade population. For details of the nonresponse weighting adjustment procedures, see the forthcoming *NAEP 1997 Arts Analysis Technical Report*.

All eighth-grade students sampled for the theatre assessment were considered eligible if they had completed at least 30 classroom hours of instruction in theatre by the end of the 1996-97 school year. There were no course requirements for eligibility for the eighth-grade students sampled for the visual arts assessment. In music, there were no course requirements for eighth-grade students sampled to complete the blocks designed for the general student population. Students sampled for the supplementary Creating/Performing blocks in music were selected from the sample of students initially selected from the general student population. Eligibility for the supplemental Creating/Performing blocks was limited to students who indicated that they either sang in a school chorus, sang outside of school, took singing lessons, played a musical instrument, took instrumental lessons, or played in a band or orchestra in school or outside of school. School officials advised assessment staff on SD/LEP accommodations necessary for individual students sampled for an assessment in a given arts discipline. Students for whom recommended SD/LEP accommodations could not be made were classified as ineligible for the assessment.

For the national sample, Table A.4 contains, for public and nonpublic schools, the unweighted total number of participating schools, total number of students assessed, and the weighted school and student participation rates. In addition, Table A.4 provides the weighted percentages of students who were matched to a completed school questionnaire. For theatre, data indicating the overall teacher questionnaire participation rate and the total number of teachers completing a questionnaire are provided. In addition, the weighted percentage of theatre students and total number of students matched to a completed theatre teacher questionnaire are indicated.

The weighted school participation rates are calculated from the number of schools that were initially selected for the assessment. For each arts discipline, the numerator of this rate is the sum of the number of students represented by each initially selected school that participated in the assessment. The denominator is the sum of the number of students represented by each of the initially selected schools that had eligible students enrolled. The denominator included both participating and nonparticipating schools.

The weighted percentages of students who participated in the assessment reflect the percentage of the eligible student population from participating schools within the jurisdiction, and represent the students who participated in the assessment in either an initial session or a make-up session. The denominator of this rate is the sum, across all assessed students by arts discipline, of the number of students represented by each selected student who was eligible to participate, including students who did not participate.

The weighted school and student participation rates for the theatre sample are lower than those typically obtained in NAEP, and, as is evident in Table A.4, are substantially lower than those obtained for the music and visual arts samples. As noted above, nonresponse adjustments were made to the sampling weights used in the analysis of all three arts disciplines in order to compensate for school and student nonparticipation. However, the lower participation rates in the theatre assessment do raise questions about the effectiveness of these adjustments and the possibility of bias in the results.

In order to provide some evidence on the extent of potential bias, analyses were conducted comparing the participating theatre schools to all schools in the theatre sample with respect to five demographic characteristics: metro status

(metro, nonmetro); type of locale (large city, mid-size city, urban fringe of large city, urban fringe of mid-size city, large town, small town, rural); affiliation (public, nonpublic), school type (public, Catholic, other religious, nonsectarian); and region (Northeast, Southeast, Central, West). In addition, comparisons were also made for three additional school enrollment variables: grade 8 enrollment, total school enrollment, and percentage of minority enrollment. In general, the participating schools were similar to the full sample of schools for most of the demographic and school enrollment variables. The participating schools were, however, more likely to come from large city locales and had, on average, lower grade 8 and total enrollments than nonparticipating schools.

Data Collection and Scoring

Materials from the NAEP 1997 arts assessment were shipped to National Computer Systems, where trained staff evaluated the responses to the constructed-response questions using scoring rubrics prepared by Educational Testing Service (ETS). Each constructed-response question had a unique scoring guide that defined the criteria used to evaluate students' responses. The extended constructed-response questions generally were evaluated with four- or five-level scoring guides, while the short constructed-response questions generally were scored with two- and three-level scoring guides. Across arts disciplines, the responses of students who *skipped* a question or performance task (but who answered questions positioned later in the test booklet or in the Creating/Performing block) were scored as incorrect for unanswered multiple-choice questions or as Level 1 (Inadequate/Unsuccessful/Unacceptable) for unanswered constructed-response items or performance tasks. Data for students who *did not reach* a given question in a test booklet or in a Creating/Performing block were excluded from analysis for that question. A student's response for a given question was classified as *not reached* if a student failed to answer the question and all others following it in a given test booklet or Creating/Performing block.

For the arts assessment, more than 240,000 constructed responses were scored. This number includes rescoring to monitor inter-rater reliability. One hundred percent of student responses for theatre performing tasks were rescored in order to determine reliability rates. For all other types of questions and tasks across all arts disciplines, 25 percent of the student responses were rescored. The overall percentages of exact agreement for the 1997 national reliability sample were 89.6 percent in music, 84.6 percent in theatre, and 86.2 percent in visual arts.

Data Analysis and IRT Scaling

After the assessment information had been compiled in the NAEP database, the data were weighted according to the sample design and the population structure. The weighting for the samples reflected the probability of selection for each student as a result of the sampling design and adjustment for nonresponse. Through poststratification, the weighting assured that the representation of certain subpopulations corresponded to figures from the U.S. Census and the Current Population Survey.[5]

Analyses were then conducted to determine the percentage of students who gave various responses to each cognitive and background question. Item response theory (IRT) was used to estimate average proficiency for the nation and various subgroups of interest within the nation. Mean percent-correct scores were developed for items classified as Creating, Performing, or Creating/Performing, and percentages of each response to each item computed.

5 For additional information about the use of weighting procedures in NAEP, see Johnson, E.G. (1989). "Considerations and techniques for the analysis of NAEP data." *Journal of Educational Statistics*, 14(4), 303-334.

IRT models the probability of answering a question correctly as a mathematical function of proficiency or skill. The main purpose of IRT analysis is to provide a common scale on which performance can be compared across groups, such as those defined by age, assessment year, or subpopulations (e.g., race/ethnicity or gender) and to make it possible to assess groups on more material than would be practical to give to any one student.

In the NAEP design, students do not receive enough questions about a specific topic to permit reliable estimates of individual performance. Traditional test scores for individual students, even those based on IRT, would contribute to misleading estimates of population characteristics, such as subgroup averages and percentages of students at or above a certain proficiency level. Instead, NAEP constructs sets of plausible values designed to represent the distribution of proficiency in the population. A plausible value for an individual is not a scale score for that individual but may be regarded as a representative value from the distribution of potential scale scores for all students in the population with similar characteristics and identical patterns of item response. Statistics describing performance on the NAEP scales are based on these plausible values. These statistics estimate values that would have been obtained had individual proficiencies been observed; that is, had each student responded to a sufficient number of cognitive questions so that his or her proficiency could be precisely estimated.[6]

For the 1997 music, theatre, and visual arts Responding exercises, separate IRT scales were constructed. Because of the separate scaling, each of the three arts assessments employed slightly different steps in data analysis and IRT scaling. The steps for each subject area are described in detail in the forthcoming *NAEP 1997 Arts Analysis Technical Report*. Because of concerns about multidimensionality, and because there were too few items to create IRT scales for Creating, Performing, or Creating/Performing, only items in the Responding categories of the arts framework were entered into the IRT scaling procedure. A single IRT scale was created for each of the three fields of art. (In music, it was necessary to fit two related scales and then combine them into a composite. The two interim music subscales differed in the dependence of their tasks on music notation and technical vocabulary.) Creating items in visual arts and music, Creating/Performing items in theatre, and Performing items in music were formed into separate percent-of-total-possible-points averages, with mean percent correct scores reported at various levels. Certain theatre items which combined aspects of Responding and Creating (e.g., "draw a set design for this play") did not fit the Responding IRT scale[7], and so were not included in that scale, but were reported on an item-level basis. All of the Creating and Performing tasks from the two music blocks that were administered to a subsample of students in the music assessment were also reported on an item-level basis. In addition, items with logical dependencies (e.g., write a new ending to a script, followed by a discussion at how that ending accomplished one's goals) were separated, with the discussions included in the IRT scale. The endings on which they depended were put in the Creating/Performing averages. This appeared to introduce no worse a violation of the local independence assumption, which underlies IRT models, than does presenting a series of questions based on a single reading passage, and the IRT scaling led to acceptable fit to the models. (Local independence assumes that, given an examinee's ability, his or her responses depend on no other dimension. If this requirement is not met, fitting a unidimensional scale to the data results in some degree of error due to model misfit.)

For each of the three subjects in the arts assessment, a Responding IRT scale with a mean of 150 and a standard deviation of 35 was created, so that the great majority of

6 For theoretical and empirical justification of the procedures employed, see Mislevy, R.J. (1988). "Randomization-based inferences about latent variables from complex samples." *Psychometrika*, 56(2), 177-196.

For computational details, see National Assessment of Educational Progress. (1990). *Focusing the New Design: NAEP 1988 Technical Report and the 1990 NAEP Technical Report.* (Princeton, NJ: Educational Testing Service).

7 Item fit was assessed by assessing fit statistics and by visual inspection of empirical vs. model-based plots.

students had scores between 45 and 255, using a generalized partial-credit (GPC) model.[8] Developed by ETS and first used in 1992, the generalized partial-credit model permits the scaling of tasks scored according to multi-point rating schemes. The model takes full advantage of the information available from each of the student response categories used for these more complex performance tasks. It also graphically identifies pairs of adjacent scoring categories on which raters did not make reliable distinctions, allowing the analyst to collapse such categories and yield better fit to the model, and more importantly, empirically justify final scoring rubrics. Although the mean of the Responding scale for each subject has been set to 150, the scales are measuring different accomplishments. Comparisons cannot be made between student results on any pair of Responding scales, even though the scales share the same mean (150). In other words, a score of 165 in visual arts is not necessarily "better" than a score of 160 in music.

It is useful to illustrate the level of performance of students with a given scale score in a given subject by identifying questions likely to be answered correctly by students with that scale score. This process is known as "mapping." The position of a question on the Responding scale for each arts area represents the scale score attained by students who had: (1) at least a 65 percent probability of reaching a given score level on a constructed-response question, or (2) at least a 74 percent probability of correctly answering a multiple-choice question.

Readers should note that the percentages of 65 and 74 are a measurement convention adopted by NAEP. Other surveys have used 80 percent and 50 percent as measurement conventions. The probability conventions of 65 percent and 74 percent were established in 1994 for the NAEP assessments based, in part, on an intuitive judgement that they would provide a reasonable picture of students' knowledge and skills. Sixty-five and 74 percent represent a compromise between 50 percent (which was rejected on the grounds that having a 50/50 chance of getting the exercises right shows an insufficient degree of mastery) or 80 percent (which was rejected because students below that percentage still show some level of ability that would be ignored by such a stringent criterion).

Cautions in Interpretations

As described earlier, the NAEP scales for all the subjects make it possible to examine relationships between students' achievement and a variety of background factors measured by NAEP. The fact that a relationship exists between achievement and another variable, however, does not reveal the underlying cause of the relationship, which may be influenced by a number of other variables. Similarly, the arts assessments do not capture the influence of unmeasured variables. The results are most useful when they are considered in combination with other information about the student population and the educational system, such as emphases in instruction, changes in the school-age population, and societal demands and expectations placed on learning and performance.

NAEP Reporting Groups

In this report, results are provided for groups of students defined by shared characteristics — region of the country, gender, race/ethnicity, and parental education. Results are reported for subpopulations only when sufficient numbers of students are assessed and adequate school representation criteria are met. For public school students, the minimum requirement is 62 students in a particular subgroup from at least 5 primary sampling units (PSUs).[9] For nonpublic school students, the minimum requirement is 62 students from at least 5 PSUs. The data for all students, regardless of whether their subgroup was reported separately, were included in computing overall results. Definitions of the subpopulations referred to in this report follow.

Region

Results are reported for four regions of the nation: Northeast, Southeast, Central, and West. Figure A.2 shows how states are subdivided into these regions. All 50 states and the District of Columbia are listed. Territories and the two Department of Defense Educational Activities jurisdictions are not assigned to any region.

8 Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

9 For the national assessment, a PSU is a selected geographic region (a county, group of counties, or a metropolitan statistical area).

Figure A.2 Regions of the Country

THE NATION'S REPORT CARD NAEP

Northeast	Southeast	Central	West
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
District of Columbia	Florida	Iowa	California
Maine	Georgia	Kansas	Colorado
Maryland	Kentucky	Michigan	Hawaii
Massachusetts	Louisiana	Minnesota	Idaho
New Hampshire	Mississippi	Missouri	Montana
New Jersey	North Carolina	Nebraska	Nevada
New York	South Carolina	North Dakota	New Mexico
Pennsylvania	Tennessee	Ohio	Oklahoma
Rhode Island	Virginia*	South Dakota	Oregon
Vermont	West Virginia	Wisconsin	Texas
Virginia*			Utah
			Washington

*NOTE: The part of Virginia that is included in the Washington, DC metropolitan area is included in the Northeast region; the remainder of the state is included in the Southeast region.
SOURCE: National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1997 Arts Assessment.

Gender

Results are reported separately for males and females.

Race/Ethnicity

The race/ethnicity variable is derived from two questions asked of students and, where necessary, school records, and it is used to compare the performance of race/ethnicity subgroups. Two questions from the set of general student background questions were used to determine race/ethnicity.

If you are Hispanic, what is your Hispanic background?

- (A) I am not Hispanic.
- (B) Mexican, Mexican American, or Chicano
- (C) Puerto Rican
- (D) Cuban
- (E) Other Spanish or Hispanic background

For the question above, students who responded by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the question, or provided information that could not be classified, responses to the following question were examined to determine their race/ethnicity.

Which best describes you?

- (A) White (not Hispanic)
- (B) Black (not Hispanic)
- (C) Hispanic ("Hispanic" means someone who is from a Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or other Spanish or Hispanic background.)
- (D) Asian ("Asian" means someone who is from a Chinese, Japanese, Vietnamese, or other Asian background.)
- (E) Pacific Islander ("Pacific Islander" means someone who is from a Filipino, Hawaiian, or other Pacific Island background.)
- (F) American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)
- (G) Other

Students' race/ethnicity was then assigned on the basis of their responses. For the question above, students who filled in the seventh oval ("Other"), or who did not respond at all, race/ethnicity was assigned as determined by school records.

Race/ethnicity could not be determined for students who did not respond to either of the demographic questions and whose schools did not provide information about race/ethnicity.

Details of how race/ethnicity classifications were derived are presented so that readers can determine how useful the results are for their particular purposes. Also, some students indicated that they were from a Hispanic background (e.g., Puerto Rican or Cuban) and that a racial/ethnic category other than Hispanic best described them. These students were classified as Hispanic based on the rules described above.

Parents' Highest Level of Education

The variable representing the level of parental education is derived from responses to two questions from the set of general student background questions. Students were asked to indicate the extent of their mother's education.

How much education did your mother receive? ("Mother" can be a mother, stepmother, or female guardian.)

- (A) She did not finish high school.
- (B) She graduated from high school.
- (C) She had some education after high school.
- (D) She graduated from college.

Students were asked a similar question about their father's education level.

How much education did your father receive? ("Father" can be a father, stepfather, or male guardian.)

- (A) He did not finish high school.
- (B) He graduated from high school.
- (C) He had some education after high school.
- (D) He graduated from college.

The information was combined into one parental education reporting variable through the following process. If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. If the student did not respond for either parent, the student was recorded as having provided no response.

Guidelines for Analysis and Reporting

This report describes performance of eighth graders and examines the results for various groups of students within this population (e.g., those who have certain demographic characteristics or who responded to a specific background question in a particular way). It also examines the results for individual demographic groups and individual background questions. However, it does not include an analysis of the relationships among combinations of these subpopulations or background questions.

Estimating Variability

The statistics presented in this report are estimates of group and subgroup performance based on samples of students, rather than the values that could be calculated if every student in the nation answered every assessment question. It is therefore important to have measures of the degree of uncertainty of the estimates. Accordingly, in addition to providing estimates of percentages of students and their average scale score, this report provides information about the uncertainty of each statistic.

Two components of uncertainty are accounted for in the variability of statistics based on IRT scale or mean percent correct scores: the uncertainty due to sampling only a small number of students relative to the whole population and the uncertainty due to sampling only a relatively small number of questions. The variability of estimates of percentages of students having certain background characteristics or answering a certain cognitive question correctly is accounted for by the first component alone. Because NAEP uses complex sampling procedures, conventional formulas for estimating sampling variability that assume simple random sampling are inappropriate. For this reason, NAEP uses a jackknife replication procedure to estimate standard errors. The jackknife standard error provides a reasonable measure of uncertainty for any information about students that can be observed without error, but each student typically responds to so few questions within any content area that the scale score for any single student would be imprecise. In this case, using plausible values technology makes it possible to describe the performance of groups and subgroups of students, but the underlying imprecision that makes this step necessary adds an additional component of variability to statistics based on NAEP scale scores.[10]

The reader is reminded that, like those from all surveys, NAEP results are also subject to other kinds of errors including the effects of necessarily imperfect adjustments for student and school nonresponse and other largely unknowable effects associated with the particular instrumentation and data collection methods used. Nonsampling errors can be attributed to a number of sources: inability to obtain complete information about all selected students in all selected schools in the sample (some students or schools refused to participate; some students participated but answered only certain questions); ambiguous definitions; differences in interpreting questions; inability or unwillingness to give correct information; mistakes in recording, coding, or scoring data; and other errors of collecting, processing, sampling, and estimating missing data. The extent of nonsampling errors is difficult to estimate. By their nature, the impacts of nonsampling errors cannot be reflected in the data-based estimates of uncertainty provided in NAEP reports.

Drawing Inferences from the Results

The use of confidence intervals, based on the standard errors, provides a way to make inferences about the population averages and percentages in a manner that reflects the uncertainty associated with the sample estimates. An estimated sample scale score average ± 2 standard errors represents about a 95-percent confidence interval for the corresponding population

10 For further details, see Johnson, E.G. & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, 17(2), 175-190.

quantity. This means that with 95-percent certainty, the average performance of the entire population of interest is within about ± 2 standard errors of the sample average.

As an example, suppose that the average theatre Responding scale score of students in a particular group was 256, with a standard error of 1.2. A 95-percent confidence interval for the population quantity would be as follows:

Average ± 2 standard errors
= 256 ± 2 (1.2)
= 256 ± 2.4
= (256 - 2.4, 256 + 2.4)
= (253.6, 258.4)

Thus, one can conclude with close to 95-percent certainty that the average scale score for the entire population of students in that group is between 253.6 and 258.4.

Similar confidence intervals can be constructed for percentages, provided that the percentages are not extremely large or extremely small. For percentages, confidence intervals constructed in the above manner work best when sample sizes are large, and the percentages being tested have magnitude relatively close to 50 percent. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size and/or if "extreme" percentages are being compared. For reporting purposes, groups' percentages, P, were treated as "extreme" if:

$$P < P_{lim} = \frac{200}{N_{EFF} + 2},$$

where the effective sample size,

$$N_{EFF} = \frac{P(100 - P)}{(SE)^2},$$

and SE is the jackknife standard error of P. This "rule of thumb" cutoff leads to flagging a large proportion of confidence intervals that would otherwise include values < 0 or > 1. Similarly, at the other end of the 0 - 100 scale, a percentage is deemed extreme if (100 - P) < P_{lim}. In either extreme case, the confidence intervals described above are not appropriate, and procedures for obtaining accurate confidence intervals are quite complicated. In this case, the value of P was reported, but no standard error was estimated and hence no tests were conducted. This rule was not applied to response percentages within items.

As for percentages, confidence intervals for average scale scores are most accurate when sample sizes are large. For some of the subgroups of students for which average scale scores or percentages were reported, student sample sizes could be quite small. For results to be reported for any subgroup, a minimum student sample size of 62 was required. If students in a particular subgroup were clustered within a small number of geographic primary sampling units (PSUs), the estimates of the standard errors might also be inaccurate. So, subgroup data were required to come from a minimum of five PSUs.

Analyzing Group Differences in Averages and Percentages

To determine whether there is a real difference between the average scale score (or percentage of a certain attribute) for two groups in the population, one needs to obtain an estimate of the degree of uncertainty associated with the difference between the average scale scores or percentages of these groups for the sample. This estimate of the degree of uncertainty, called the standard error of the difference between the groups, is obtained by squaring each group's standard error, summing these squared standard errors, and then taking the square root of this sum.

If zero is within the confidence interval for the differences there is no statistically significant difference between the groups. The multiplier is the .975(1-.025) percentile from a T-distribution with the degrees of freedom that vary by the values of the average scale scores, their standard errors, and the number of PSUs that contribute to the average scale

scores. (See the forthcoming *1997 NAEP Arts Analysis Technical Report* for more details.) This procedure produces a conservative estimate of the standard error of the difference, since the estimates of the group averages or percentages will be positively correlated to an unknown extent due to the sampling plan. Direct estimation of the standard errors of all reported differences would involve a heavy computational burden.

Sometimes a group of related comparisons are made, such as comparing the average scale scores for a particular region with those from another region for specific groups of students. If one wants to hold the certainty level for a specific set of comparisons at a particular level (e.g., 95-percent), adjustments (called multiple-comparisons procedures) need to be made. One such procedure, the Bonferroni method, was used to form confidence intervals for the differences for sets of comparisons. The set of comparisons is referred to as a "family," and the typical family involves all subgroups related by a certain background question.[11]

Multiple-comparisons procedures, like the Bonferroni method, are useful for controlling the overall Type I error rate for a defined set of hypothesis tests. However, especially when the number of potential comparisons which could be made is large, as in NAEP data, this protection comes at the substantial loss of power in detecting specific consistent patterns in the data. For example, more powerful and complex tests of significance designed to identify consistent patterns in the data might judge that two groups were significantly different when a Bonferroni multiple-comparisons procedure would not.

These tests do not control the overall Type I error rate when they are applied to several related subgroups, such as the students in each region of the country. For this reason, the Bonferroni method for controlling Type I error was used when related subgroups were tested. For example, when tests were conducted for means or percentages for the separate race/ethnicity groups (e.g., White, Black, and Hispanic) these tests were treated as a single family of comparisons of size 3. The significance level for each of the separate tests was adjusted by the Bonferroni procedure to yield a Type I error rate of .05 for the family of comparisons.

11 Miller, R.G. (1996). *Simultaneous statistical inference*. New York: Wiley.