DOCUMENT RESUME

ED 409 356		TM 026 879
AUTHOR	Yepes-Baraya, Mario	
TITLE	Lessons Learned from the Coding of Item Attributes for the 1996 National Assessment of Educational Progress (NAEP) Science Assessment: Grade 4 Results.	
INSTITUTION	Educational Testing Service, Princeton, NJ.	
SPONS AGENCY	National Assessment of Educational Progress, Princeton, NJ.	
PUB DATE	Mar 97	
NOTE	34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 252-27, 1997).	
CONTRACT	R999J50001	
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC02 Plus Postage.	
DESCRIPTORS	*Coding; *Elementary School Students; Grade 4; Intermediate Grades; Performance Based Assessment; *Problem Solving; *Science Tests; Test Content; Test Format; *Test Items; Thinking Skills	
IDENTIFIERS	Experts; *National Assessment of Educational Progress	

ABSTRACT

The study described in this paper is part of an effort to improve understanding of the science assessment of the National Assessment of Educational Progress (NAEP). It involved the coding of all the items in the 1996 NAEP science assessments, which included 45 blocks (15 each for grades 4, 8, and 12) and over 500 items. Each of the approximately 2,500 students participating in the assessment was given a test booklet with 3 blocks of cognitive items. One was a conceptual/problem solving block, one, a theme block, and the last, a block of items associated with a performance task. Coding the item attributes provides descriptive information for each item, each block, and the whole test. The focus of this paper is on the grade-4 blocks. Nine science experts (two NAEP experts and a science teacher for each grade level) coded the attributes in the assessment according to categories such as knowledge of principles and reasoning with content. In all, 39 attributes were assessed. Results from the coding and block analyses suggest that, overall, the 1996 NAEP science assessment is a balanced assessment with respect to the science fields involved and item format used. Reasoning with content and explanation were the most significant attributes assessed; they were found to be key to successful performance on all three types of item blocks. (Contains 2 figures, 7 tables, and 12 references.) (SLD)

Lessons Learned from the Coding of Item Attributes for the 1996 National Assessment of Educational Progress (NAEP) Science Assessment: Grade 4 Results

Mario Yepes-Baraya

Educational Testing Service
Princeton, NJ 08541

Paper presented as part of the Symposium *What Does the NAEP Science Assessment Measure?: Results from the 1996 Science Attribute Study* at the annual meeting of the National Council on Measurement in Education, Chicago, IL, March 1997.

The work reported herein was partially supported under the National Assessment of Educational Progress (Grant No. R999J50001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Author's address: Mario Yepes-Baraya, Ph.D (8-R), Educational Testing Service, Princeton, NJ 08541. Telephone: (609) 734-5357. E-mail: myepes@ets.org

Lessons Learned from the Coding of Item Attributes for the 1996 National Assessment of Educational Progress (NAEP) Science Assessment: Grade 4 Results

Mario Yepes-Baraya
Educational Testing Service, Princeton, NJ 08541

> *We--teachers, students, researchers--are working toward the development of classroom communities in which students appropriate the discourse of science: a set of socio-historically constituted practices for constructing facts, for integrating facts into explanations, for defending and challenging claims, for interpreting evidence, for using and developing models, for transforming observations into findings, and for arguing theories. From this perspective, learning in science cannot be reduced simply to the assimilation of scientific facts, the mastery of scientific process skills, the refinement of a mental model, or the correction of misconceptions. Rather learning in science is conceptualized as the appropriation of a particular way of making sense of the world, of conceptualizing, evaluating, and representing the world.* (Warren and Rosebery, 1996)

Overview and Purpose of the Study

The study described in this paper is part of a research program to improve our understanding of the NAEP science assessment and what it measures. This initiative is important because for the first time the assessment includes a variety of innovative item types and tasks designed to study students' higher-order thinking skills. Results from previous studies involving item attributes with data from the 1993 NAEP science field test have already been reported (Park & Allen, 1994; Yepes-Baraya & Allen, 1994; Allen, Park, Liang, & Thayer, April 1995; Yepes-Baraya, 1996).

While the previous studies were based on the coding of 90 items comprising six Grade 8 blocks from the 1993 NAEP science field test, the present study involved the coding of all the items in the 1996 NAEP science assessment, which included a total of 45 blocks (15 each for Grades 4, 8, and 12) and over 500 items. The specific objectives for the study were threefold: 1) to characterize items, item blocks, and the assessment as a whole in terms of a set of attributes identified in previous studies; 2) to determine coding reliabilities for the item attributes, item blocks, and item block types in the assessment; and 3) to identify questions for further research to improve the NAEP and other science assessments.

Each of the approximately 2500 students participating in the assessment was given a test booklet that included three blocks of cognitive items: 1) a conceptual/problem solving block, similar to the standard blocks of previous NAEP science assessments but containing a larger proportion of constructed-response items; 2) a theme block, in which all the items are associated with a given theme, e.g., a pond ecosystem, a model of the solar system, or the water cycle; and 3) a block of items associated with a performance task.

Each item in the assessment is characterized by the presence or absence of 39 attributes. An abbreviated description of the attributes is provided in Figure 1 (p. 18). The attributes have been classified into six categories: 1) content knowledge, 2) reasoning with content and explaining, 3) hypothesis formulation and testing, 4) processing figural information, 5) item format and reading difficulty, 6) and process skills for hands-on tasks. Content knowledge pertains to items for which certain types of knowledge (e.g., knowledge of facts or concepts, or knowledge derived from practical experience) can be used to answer an item. Reasoning with content and explaining refers to items requiring some form of deductive or inductive reasoning involving science content. Items in the third category require the formulation or testing of a hypothesis. Processing figural information describes items requiring the processing of information contained in a table, graph or figure, or the provision of a figural response. Item format and reading difficulty groups items with sentence structures and format characteristics that might facilitate or hinder answering the item correctly. Process skills for hands-on tasks refers to items requiring manipulation of equipment or materials, making observations or measurements, and other science process skills. Attribute 39 is an additional measure of reading difficulty: For each item in the assessment, the teachers doing the coding were asked to determine if the average student in their classes would find the reading level of the item difficult.

The coding of item attributes provides descriptive information for each item, each block of items, and for the assessment as a whole. The focus of analysis for this report, however, is on the Grade 4 blocks of items. The information presented in this report has implications for researchers and practitioners interested in understanding the types of knowledge, skills, and information processing required by the 1996 NAEP science assessment. Science educators, in particular the advocates of constructivist perspectives, will be able to identify elements in the assessment that can be linked to activities, investigations, or projects in which students and teachers learn together through dialogue and reasoning to make sense of the world.

The coding of item attributes is not an end in itself. A completed coding sheet is called an incidence matrix (of items by attributes) and serves as the basis for the application of Tatsuoka's rule space model (Tatsuoka, & Tatsuoka, 1989; Tatsuoka, 1983). The rule space model is a probabilistic approach to identifying patterns of examinee responses which can be used in conjunction with Item Response Theory to identify attributes that an examinee or groups of examinees have mastered at a specified probability level. The information thus

Methodology

A team of nine science experts, three for each of grades 4, 8, and 12, was responsible for individually coding the attributes in the assessment. For each grade level, the item-coding team consisted of two ETS-NAEP science experts and one science teacher. The teachers were selected on the basis of their science teaching experience and familiarity with the science curriculum taught in New Jersey public schools, as well as their interest in innovative forms of science assessment. The project team attended a one-day training session to learn about the purpose of the study, the types of items in the assessment, and the 39 item attributes previously identified. One-half of the session was devoted to discussing and becoming familiar with each attribute and the other half to practice coding with one of the blocks in the assessment. The actual coding of all the blocks in the assessment took place over a six-week period in the summer of 1996. After completing the coding of the item attributes, the science teachers in the team were asked to answer in writing a number of questions about their experience with the project. The questions are included in Appendix 1.

The completed coding sheets were checked to identify discrepancies between coders. Each discrepancy was carefully examined and satisfactorily resolved by at least two members of the team. An official coding sheet was then completed for each block of items in the assessment and used later for the creation of the computer data files. Figure 2 (p. 19) shows the coding sheet for one of the Grade 4 conceptual/problem solving blocks.

Preliminary Results

The actual coding of the blocks of items, the discrepancies between coders, and the answers to the teacher questionnaire, each provide information that is useful to characterize and better understand the blocks of items and the knowledge and skills they were designed to assess. The results presented below are preliminary and are based on the coding of the fifteen Grade 4 blocks only, although somewhat similar results would be expected for the fifteen Grade 8 and the fifteen Grade 12 blocks.

Coding of the Blocks and Comparisons Between Blocks

The coding of the items in a block results in an incidence matrix such as the one shown in Figure 2. The incidence matrix for that block shows a certain pattern of 1's and 0's (zeros have been left blank) that may be similar or different from the patterns of other Grade 4 conceptual/problem solving blocks and other types of blocks. An obvious question is, What are the similarities and differences between the incidence matrices of blocks of the same type? Table 1 (p, 22) shows a comparison of the four Grade 4 performance tasks for some of the attributes being considered in the study. Table 4 (p. 25) summarizes the information in Table 1. The overall pattern that emerges from the comparisons suggests that *performance tasks* in the assessment can be characterized as follows:

- largely homogeneous content with respect to one of the three fields of science in the assessment: physical, earth, and life science (i.e., all the items in a given block are likely to belong to only one of these fields);
- a preponderance of constructed response items, with two of the tasks having only these type of items;

- little reliance on factual knowledge, or understanding of science vocabulary;
- moderate emphasis on knowledge of concepts and/or principles, and knowledge from practical experience;
- large emphasis on knowledge of experimental procedures, and reasoning with content;
- a preponderance of items with figural information, but a moderate reliance on figural information to obtain the correct answer;
- a moderate reading load; and
- heavy emphasis on manipulating equipment/materials, recording data, and/or interpreting data.

A similar comparison was made for the three Grade 4 theme blocks in the assessment. Table 2 (p. 23) shows a comparison of the three Grade 4 theme blocks for the same set attributes used in the comparison of the performance tasks. Table 4 summarizes the information in Table 2. The overall pattern that emerges from the comparisons suggests that *theme blocks* in the assessment can be characterized as follows:

- largely homogeneous content with respect to one of the three fields of science in the assessment: physical, earth, and life science (i.e., all the items in a given block are likely to belong to only one of these fields);
- a preponderance of constructed response items;
- no reliance on knowledge of experimental procedures;
- moderate emphasis on knowledge from practical experience;
- large emphasis on factual knowledge, knowledge of concepts or principles, understanding of science vocabulary, and reasoning with content;

- a preponderance of items with figural information, but a moderate reliance on figural information to obtain the correct answer;
- a light to moderate reading load; and
- no reliance on manipulating equipment/materials, recording data, and/or interpreting data.

As shown in Table 3 (p. 24), a similar comparison was made for four of the eight Grade 4 conceptual/problem solving blocks in the assessment for the same set of the attributes used in the two previous comparisons. Table 4 summarizes the information in Table 3. The overall pattern that emerges from the comparisons suggests the following characterization for the *conceptual/problem solving blocks* in the assessment:

- heterogeneous content with respect to the three fields of science in the assessment: physical, earth, and life science (i.e., a given block is likely to have items from all three fields);
- a slight preponderance of multiple-choice items;
- little reliance on knowledge of experimental procedures;
- moderate emphasis on knowledge from practical experience;
- large emphasis on factual knowledge, knowledge of concepts or principles, understanding of science vocabulary, and reasoning with content;
- moderate emphasis on items with figural information, and moderate reliance on figural information to obtain the correct answer;
- a relatively light reading load; and
- no reliance on manipulating equipment/materials, recording data, and/or interpreting data.

Coding Reliability

As shown in Table 6 (p. 27), the reliability for the coding of the attributes ranged from 88% for the performance tasks, to 90% for the theme blocks, to 92% for the conceptual/problem solving blocks. These high overall reliabilities hide lower reliabilities for specific attributes, as shown in Table 7 (p. 28). There are six attributes with reliabilities under 85% common to the three different types of item blocks. These attributes are Attributes Nos. 4, 5, 6, 7, 20, and 30: *knowledge of principles, understanding science vocabulary, knowledge from practical experience, figural response, and intratext referentials*. Of these, *reasoning with content* is among the three attributes with the lowest coding reliability for each of the three item block types, and *knowledge of principles* and *intratext referentials* are among the three attributes with the lowest coding reliability for two of the item block types. Possible reasons for these relatively low reliabilities will be discussed in light of the data presented in Tables 1-5 and teachers' responses to the open-ended questions, reported below.

Teachers' Responses

The four participating teachers were asked to answer a number of questions about their attribute coding experience (see Appendix 1). A summary of the responses provided is presented below.

1. Attributes that are difficult to code. Three of the four respondents found Attribute 4, *knowledge of principles*, and 13, *generating a hypothesis*, difficult to code. One Grade 4 teacher said: "At times it was difficult to decide if a principle was involved, what the principle was if there was a principle, or was there a simple fourth grade level principle rather than a

more involved principle." The only other attributes identified as difficult to code were Attribute 6, *information from practical experience* (mentioned by one respondent), and Attribute 9, *inductive reasoning* (mentioned by a different respondent).

2. Items or blocks that are difficult to code. One of the respondents found the performance tasks difficult to code: "I felt that the four hands-on blocks were harder to code. The block pertaining to the observation and classification of seeds took me over two hours to sort through. I wanted to be sure that I was seeing exactly what the student was supposed to be doing." Two other respondents identified specific item types as difficult to code. One respondent mentioned constructed-response items with complex scoring rubrics, and the other one graphing items.

3. Additional item attributes. Only one of the respondents provided additional item attributes like "identify and interpret an equation," or "identify and describe a scientific process." Due to their specific nature, those attributes appear to be more appropriate for a classroom assessment than for a large-scale assessment like NAEP.

4. Quality of pool of items. Thee respondents agreed that the quality of the items in the assessment is high. One said: "The quality seemed rather high. Most of the items required reasoning along with factual knowledge. The way the items were written ... and the scoring system made them reasonable to correct and decide if the student knew what was asked. The subjects covered were varied and appropriate..." Another respondent stated: "I feel the quality of the pool was good. Many of the items had to be thought through or were hands-on. As an eighth grade science teacher, I do a lot of hands-on work. It is wonderful to see that students are tested on lab work and the scientific method." A third respondent said: "The questions are extremely well written. All questions were very clear. The concept

or fact being tested was clear. Questions with multiple tasks were stated in clear, simple terms. Effort was made not to confuse students, nor to 'trick' them..."

5. Lessons learned from the coding experience and impact on teaching/assessment. Participants stated that they found the attribute coding experience valuable and, in some cases, were able to identify direct implications for teaching/assessment. Responses from the Grade 4 teachers are presented below.

> Respondent 1: "This was a worthwhile experience as I learned a great deal. A) Hopefully I will remember the format for the questions so I can use it when creating assessments (not just science related). There were some really good ways of ascertaining understandings, concepts, etc. I feel I gained some tools for more effective assessment. B) Presently the majority of fourth grade students in my school would not fare well on this test. I do think that in a few years the fourth graders will do better because of a revised curriculum and the teachers are receiving training and support in science...E) I think the people at ETS are working hard and carefully to develop methods of assessing students that are as fair and thorough as possible."

> Respondent 2: "This was a very valuable activity for me. I personally learned that the structuring of a question is very important in assessing what kind of information you can get from that question. Looking back on some of the activities that I did in the classroom last year, I can now see many things that need to be revisited so as to get the most out of them..."

Discussion and Conclusions

As stated earlier, the study described in this paper is part of a research program to improve our understanding of the NAEP science assessment and what it measures. Development of the item attributes presented in this report began in 1992. Over the years, the attributes have been tested and refined: some attributes have been added, others have been deleted, and others yet have been modified. Most of the attributes in the current list were part of a validation study that included think-aloud tasks with middle school science students (Yepes-Baraya, 1996). Although the current list includes attributes that are general enough to encompass all NAEP science assessment items, it may not be fully applicable to assessment situations different from NAEP.

The coding of item attributes provided descriptive information to characterize any given item, item blocks, and the assessment as a whole in terms of the knowledge, skills, and information processing required to demonstrate science proficiency. In this section, we will describe the characterization that emerges from the coding and explore some of the implications for further research on test development in science, instruction, and assessment.

Characterization of the blocks and the overall assessment

As summarized in Table 5 (p. 26), performance on the science tasks is largely a function of *knowledge and application of experimental procedures* (Attributes 2, and 35-37), as would be expected, *reasoning with content* and *explaining* (Attributes 7 and 12), and *processing figural information* attributes. These latter attributes play an important role in task performance because tasks usually have a chart or table that needs to be completed by the examinee with observations or measurements. This chart or table is a focal point for other items involving data interpretation and/or explanations. To a lesser extent, task performance is a function of *knowledge of concepts/principles* (Attributes 3 and 4) and *knowledge from practical experience* (Attribute 6). Additionally, task performance tends not to be a function of *factual knowledge* (Attribute 1) or *understanding of science vocabulary* (Attribute 5). This lessened importance of content knowledge in task performance is important because tasks were designed to assess primarily science process skills.

In contrast to performance on the tasks, performance on the theme blocks and the conceptual/ problem solving blocks is not a function of *knowledge or application of experimental procedures*. It is largely a function of *knowledge of facts, concepts/principles, and science vocabulary*, as well as *reasoning with content* and *explaining*.

The above observations and the information summarized in Table 5 suggest that, overall, the 1996 NAEP science assessment is a balanced assessment with respect to the science fields involved and item format used, with multiple-choice items being more prevalent in conceptual/problem solving blocks. Perhaps the most significant attributes in our list are *reasoning with content* and *explaining*: they are key to successful performance on all three types of item blocks in the assessment. A significant aspect of the assessment as a whole is the relatively low to moderate *reading load*: an effort has been made to reduce and eliminate sources of construct-irrelevant variance (Messick, 1994). Also significant is the use of *figural information* in the assessment: while all three item block types include items with figural information, in many instances this information is not required to answer then item correctly. In other words, there is a built-in redundancy in many items and blocks that provides examinees with multiple paths to demonstrate science proficiency.

Some implications for further research in science test development, instruction and assessment

The results presented in this report are based entirely on the Grade 4 portion of the 1996 NAEP science assessment. Similar analyses should be conducted to determine whether analogous patterns hold true for the Grade 8 and Grade 12 components of the assessment. Another research question is the extent to which the set of attributes selected for the present study predict item difficulty once the regression and other statistical analyses are performed. The main reasons why these attributes were selected were their prevalence in a number of science frameworks examined in preparation for this research (Yepes-Baraya & Allen, 1994), as well as their prevalence in the NAEP science Grade 4 blocks relative to other less common attributes.

From an instructional perspective, the present study raises issues that are pertinent to the improvement of science instruction and assessment. One issue is the usefulness of the attributes in the professional development of teachers and in the classroom. The attributes in this study represent core cognitive elements that need to be mastered to demonstrate problem-solving ability in science (Yepes-Baraya, 1996; Sugrue, 1995). Relatively low coding reliabilities and teacher responses to the questionnaire suggest that teachers found Attributes Nos. 4 and 7, *knowledge of principles* and *reasoning with content* particularly difficult to code. The reason for this difficulty is not clear, but it might indicate an atomistic approach to science teaching: too much emphasis on factual knowledge at the expense of concepts, principles, and reasoning to identify significant connections. As one of the teachers stated, "At times it was difficult to decide if a principle was involved, what the principle was if there was a principle, or was there a simple fourth grade level principle rather than a more involved principle." Thus, the attributes might be used in the context of teacher preparation or teacher enhancement programs to help teachers align standards, curriculum and assessment, and to develop instruction that encourages reasoning, dialogue, and reflection. As another participating teacher said, "I use the upper levels of Bloom's Taxonomy with most of my objectives but I felt much more comfortable with the grouping of attribute clusters that we used in this survey. The way that these attributes are clustered, you can easily see if the assessment item meets the curriculum criteria."

Acknowledgements

The author gratefully acknowledges Christine O'Sullivan of ETS/NAEP for her continued support, Fiona Herr of ETS/NAEP for her valuable assistance and ideas in conducting this study, and Jean V. Yepes of the Assessment Division of ETS for her thoughtful comments during the preparation of this report. This effort has been supported by NAEP and ETS. Opinions expressed are those of the author and not necessarily ETS or NAEP.

References

Allen, N.L., Park, C., Liang, J., & Thayer, D. (April 1995). *Relationships between test specifications, task demands, and item attributes in a large-scale science assessment*. Paper presented as part of the symposium Large Scale Science Performance Assessment and Results: Informing Test and Score Development at the annual meetings of the AERA and NCME, San Francisco.

Messick, S. (December 1994). *Alternative modes of assessment, uniform standards of validity*. (Research Report RR-94-60). Princeton, NJ: Educational Testing Service.

Park, C. & Allen, N.L. (1994). *Relationships between test specifications, item responses, task demands, and item attributes in a large-scale science assessment*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Sugrue, B. (Fall 1995). A theory-based framework for assessing domain specific problem-solving ability. *Educational Measurement: Issues and Practices, 3*, 29-36.

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R.L. Glaser, A.M. Lesgold, & M.G. Shafto (Eds.) *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Erlbaum.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20* (4).

Tatsuoka, M.M., & Tatsuoka, K.K. (1989). Rule space. In Kotz & Johnson (Eds.) *Encyclopedia of statistical sciences*. New York: Wiley.

van Someren, M.W., Barnard, Y.F., & Sandberg, J.A.C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. San Diego: Academic Press.

Warren, B. & Rosebery, A.S. (1996). "This question is just too easy!" Students' perspectives of accountability in science. In L. Schauble & R. Glaser (Eds.) *Innovations in Learning* (pp. 97-125). Mahwah, NJ: Erlbaum.

Yepes-Baraya, M. (April 1996). *A cognitive study based on the National Assessment of Educational Progress (NAEP) science assessment*. Paper presented as part of the symposium School-Based Special Studies of a Large-Scale Performance Assessment at the annual meetings of the AERA and NCME, New York, NY.

Yepes-Baraya, M. (April 1995). *Task analysis of science performance tasks and items: Identifying relevant attributes*. Paper presented as part of the symposium Large Scale Science Performance Assessment and Results: Informing Test and Score Development at the annual meetings of the AERA and NCME, San Francisco, CA.

Yepes-Baraya, M. & Allen, N.L. (1994). *The process of identifying item attributes related to item performance for the 1993 National Assessment of Educational Progress (NAEP) science field test*. Princeton, NJ: Educational Testing Service.

Figure 1. Item Attributes

Content knowledge
1. Can knowledge of facts be used to answer the item?
2. Can knowledge of experimental procedures be used to answer the item?
3. Can knowledge of concepts be used to answer the item?
4. Can knowledge of principles be used to answer the item?
5. Does item have science vocabulary that must be understood to answer item?
6. Could the info. required to answer item have been gained through practical experience?

Reasoning and explaining

7. Can reasoning from general concept/principle/law to specific conclusion be used?
8. Can tracing cause-effect from one component to another in system needed to answer item?
9. Can formal inductive reasoning be used to answer item?
10. Does item require identifying or describing a procedure to solve a problem?
11. Can thinking with models/analogies be used to answer item?
12. Does item require that a response be given and the response be justified?

Hypothesis formulation and testing

13. Is generation of hypothesis/prediction necessary to answer item?
14. Does item require ident. of variables/controls in design of test for hypothesis?
15. Does item require generating operationalized procedures for testing a hypothesis?
16. Does item require use of multiple control groups in design of test for hypothesis?

Processing figural information

17. Does item have a TGF* already completed/needs to be completed?
18. Does item refer directly or indirectly to info. in a completed & separate TGF (g/s)?
19. Does item refer to info. in a tTGF* (s)* separate from stem?
20. Does item have (or refers to info. in) a completed TGF (g/s)*?
21. When present, is it possible to use info. in completed TGF (g/s) to answer item?
22. Is it necessary to use info. in completed TGF (g/s) to answer item?
23. Is some of the info. needed to answer item in TGF (s)?
24. Is all info. needed to answer item in tTGF in block with item? [All info. is (g)]
25. Is all info. needed to answer item in tTGF in block with item? [Some info. is (s)]
26. Does response require a TGF to be drawn or completed?
27. Does response require a GF to be drawn or completed?

Item format and reading difficulty

28. Is item a 5 or 4-category constructed-response item?
29. Is item a 3 or 2-category constructed-response item?
30. Does item stem have at least 1/2/3 intratext referentials (e.g., it, this, these)?
31. Does item stem have at least 1/2/3 clauses with fronted structures?
32. Must response meet all conditions specified in stem?
33. Does item have exceptions/negations that make item complex?
34. Can item be solved by choosing the odd option out?

Process skills for hands-on tasks

35. Does item require the manipulation of equipment/materials?
36. Does item require the recording of data (observations or measurements)?
37. Does item require interpreting data collected or making inferences from this data?
38. Does item require performing numerical calculations with data collected?

*TGF = table, graph, or figure (g) = given
tTGF = text, table, graph, or figure (s) = student-generated

Figure 2. Incidence Matrix for the Coding of A Grade 4 Conceptual/Problem Solving Block in the 1996 NAEP Science Assessment

		CONTENT KNOWLEDGE						REASONING WITH CONTENT						HYPO.FORM.& TEST.			
Attribute No.--> Item No.	& Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1		1				1										
2	2	1		1		1											
3	1	1		1	1		1										
4	2	1		1	1	1		1				1					
5-2	2	1		1	1		1	1					1				
5-1a	2	1		1	1		1	1					1				
5-1b	2	1		1	1		1	1					1				
6	2	1		1								1					
7	2	1		1								1					
8	2	1		1		1	1		1						1		
9-1	1		1	1							1		1		1		
10-2	3		1	1							1		1		1		
10-1	3		1	1							1				1		
11-2	3		1	1							1		1		1		
11-1	3		1	1							1				1		

Figure 2 (continued). Incidence Matrix for the Coding of A Grade 4 Conceptual/Problem Solving Block in the 1996 NAEP Science Assessment

		PROCESSING FIGURAL INFORMATION										
Attribute No.--> Item No.	& Scale	17	18	19	20	21	22	23	24	25	26	27
1	1	1			1	1	1		1			
2	2											
3	1	1			1	1	1					
4	2											
5-2	2	1			1	1	1					
5-1a	2	1			1	1	1					
5-1b	2	1			1	1	1					
6	2											
7	2											
8	2											
9-1	1	1			1	1	1					
10-2	3	1	1	1	1	1	1					
10-1	3	1	1	1	1	1	1					
11-2	3	1	1	1	1	1	1					
11-1	3	1	1	1	1	1	1				1	

Figure 2 (continued). Incidence Matrix for the Coding of A Grade 4 Conceptual/Problem Solving Block In the 1996 NAEP Science Assessment

		ITEM FORMAT AND READING DIFFICULTY							PROCESS SKILLS				*
Attribute No.--> Item No.	& Scale	28	29	30	31	32	33	34	35	36	37	38	39
1	1			1		1							
2	2					1							
3	1					1							
4	2		1	1		1							
5-2	2		1	1		1							
5-1a	2		1	1									
5-1b	2		1	1									
6	2					1							
7	2					1							
8	2			1	1	1							
9-1	1		1			1							
10-2	3		1			1							
10-1	3		1										
11-2	3		1			1							
11-1	3		1								1		

* As described on page 2, Attribute 39 is an additional measure of reading difficulty.

Table 1. Comparison of Four Grade 4 Performance Tasks in the 1996 NAEP Science Assessment

ATTRIBUTES	TASK 1	TASK 2	TASK 3	TASK 4
Science Field & No. of items	All 7 - Life	All 7 - Physical	7/11 - Phys. 4/11 - Earth	All 5 - Phys.
Item Format	All - CR	5/7 - MC	8/11 - CR	All - CR
1-Facts	None	None	None	All but 2
2-Experimental Procedures	None	All but 2	All but 1	Only 1
3-Concepts/ 4-Principles	All	3/7	5/11	3/5
5-Science Vocabulary	None	None	None	Only 1
6-Practical Experience	All	4/7	4/11	Only 1
7-Reasoning 12-Explaining	All	3/7	5/11	All
Figural Information	All but 1	All but 2	All	All
Figural-Necessary	All but 2	None	Only 2	Only 1
Reading load	Moderate to high	Moderate	Moderate to high	Moderate to high
35-Manipulation/ 36-Data Recording/ 37-Data Interpret.	All	All but 1	All but 1	All but 1

Table 2. Comparison of Three Grade 4 Theme Blocks in the 1996 NAEP Science Assessment

ATTRIBUTES	THEME 1	THEME 2	THEME 3
Science Field & No. of items	All 10 - Earth	All 8 - Life	All 9 - Life
Item Format	All but 2 - CR	All but 2 - CR	All but 2 - CR
1-Facts	All	All	All
2-Experimental Procedures	None	None	None
3-Concepts/ 4-Principles	All	All	All
5-Science Vocabulary	All	All but 1	All but 2
6-Practical Experience	None	All but 2	Only 3
7-Reasoning 12-Explaining	All	All	All but 2
Figural Information	All but 1	All	All
Figural-Necessary	All but 3	4/8	Only 2
Reading load	Moderate	Moderate	Moderate to high
35-Manipulation/ 36-Data Recording/ 37-Data Interpret.	None	None	None

Table 3. Comparison of Four Grade 4 Conceptual/Problem Solving Blocks in the 1996 NAEP Science Assessment

ATTRIBUTES	C/PS BLOCK 1	C/PS BLOCK 2	C/PS BLOCK 3	C/PS BLOCK 4
Science Field & No. of items	2 - Physical 5 - Earth 4 - Life	3 - Physical 7 - Earth 1 - Life	6 - Physical 1 - Earth 3 - Life	3 - Physical 4 - Earth 4 - Life
Item Format	6/11 - MC	6/11 - MC	6/10 - MC	6/11 - MC
1-Facts	All	All but 1	All but 1	All but 2
2-Experimental Procedures	None	Only 3	None	None
3-Concepts/ 4-Principles	All	All	All	All but 2
5-Science Vocabulary	All but 2	Only 3	All but 2	6/11
6-Practical Experience	6/11	All but 3	Only 1	4/11
7-Reasoning 12-Explaining	All but 1	6/11	All but 3	6/11
Figural Information	4/11	6/11	Only 2	4/11
Figural-Necessary	4/11	5/11	Only 2	4/11
Reading load	Low	Low	Moderate	Low
35-Manipulation/ 36-Data Recording/ 37-Data Interpret.	None	None	None	None

Table 4. Comparison of Three Different Types of Grade 4 Blocks of Items in the 1996 NAEP Science Assessment

ATTRIBUTES	TASKS (4 blocks)	THEMES (3 blocks)	C/PS (4 blocks)
Science Field & No. of items	3 - All - 1 type 1- 7/4 - P/E	3 - All - 1 type	4 - All - Mix P/E/L
Item Format	2 - All - CR 1 - 8/11 - CR 1 - 5/7 - MC	3 - All but 2 - CR	4 - Slightly over half - MC
1-Facts	3 - None 1 - All but 2	3 - All	1 - All 3 - All but 1
2-Experimental Procedures	2 - Almost all 1 - Only 1 1 - None	3 - None	3 - None 1 - Only 3
3-Concepts/ 4-Principles	1 - All 3 - @ half	3 - All	3 - All 1 - All but 2
5-Science Vocabulary	3 - None 1 - Only 1	1 - All 1 - All but 1 1 - All but 2	2 - All but 2 1 - @ half 1 - Only 1
6-Practical Experience	1 - All 2 - @ half 1 - Only 1	1 - All but 2 1 - Only 3 1 - None	1 - All but 3 2 - @ half 1 - Only 1
7-Reasoning 12-Explaining	2 - All 2 - @ half	2 - All 1 - All but 2	1 - All but 1 1 - All but 3 2 - @ half
Figural Information	2 - All 2 - Almost all	3 - All	3 - @ half 1 - Only 2
Figural-Necessary	1 - All 1 - All but 2 1 - Only 2 1 - None	1 - All but 3 1 - @ half 1 - Only 2	3 - @ half 1 - Only 2
Reading load	2 - Moderate to high 2 - Moderate	3 - Moderate	4 - Low to moderate
35-Manipulation/ 36-Data Recording/ 37-Data Interpret.	1 - All 3 - All but 1	3 - None	None

Table 5. Relative Prevalence* of Selected Attributes for Three Different Types of Grade 4 Blocks in the 1996 NAEP Science Assessment

ATTRIBUTES	TASKS (4 blocks)	THEMES (3 blocks)	C/PS (4 blocks)
Homogeneous content by science field	HI	HI	LO
Constructed-response items	HI	HI	MED
1-Facts	LO	HI	HI
2-Experimental Procedures	HI	NO	LO
3-Concepts/ 4-Principles	MED	HI	HI
5-Science Vocabulary	LO	HI	HI
6-Practical Experience	MED	MED	MED
7-Reasoning 12-Explaining	HI	HI	HI
Figural Information	HI	HI	MED
Figural-Necessary	MED	MED	MED
Reading load	MED	LO to MED	LO
35-Manipulation/ 36-Data Recording/ 37-Data Interpret.	HI	NO	NO

*HI refers to high prevalence, MED to moderate prevalence, LO to low prevalence, and NO to the absence of items with a given attribute.

Table 6. Coding Reliability for Three Different Types of Grade 4 Blocks in the 1996 NAEP Science Assessment

	TASKS	THEME BLOCKS	C/PS BLOCKS
Number of blocks	4	3	4
N* = Number of items in these blocks	66	63	72
C = Number of cells in matrix (N X 36**)	2376	2268	2592
D = Mean number of discrepancies per coder	293	217	215
Reliability = 100 X (D/C)	88%	90%	92%

*For constructed-response items, this number includes all item levels for partial credit

**Although there are 39 item attributes, only 36 were included in calculating the reliabilities. For logistics reasons, Attributes 18 and 19 were coded independently by ETS researchers, and Attribute 39 was coded only by the three participating teachers.

Table 7. Attributes with the Lowest* Coding Reliability for Three Different Types of Grade 4 Blocks in the 1996 NAEP Science Assessment

No.** ATTRIBUTES	TASKS	THEMES	C/PS BLOCKS
1 - Knowledge of facts	81		
4 - Knowledge of principles	81	62***	65***
5 - Understanding science vocabulary	66***	75	81
6 - Practical experience	83	83	80
7 - Reasoning with content	71***	66***	67***
9 - Inductive reasoning	83		
12 - Justification of response		83	
20 - Figural response	72	82	84
21 - Figural response	72	80	
22 - Figural response (necessary)	82	78	
23 - Figural response	71***		
30 - Intratext referentials	74	66***	79***
37 - Interpreting data collected	74		

*Reliabilities under 85% are reported.

**Attribute numbers are the same as the numbers on Figure 1.

***These are the three attributes with the lowest coding reliability for each of the three types of item blocks.

Appendix 1

SCIENCE ATTRIBUTE STUDY

<u>Questions for Item Raters</u>

On a separate sheet of paper, please provide complete answers to the following questions.

1. List all **attributes** that you found difficult to code (refer by number to the blue document, Guidelines for Coding Items) and explain the reason for the difficulty.

2. List all **blocks of items** or **specific items** that you found difficult to code (refer to the block identification number and item number, if necessary) and explain the reason for the difficulty.

3. Other than the attributes in the blue document (Guidelines for coding Items), can you think of **other attributes** that would explain the level of a difficulty of the pool of science assessment items that you worked with?

4. As a teacher of science, how would you judge the **quality** of the pool of science assessment items that you worked with? What is your **criteria** for making this judgement?

5. Please describe your experience as you coded the items in the blocks that you received. Did you learn anything of value? Will this task have an impact in your teaching and/or assessment of science?

6. Additional comments

NCME 1997 TM026879

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Lessons Learned from the Coding of Item Attributes for the 1996 National Assessment of Educational Progress (NAEP) Science Assessment: Grade 4 Results	
Author(s): Mario Yepes-Baraya	
Corporate Source: Educational Testing Service	Publication Date: March 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

[X] **Check here** — Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

[ ] **or here** — Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: [signature]	Position: Research Scientist
Printed Name: Mario Yepes-Baraya	Organization: Educational Testing Service
Address: Rosedale Road, Princeton, NJ 08541	Telephone Number: (609) 734-5357
	Date: March 28, 1997

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 24, 1997

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our process of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: NCME 1997/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an NCME chair or discussant, please save this form for future use.

ERIC® Clearinghouse on Assessment and Evaluation