DOCUMENT RESUME

ED 371 010 TM 021 641

AUTHOR	Park, Chung; Allen, Nancy L.
TITLE	Relationships between Test Specifications, Item Responses, Task Demands, and Item Attributes in a Large-Scale Science Assessment.
SPONS AGENCY	Educational Testing Service, Princeton, N.J.
PUB DATE	Apr 94
NOTE	40p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Communication (Thought Transfer); Comparative Analysis; Construct Validity; Content Validity; *Difficulty Level; Elementary Secondary Education; Grade 4; Grade 8; Grade 12; Hypothesis Testing; Knowledge Level; National Surveys; *Responses; *Science Tests; Test Construction; Test Format; *Test Items; Test Validity; Thinking Skills
IDENTIFIERS	*Large Scale Programs; *National Assessment of Educational Progress; Task Characteristics; Test Specifications

ABSTRACT

This study is part of continuing research into the meaning of future National Assessment of Educational Progress (NAEP) science scales. In this study, the test framework, as examined by NAEP's consensus process, and attributes of the items, identified by science experts, cognitive scientists, and measurement specialists, are examined. Preliminary information about item responses was available from the 1993 field test of NAEP science items, which involved 3,908 4th graders, 3,585 8th graders, and 3,041 12th graders. Five-hundred sixteen 8th graders were administered a particular test booklet that was examined in detail. Information from the other test booklets will eventually be compared with these results. Each test item and the attributes associated with it has been evaluated for attributes in the major categories of (1) specific knowledge; (2) item format and vocabulary; (3) reasoning; (4) hypothesis testing and design of the hypothesis test; (5) explanation; and (6) communication. Further research will explore the relationship of item attributes and test framework to actual student responses. Nine tables and five figures present analysis results. (Contains 2 references.) (SLD)

Relationships between Test Specifications, Item Responses, Task Demands, and Item Attributes in a Large-Scale Science Assessment

Chung Park

University of Massachusetts at Amherst

Nancy L. Allen

Educational Testing Service, Princeton, NJ

March 21, 1994 DRAFT

April 6, 1994

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, 1994. This work was supported by research allocation funds from the Educational Testing Service.

Abstract

This study is a part of continuing research into the meaning of future NAEP science scales. In this study we examine the test framework, as developed by NAEP's consensus process, and attributes of the items, identified by science experts, cognitive scientists, and measurement specialists. Preliminary information about item responses was available from the 1993 field test of NAEP science items. The examination of these three pieces of information is important because the next NAEP assessment of science will include a hands-on manipulative task component, as well as innovative theme-related items.

Relationships between Test Specifications, Item Responses, Task Demands, and Item Attributes in a Large-Scale Science Assessment

Recently, the development of performance standards and the push for authentic assessment have been highly publicized as ways to improve education. In order for performance standards and authentic assessment to have this result, it is important that we understand what the items in assessments measure. Understanding of the learning process, the subject-area of interest, test development and cognitive psychology can help us to evaluate what an assessment measures. These perspectives offer approaches that can lend credence to performance standards and validity to assessments that have components that might be considered authentic--or at least, performance-based.

Over many years, the National Assessment for Educational Progress (NAEP) has developed a process to create a framework and more specific test specifications for NAEP subject-area assessments. This process builds consensus among subject-area specialists, educators, and measurement experts. The final products are meant to reflect the most forward-looking concepts of learning in the specific subject area and to keep a reasonable amount of continuity across time to adequately measure trends in learning. Educators and test developers develop items to fit the test specifications and test developers select items for the final assessment using the guidelines in the specifications and framework. Provisional assessments are refined using information from a field test.

Recently, as part of this process, NAEP developed a science assessment with a performance-based component. The inclusion of a large proportion of time devoted to hands-on manipulative tasks has led to a concern that the most information about student performance be gleaned from that testing time. Scoring the items related to the performance tasks in a traditional way that considers only student responses to the items may not provide maximum information about student learning, because that type of scoring does not take into account what we know about the relationships between tasks and what attributes of the items influence student responses.

In order to prepare for an analysis of options in scoring a large-scale assessment with testing situations as varied as independent multiple-choice items to constructed-response items associated with manipulative science tasks, we have examined the test framework and specifications, the item responses from a field test of the items, and item attributes as identified by subject-trained test developers and cognitive scientists. Figure 1 contains the information available for this study. A fourth piece of information that would provide information about what is actually being measured is a protocol analysis. Although a protocol analysis was not used here, it would provide more information about item attributes than can be provided by subject-area and testing experts.

----------Insert Figure 1 about here-----------

The connection between the three pieces of information about items (the test framework, the item responses, and the item attributes) represents a new approach in practical measurement. In the past, experience with items have been compartmentalized into either an examination of item responses (item analysis and scoring) by psychometricians or an examination of the categorization of items into the framework of the test specifications as a part of the test development process. Attributes of the items that contribute to specific item responses have been examined primarily in a research environment, as opposed to test production or test interpretation settings.

In our examination of the relationships between the test framework, the item responses, and the item attributes, we focused on four sets of questions. They are:

1. How are the item attributes related to one another? How are the framework categories related to one another?
2. What is the connection between the item attributes and the test framework?
3. How do the item attributes and test framework relate to whether the item is associated with the hands-on performance task or not? How do the item attributes and test framework relate to whether the item is an extended constructed response item or not? How do the item attributes and test framework relate to whether the item appears early in one of the blocks of items presented to the student?
4. What is the connection between item responses, item attributes, and the test framework?

The first set of questions relate to defining what the item attributes and framework categories mean. Next a question about the relationships between the item attributes and test specifications and framework is raised. The third set of questions try to get at whether the items that are associated with the performance tasks, the extended constructed response items, or the items early in each block tend to have certain characteristics. The final question begins to look at approaches to summarizing the responses of students. In future research this question will be examined in detail. In addition, in future research, a fifth question will be addressed, "What are the options for reporting results of a NAEP Science Assessment that includes many constructed response items and items associated with hands-on performance tasks?."

The Framework and Specifications as Developed for the 1993 NAEP Science Field Test

The framework and specifications as developed for the 1993 NAEP Science field test were developed by the Council of Chief State School Officers and the National Assessment Governing Board (*Science Framework for the 1994 National Assessment of Educational Progress*, Pre-publication draft; *Science Assessment and Exercise Specifications for the 1994 National Assessment of Educational Progress*, Pre-publication draft). The consensus process used to develop the framework and specifications involved curriculum specialists, science teachers, local science supervisors, state supervisors, administrators, and parents. It also involved representatives of scientific associations, business and industry, and unions. Finally, cognitive psychologists and science educators were involved. The framework and specifications emphasize what is considered essential learning in science and recommend the use of innovative assessment techniques. Recognition is made that the various constituencies listed above hold diverse views about science assessment. Lack of agreement about the definition of scientific literacy, themes that span all subdivisions of science, ideal science instruction, and important outcomes of instruction inhibits the public's understanding of what science education is all about. Research leading to more general agreement is needed. (*Science Framework for the 1994 National Assessment of Educational Progress*, Pre-publication draft p. 3)

The framework for the assessment is composed of a matrix with two major dimensions, fields of science and knowing and doing science. The fields of science include the earth, physical, and life sciences. Astronomy, geology, meteorology, and oceanology are parts of earth science. Physics and chemistry are physical sciences. Biology, health, and nutrition are aspects of life science. The knowing and doing dimension is related to thinking skills, and includes conceptual understanding, scientific investigation, and practical reasoning. Two framework components that are external to the matrix are the nature of science, and ,emes. These two components integrate earth, physical, and life sciences. Overarching science themes include models, systems, and patterns of change. Historical development of science and technology, the habits of mind used in science, and methods of problem solving are parts of the nature of science. The framework is represented in Figure 2.

---------Insert Figure 2 about here---------

The distribution of time spent on assessment items across the fields of science and the knowing and doing dimension vary by grade. For grades 4 and 12, each of the fields of science have equal importance. For grade 8, 40% of the testing time is spent on life science items, while the remaining assessment time is split evenly between the other two fields of science. For each of the grades, 45% of the assessment time is spent on items that evaluate conceptual understanding. For grade 4, an equal amount of time is spent on scientific investigation, while only 10% of the assessment is spent on practical reasoning. For grades 8 and 12, 30% of the time is spent on scientific investigation and 25% of the assessment time is spent on practical reasoning items. (*Science Assessment and Exercise Specifications for the 1994 National Assessment of Educational Progress*, Pre-publication draft pp. 4 & 5) The specifications provide that multiple choice, short and extended open-ended paper and pencil, and open-ended performance items should be used for measuring all of the ways of knowing and doing science.

Each of the booklets used in the 1993 field test of NAEP items contained three blocks of items. The booklets contained a block devoted to a science theme, a block containing a performance task, and a block of other items. Every block contained multiple-choice, short constructed-response, and extended constructed response items, and the performance task block always appeared in the final position in the booklet. The number of items in each block varied from 4 to 17. We examined one of the eighth-grade booklets in detail.

In our analyses, the items were categorized as pertaining to the fields of science, knowing and doing categories, types of themes, and nature of science/technology. The aspects of the framework and specifications that were specified are:

Fields of Science	
Earth Science	ES
Physical Science	PS
Life Science	LS
Knowing and Doing Science	
Scientific Investigation	SI
Conceptual Understanding	CU
Practical Reasoning	PR
Themes	
Systems	SYS
Models	MOD
Patterns of Change	PC
Not Applicable	NA
Nature of Science/Technology	
Nature of Science	NS
Nature of Technology	NT
Not Applicable	NA

These categories of items were compared to categories specified a posteriori by subject-area experts. They were also related to item responses by the prediction of the difficulty of the items.

The Item Responses

The 1993 field test of science items was completed between January and March of 1993.

In all, 3908 fourth graders, 3585 eighth graders, and 3041 twelfth graders participated in the main part of the science field test. Approximately 350 students on average received each booklet. Grade 4 students were given 20 minutes to complete each block, while students in grades 8 and 12 were given 30 minutes to complete each block. Approximately twice as many exercises were included in the field test of fourth-grade items than would be required for a regular NAEP assessment. At grades 8 and 12, approximately 40% more exercises than are needed were field tested. Following the field test administration, the constructed-response items were scored by professional scorers in Iowa City, Iowa, at the headquarters of National Computer Systems under the supervision of ETS staff. Prior to the scoring, staff from ETS and members of the Science Instrument Development Committee met to review students' responses, finalize scoring guides, and select exemplar responses for use as anchor papers and training papers in training the professional scorers. Each day during the five week professional scoring period, scorers were trained on specific items and then scored those items. Following the field test scoring, a complete item analysis of the field test was conducted at ETS. These results were used to tentatively select items for the next NAEP science assessment, and were used in the current study.

Five hundred and sixteen eighth grade students were administered the booklet that we examined in detail (weighted N = 375.5). There are 16 items (3 multiple-choice and 13 constructed-response items) in the first block of the booklet. This block contains items measuring knowledge about a variety of fields of science and ways of knowing and doing science. The second block contains items related to the theme of an ecosystem. There are 13 items (7 multiple-choice and 6 constructed-response items) in this block. The third block contains 8 items (1 multiple-choice and 7 constructed-response items) pertaining to a performance task. Future analyses will replicate the current study, and make use of the data from other booklets and for other grade levels.

The Item Attributes

In order to more fully understand what is being measured in the 1993 NAEP science field test, a committee of five science experts, measurement statisticians, and cognitive scientists developed a list of item characteristics or attributes that were deemed to influence the responses to the items in the booklet under consideration. First, committee members examined frameworks and specifications for a variety of science assessments and curriculum guides for a variety of science education programs. With these as a basis, the items within the booklet were examined in detail to identify what contributed to the way students would respond to the items. As stated earlier, additional information could be gained from student response protocols where students are asked what they are thinking about as they respond to the items. However, due to time and budgetary constraints protocols are not available. From the detailed notes about each item, item attributes were identified and questions about the items were developed to aid in categorizing items. These questions were used to categorize items as having specific attributes or not. Each item and the attributes associated with it were evaluated, and the list of item attributes and the questions associated with them were revised until items could be categorized easily by subject-area experts. Several item attributes that were not found in the items of the studied booklet were identified as a part of the comprehensive viewpoint used in the process.

The item attributes fell into six major categories: specific knowledge, item format and vocabulary, reasoning, hypothesis testing and the design of a test for a hypothesis, explanation, and communication. Specific knowledge implies that knowledge must be brought to the task by the student, as opposed to being provided in the text of the item. Item format and vocabulary pertains to information provided by the item. Reasoning involves the thought processes necessary to respond well to an item. Hypothesis testing and the design of hypothesis tests is a basic part of scientific thinking. The requirement of an explanation forces students to justify or compare responses. Finally, communication of scientific information can be a part of the science tasks. The full list of questions that define item attributes follows.

Coding Questions that Match the Skills that Items may Require

Specific knowledge

1) Is knowledge of facts necessary to answer the item using a reasonable strategy? Can knowledge of facts be used to answer item? For items in this category, students don't have to understand facts; only to remember them.
2) Is knowledge of science procedures necessary to answer the item using a reasonable strategy? (i.e., knowledge of lab procedures or experimental design).
3) Is knowledge of concepts necessary to answer the item using a reasonable strategy? This is often denoted by a noun.
4) Is knowledge of principles necessary to answer the item using a reasonable strategy? (an assertion: a law or a theory)? A key issue with this attribute is to differentiate between a concept and a principle, law, or theory. For example the kinetic theory of gases includes a number of assumptions and principles which in turn are based on a number of concepts. Perhaps, greater complexity and a hierarchical arrangement are the distinguishing features of this attribute.
5) Is knowledge about relationships between facts, procedures, concepts or principles necessary to answer the item using a reasonable strategy?

Item Format and Vocabulary

6) Does the item contain a table, graph or figure?
7) Does the item refer, directly or indirectly, to a table, graph or figure contained in the same block of items (but separate from item)?
8) Does the item contain or refer to a table, graph or figure that is complex? Does the item contain or refer to a table, graph or figure that is dynamic, multiple and/or abstract rather than static, single or concrete?
9) Is a table, graph or figure necessary to answer the item using some conceivable strategy? Is it possible to use a table, graph or figure to answer the item?
10) Is a table, graph or figure necessary to answer the item using every strategy? Is it necessary to use a table, graph or figure to answer the item no matter which strategy is used?
11) Does the item contain science terminology or vocabulary that must be understood in order to answer the question correctly?
12) Must the response meet all the conditions found in the stem? Multiple-

choice questions are always coded "yes". Plurals or "name two things" in a constructed-response question often distinguishes between score category levels. This question combines information in the item and the item response.

13) Does the stem contain hypotheticals (what if), exception, negation or other text phrases that make the task complex? Also, "Suppose ..."
14) Does the item require comprehension of every option of a multiple-choice item in order to define the possible correct answers?
15) Does the item refer, directly or indirectly, to a student-generated table, figure, or text, separate from the stem?
16) Is the reading level complex? For example, does the item contain at least an imbedded independent clause?
17) Does the item require information that can be gained through practical experience (not formally instructed)?
18) Does the item require only information found in the item itself? Information does not include procedural knowledge, which may be needed in addition to the information provided in the item.
19) Is all the information necessary to answer the question available in the text, a table, a graph or a figure in the block with the item? Information does not include procedural knowledge, which may be needed in addition to the information provided in the block of items.

(coded as 41)
Is all the information necessary to answer the question available in the block with the item or generated directly by the student for the assessment as part of a performance task? The responses for this question and question 19 were identical for all items in Booklet S21. Therefore, this question as ignored in the analysis.

(coded as 39)
Can the item be solved by elimination of options?

(coded as 40)
Can information from the options be used to constrain or inform the definition of the task? The item must be a multiple-choice item, if the response is yes.

Reasoning

20) Is deductive reasoning necessary to answer the item using a reasonable strategy? This may include analysis of attribute (part-whole) relationships.
21) Is reasoning from a general concept, principle or law to a specific conclusion necessary to answer the item using a reasonable strategy?
22) Is tracing a cause-effect from one component to another within a system necessary to answer the item using a reasonable strategy?
23) Is formal inductive reasoning necessary to answer the item using a

reasonable strategy?

24) Is the application of a concept or principle necessary to answer the item using a reasonable strategy? (Application of concept or principle, as opposed to understanding)
25) Can thinking with or about models be used to answer the item using a reasonable strategy?

Hypothesis testing and design of a test for a hypothesis

26) Is the generation of a hypothesis or prediction necessary to answer the item? The hypothesis or prediction must refer to the future or to changed conditions.
27) Does the item require the identification of variables or control groups involved in a design of a test for a hypothesis?
28) Does the item require the generation of specifically operationalized procedures to be used in testing a hypothesis?
29) Does the item require the use of a control group in a design of a test for a hypothesis?
30) Does the item require the use of multiple control groups in a design of a test for a hypothesis?

Explanation

31) Does the item (specifically, each category) require a reason or justification for a response?
32) Does the item require an explanation comparing an attribute against a standard? (Why is your answer a good one, rather than why did you say that?) No items in Booklet S21 required this.
33) Are the alternatives reasons or explanations? Constructed-response items are always coded "no".
34) Does the item require generation of a number of (not just one) possible scientific explanations? No items in Booklet S21 required this.

Communication

35) Is the item an extended constructed-response item?
36) Is the item a short constructed-response item?
37) Does the item require drawing a diagram?
38) Does the item require filling in a table?

(coded as 42)

Does the item require constructing a graph? No items in Booklet S21 required this. This question was ignored in the analysis.

Method

The dataset that we focus on in this paper was one booklet from the 1993 NAEP Science field test. That booklet contained 37 items (11 multiple-choice and 26 constructed-response items, and 8 items associated with hands-on performance task) in three blocks. The booklet was administered to 516 student in grade 8. The datasets for attributes and framework categories were developed by the consensus of science experts, measurement statisticians, and test developers. Data consisted of 0/1 codings for each item. Here, 1 means that an item has the attribute or falls in the category of the framework; 0 means that an item does not have the attribute or does not fall into the category of the framework. Table 1 contains the number of items having each attribute or the number of items falling in a framework category.

---------Insert Table 1 about here----------

The first part of the data analysis was concerned with answering question set 1 and question 2. These questions ask about the association among item attributes, among framework categories, and between the item attributes and the framework categories. To explore these relationships, we examined principal component analyses for each association of interest. After a number of principal components were selected, they were rotated using a varimax rotation. In these analyses, three attributes (attributes 29, 32, and 34) and two categories of the framework (models and nature of technology) were excluded because none of the items required them. One attribute (attribute 12) was excluded because all of the items required it.

To explore question set 3, we identified item attributes and framework categories that we had reason to think were related to three specific groups of items. The three groups of items of interest were the items associated with the hands-on performance task, the extended constructed response items, and the items at the beginning of each block. We examined two-by-two contingency tables for each attribute or framework category and each of the three groups of items of interest.

In the last part of the data analysis, we used regression analysis to examine the relationship between item responses, item attributes, and the test framework (question 4). In these analyses, item easiness (p-value for dichotomous items and scaled mean score for polytomous items) was used as the dependent variable. The 36 item attributes or the 9 categories of the framework were used as the predictors.

Results

The results are presented according to the order of question sets.

Question set 1: How are the item attributes related to one another? How are the framework categories related to one another?

Table 2 presents the factor loadings of the attributes from the eight factor solution of the varimax rotation of the principal component analysis. The eight factor solution was selected after examining the scree plot in Figure 3. The value of the eighth eigenvalue is 1.3.

----------Insert Table 2 and Figure 3 about here----------

Six attributes have factor loadings with large magnitudes for factor one. They are listed as the first six attributes in Table 2. These attributes are knowledge of science procedures (attribute 2), complex reading level (attribute 16), thinking with or about models (attribute 25), identification of variables or control groups involved in a design of a test for a hypothesis (attribute 27), generation of specifically operationalized procedures to be used in testing a hypothesis (attribute 28), and the use of multiple control groups in the design of a test of a hypothesis (attribute 30). Several of these attributes are concerned with setting up scientific hypothesis tests. It is interesting that these attributes are related to complex reading level and thinking with or about models.

Seven attributes have factor loadings with large magnitudes for factor two. They are listed as the second group of attributes in Table 2. Six of these attributes are knowledge of facts (attribute 1), knowledge of concepts (attribute 3), knowledge about relationships between facts, procedures, concepts or principles (attribute 5), reasoning from a general concept, principle or law to a specific conclusion (attribute 21), tracing a cause-effect from one component to another within a system (attribute 22), and the application of a concept or principle (attribute 24). These attributes are related to reasoning and knowledge. The item attribute of having all of the information needed to answer the item available in the text, or in a table, a graph or a figure in the block with the item (attribute 19) is related to these reasoning and knowledge attributes in a negative way.

Three attributes have factor loadings with large magnitudes for factor three. Items having these attributes contain a table, graph or figure (attribute 6), require only information found in the item itself (attribute 18), and have options that can be used to inform the definition of the task (attribute 40). All of these attributes are concerned with information available within an item.

Five attributes have factor loadings with large magnitudes for factor four. Four of these attributes have negative factor loadings. Items having these attributes require comprehension of every option of a multiple-choice item (attribute 14), require deductive reasoning (attribute 23), have alternatives that are reasons or explanations (attribute 33), and can be solved by elimination of options (attribute 39). These attributes are concerned with deductive reasoning. Items that require formal inductive reasoning (attribute 30) load at the other end of the scale for factor four.

Five attributes have factor loadings with large magnitudes for factor five. Four of these attributes involve tables, graphs or figures. Only the attribute that an item requires a reason or justification for a response (attribute 9) does not. Two attributes are associated with factor six. They are having a complex expression in the item's stem (attribute 13) and requiring the generation of a hypothesis or prediction (attribute 26). Four attributes are associated with factor seven. Two have positive factor loadings. Items with these two attributes have a table, graph or figure related to them but separate from them (attribute 7), and have a complex table, graph or figure associated with them (attribute 8). The other two attributes are knowledge of principles (attribute 4), and knowledge of information gained through practical experience (attribute 17). Four attributes are associated with factor eight. Three of these have positive factor loadings. They are science terminology or vocabulary (attribute 11), an extended constructed response format (attribute 35), and drawing a diagram (attribute 37). The fourth attribute is a short constructed response format (attribute 36).

After examining the scree plot in Figure 4, a three factor varimax rotation of the principal component extraction was selected for the framework categories. Table 3 presents the factor loadings of the framework categories for the three factors. Factor one is defined by the physical sciences (PS) at one end of the scale, and the life sciences (LS) at the other end of the scale. In this booklet, items in the theme block measuring knowledge about systems (SYS) and items measuring conceptual understanding (CU) are related to the life sciences (LS). Two framework categories have positive loadings on factor two. They are the earth sciences (ES) and patterns of change (PC). Practical reasoning (PR) has a negative loading on this factor. Factor three shows that nature of science (NS) and scientific investigation (SI) are related.

----------Insert Figure 4 and Table 3 about here----------

Question 2: What is the relationship between the item attributes and the test framework?

Table 4 presents the factor loadings of the attributes and the framework together. The eight factor solution was selected after examining the scree plot in Figure 5. The value of the eighth eigenvalue was 1.7. Factors in this solution closely matched factors in the eight factor solution for the attributes alone.

---------Insert Table 4 and Figure 5 about here---------

Factor one in the current analysis had a structure similar to that of factor one in the solution described for the attributes alone. The same six attributes had large positive loading for both solutions. In the current analysis, items categorized as nature of science items were also associated in a positive was to this factor. It is clear that the framework category nature of science is related to setting up scientific hypothesis tests.

Factor two in the current analysis had a structure similar to that of factor two in the attribute solution described above. The same six reasoning and knowledge attributes had positive loadings for both solutions and items containing all information necessary to answer the item (attribute 19) had a negative loading for both solutions. Scientific investigation had a factor loading similar to that for items containing all information necessary to answer the item (attribute 19). This reflects the fact that items measuring scientific investigation must provide enough information for students to do an investigation. Scientific investigation also had a reasonably high loading on factor one of the current analysis, indicating that scientific investigation is related to hypothesis testing. In particular, it is related in terms of factor one to attribute 2, the knowledge of science procedures.

Factor three in the current analysis has a structure similar to that of factor five in the attribute-only solution. The only attribute that loaded heavily on factor five in the attribute-only solution, but does not have a loading with a large magnitude for factor three in the current analysis is the attribute where items require a reason or justification for a response (attribute 31). In other words, this factor represents inclusion of tables, graphs or figures for items. In this booklet, the framework category physical science is positively related to tables, graphs, and figures.

Factor four in the current analysis has a structure similar to that of factor seven in the attribute-only solution, other than attribute 17 (practical experience). This factor is also related to factors one and two in the framework-only analysis. In the framework-only analysis life science and systems were similar to one another, but different from earth science and patterns of change. For factor four in the current analysis, these two sets of framework categories differ in the sign of the loadings. Earth science and patterns of change are related to the attribute knowledge of principles (attribute 4), while life science and systems are related to tables, graphs, or figures that are complex (attribute 8) or external to the item (attribute 7). These relationships are likely to be particular to the booklet we studied.

Factor six in the current analysis has a structure similar to that of factor eight in the attribute-only analysis. In the current analysis, conceptual understanding is related positively to science terminology (attribute 11), the extended constructed response format (attribute 35), and drawing a diagram (attribute 37). Practical reasoning is positively related to requiring a justification for a response (attribute 31), and the short constructed response format (attribute 36).

The other factors in the current analysis, primarily, grouped attributes with each other, rather than with framework categories. Factor five in the current solution was most like factor four in the attribute-only solution, factor seven in the current solution was most like factor three in the attribute-only solution, and factor eight in the current solution was most like factor six in the attributes-only solution.

Question set 3: How do the item attributes and the test framework relate to whether the item is associated with the hands-on performance task or not? How do the item attributes and the test framework relate to whether the item is an extended constructed response item or not? How do the item attributes and the test framework relate to whether the item appears early in one of the blocks of items presented to the student?

Table 5 contains the crosstabulations between items associated with the hands-on performance task and certain attributes. We hypothesized that the items associated with the performance task would most likely measure knowledge about relationships between facts, procedures, concepts or principles (attribute 5), and application of concepts or principles (attribute 24). We thought they would most likely have an extended constructed response (attribute 35), or a short constructed response format (attribute 36), and require drawing a diagram (attribute 37) or filling in a table (attribute 38). We also hypothesized that they would not likely measure knowledge of facts (attribute 1).

----------Insert Table 5 about here----------

For the booklet we examined, only one of these hypotheses appeared to be correct. The items associated with the performance task are not likely to measure knowledge of facts (attribute 1). They are also reasonably unlikely to measure knowledge about relationships between factors, procedures, concepts, or principles (attribute 5), and the application of concepts or principles (attribute). Because of the use of all item types in all blocks of items, including those associated with the performance task, there is little relationship between extended constructed response (attribute 35) and short constructed response (attribute 36) formats and the items in the performance task block. These items are also not related strongly to drawing a diagram (attribute 37) or filling in a table (attribute 38).

Table 5 also contains crosstabulations between items associated with the hands-on performance task and the framework categories. We hypothesized that, for this booklet, the items associated with the performance task would be positively related to physical science, conceptual understanding, and scientific investigation. For this booklet, the performance task block was positively related to the physical sciences and scientific investigation. We would expect that the relationship with scientific investigation would be replicated for the other booklets. Although we would like items associated with the performance task to measure conceptual understanding, they are unlikely to do so. This is due to the design of the test framework, where items are categorized as either scientific investigation, conceptual understanding or practical reasoning items, but not more that one of these.

Table 6 contains crosstabulations between extended constructed response items and the item attributes. We hypothesized positive relationships between extended constructed response items and knowledge about relationships between facts, procedures, concepts or principles (attribute 5), complex reading level (attribute 16) and requiring a reason or justification for a response (attribute). We hypothesized a negative relationship between extended constructed response items and knowledge of facts (attribute 1). The only hypothesis that seems to be correct is that the extended constructed response format is related to a complex reading level (attribute 16). The other relationships are not strong.

----------Insert Table 6 about here----------

Crosstabulations between extended constructed response items and specific framework categories are also in Table 6. Our expectation was that the extended constructed response format was not more likely for any of the aspects of the framework that for any other. The largest relationships were those with life science and practical reasoning. These relationships are not strong.

Table 7 contains the crosstabulations between the items at the beginning of each block and certain attributes. It was hypothesized that items in the first third of a block might be easier than items placed later in a block. For that reason, we expected a positive relation between items at the beginning of the block and knowledge of facts (attribute 1), and a negative relationship with knowledge about relationships between facts, procedures, concepts or principles (attribute 5), complex reading level (attribute 16), requiring a reason or justification for an answer (attribute 31), an extended constructed response format (attribute 35), and a short constructed response format (attribute 36). Only the relationship between items in the beginning of a block and knowledge of facts was reflected in the data. The lack of other hypothesized relationships may be due to the fact that those relationships hold only for the first or first two items in a block. This analysis categorized the first third of the items in a block as appearing at the beginning.

----------Insert Table 7 about here----------

In order to examine the relationships between knowing and doing science categories, Table 7 contains the crosstabulations between the items at the beginning of each block and scientific investigation, conceptual understanding, and practical reasoning. For this block, conceptual understanding had a strong relationship to position within the block and scientific investigation had a negative relation to position within the block.

Question set 4: What is the relationship between item responses, item attributes, and the test framework?

Table 8 contains the results of a regression analysis used to investigate the relationship between item responses and item attributes. After a preliminary backward selection of variables, 98% of the variance in the item mean was explained by 27 attributes. Among the first category of attributes (specific knowledge), knowledge of principles (attribute 4), and knowledge about relationships between facts, procedures, concepts or principles (attribute 5) were included in this regression. Note that attributes 1, 2, and 3 were selected out using the backwards selection procedure. Ten attributes from the second category (item format and vocabulary) were selected for the model. Attributes 7, 9, and 12 were not included in the model. Among the reasoning category of item attributes, four attributes were included in the model. Attributes 21 and 24 were excluded. Of the hypothesis testing attributes, two items were included in the regression model. In the explanation category of attributes two items were selected for the prediction equation of item mean score. Four attributes in the communication category were included in the model. When this analysis was replicated using only attributes with more than one items associated with them, R^2 values were in the .80 to .90 range.

----------Insert Table 8 about here--------

Table 9 contains the results of a regression analysis predicting item mean score from the framework categories. After a backward selection procedure, earth science, patterns of change, nature of science, and practical reasoning predicted the item mean score with an R^2 of .30 and an adjusted R^2 of .21.

----------Insert Table 9 about here----------

These analyses, with the input from science experts have contributed to decisions about which item characteristics will be retained for future analyses. Subsequent analyses will replicate these analyses for other booklets.

Conclusion

As part of an on-going research agenda, this study has provided preliminary information about the framework of the next NAEP science assessment. It also provided information about item attributes that science experts believe to be related to the responses that students give to these items. Future research will further explore the relationship of the item attributes and test framework to actual student responses. In addition, information from the other booklets in the 1993 field test of items will be compared to these results.

References

Science assessment and exercise specifications for the 1994 National Assessment of Educational Progress, Pre-publication draft, NAEP Science Consensus Project, Council of Chief State School Officers and the National Assessment Governing Board, U. S. Department of Education.

Science framework for the 1994 National Assessment of Educational Progress, Pre-publication draft, Council of Chief State School Officers for the National Assessment Governing Board, contract number RS90018001.

Table 1
Item Frequencies for each Item Attribute and Framework Category

Item Attributes

Attribute	1	2	3	4	5	6	7	8	9	10
# of items not having attribute	8	33	11	36	13	30	11	11	16	22
# of items having attribute	29	4	26	1	24	7	26	26	21	15

Attribute	11	12	13	14	15	16	17	18	19	20
# of items not having attribute	22	0	29	32	31	34	31	36	35	4
# of items having attribute	15	37	5	6	3	6	1	2	33	26

Attribute	21	22	23	24	25	26	27	28	29	30
# of items not having attribute	11	20	36	18	33	30	35	35	37	34
# of items having attribute	26	17	1	19	4	7	2	2	0	3

Attribute	31	32	33	34	35	36	37	38	39	40
# of items not having attribute	22	37	34	37	31	17	35	36	33	35
# of items having attribute	15	0	3	0	6	20	2	1	4	2

Framework Categories

Category	ES	PS	LS	SI	CU	PR	SYS	MOD	PC	NS	NT
# of items not in category	31	26	16	27	15	32	18	37	35	34	37
# of items in category	6	11	21	10	22	5	19	0	2	3	0

Table 2
Factor Loadings of the Varimax Rotation of the First Eight Principal Components of the Item Attributes

Attributes	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Att 30	0.95235	0.11235	-0.02310	0.02262	-0.13100	-0.11776	0.05163	-0.04837
Att 2	0.87159	-0.01589	-0.09009	0.00911	-0.02289	-0.10867	0.07694	0.03805
Att 28	0.87135	0.07651	-0.02876	0.02856	0.02276	0.00860	-0.03733	0.12869
Att 16	0.71038	0.15489	-0.04104	0.05452	0.10045	0.15525	0.03109	0.05807
Att 27	0.67212	0.09598	-0.02132	0.01489	-0.15686	-0.13204	0.06483	-0.10549
Att 25	0.62422	0.11040	0.12787	0.08922	0.12066	0.32622	-0.25249	0.09950
Att 3	0.10966	0.86943	-0.10180	0.03031	-0.26547	0.12178	-0.12511	-0.09525
Att 21	0.08446	0.82099	-0.05313	0.41292	-0.26169	0.06901	-0.08295	-0.09289
Att1	0.01327	0.77134	-0.12794	0.01411	-0.44846	0.06859	-0.16088	0.19310
Att 24	0.24821	0.66218	-0.08643	-0.04213	0.03184	0.26358	-0.40807	-0.03246
Att 5	0.18988	0.63704	-0.20887	-0.03567	-0.26244	0.17654	0.02333	0.00432
Att 22	0.28257	0.53846	-0.03008	0.26036	-0.16416	0.43330	-0.16049	0.03950
Att 19	-0.01175	-0.49750	-0.06543	0.12534	-0.05250	0.01091	-0.01155	-0.04805
Att 18	-0.02442	-0.11370	0.87965	-0.10575	0.07335	-0.07160	0.10012	-0.05185
Att 40	0.01353	-0.10581	0.73930	-0.07925	0.00591	0.29636	-0.12933	-0.00205
Att 6	-0.12292	0.06161	0.45037	-0.02124	0.36253	-0.16705	-0.18535	0.24649
Att 23	-0.03509	0.06980	-0.09331	-0.84335	-0.05744	-0.02339	-0.09834	-0.00979
Att 30	0.11792	0.21748	-0.38532	0.59955	0.16632	0.06381	0.03436	-0.33922
Att 39	-0.07923	-0.02539	0.43258	-0.59566	0.10600	-0.20676	0.18598	-0.07700
Att 33	-0.05296	0.02450	0.51589	-0.58185	0.05619	-0.13037	-0.15353	-0.03382
Att 14	-0.06407	-0.17104	0.46405	-0.54824	-0.26914	0.35166	-0.15684	0.07119
Att 9	0.21626	-0.22690	0.14670	0.17745	0.73956	0.02460	-0.07537	0.10815
Att10	-0.16403	-0.20534	0.05179	0.09166	0.70580	-0.06628	0.19430	-0.00936
Att 15	-0.04137	-0.35318	-0.21668	-0.06829	0.58080	-0.10419	0.20758	-0.15757
Att 31	-0.26543	0.27977	-0.25082	0.19801	0.34035	0.23354	0.07080	-0.33044
Att38	-0.03075	-0.11280	0.04899	-0.10166	0.30601	-0.13597	0.14836	-0.09504
Att 13	-0.12396	0.30464	0.03331	0.10933	-0.07073	0.86752	0.11663	-0.14066
Att 26	-0.11736	0.26528	0.00036	0.06114	-0.20472	0.82986	0.20713	-0.13997
Att 7	0.17428	-0.16245	-0.37973	0.23552	-0.00970	0.02954	0.87517	0.01406
Att 8	-0.42659	-0.15740	0.00082	0.28215	0.26129	0.08956	0.56380	0.06151
Att 17	-0.18272	0.15311	-0.07061	0.24381	-0.17765	-0.33292	-0.39821	-0.11618
Att4	-0.07306	0.06276	-0.05624	0.06935	-0.19380	-0.12293	-0.32062	-0.00070
Att 37	-0.11203	0.03283	-0.03745	0.03308	-0.00444	-0.08154	0.15010	0.74483
Att 35	0.48649	0.05892	-0.07667	0.09966	0.14233	-0.11021	0.01955	0.70266
Att 36	-0.22751	0.09070	-0.28211	0.24993	0.24210	0.03526	0.28098	-0.57477
Att 11	-0.30025	0.19633	-0.07226	-0.21587	-0.23599	0.12970	-0.01259	0.33911

Table 3
Factor Loadings of the Varimax Rotation of the First Three Principal Components of the Framework Categories

Categories	Factor 1	Factor 2	Factor 3
PS	-.91	-.18	-.13
LS	0.88	-.41	0.15
SYS	0.74	-.44	-.42
CU	0.54	0.47	-.52
ES	-.17	0.82	-.15
PC	-.01	0.72	-.11
PR	-.02	-.56	-.22
NS	0.19	0.04	0.93
SI	-.57	-.09	0.74

Table 4
Factor Loadings of the Varimax Rotation of the First Eight Principal Components of the Item Attributes and the Framework Categories

Attributes	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Att 30	0.95257	0.05630	-0.11742	0.02308	0.00870	-0.01039	-0.05907	-0.07209
NS	0.95257	0.05630	-0.11742	0.02308	0.00870	-0.01039	-0.05907	-0.07209
Att 2	0.88706	-0.05557	0.02304	-0.03348	0.00827	0.06261	-0.13797	-0.02551
Att 28	0.87592	0.09082	-0.00965	0.00611	0.06084	0.09914	0.03914	0.00370
Att 27	0.74139	0.01806	-0.15058	0.02965	-0.02932	-0.07196	-0.09628	-0.10105
Att 16	0.73389	0.17841	0.02533	0.10457	0.10696	-0.03310	0.05001	0.14190
Att 25	0.62238	0.19700	0.11195	-0.17287	0.14056	0.06353	0.28563	0.32173
Att 3	0.11362	0.81856	-0.33562	-0.06229	0.07427	-0.10405	-0.10155	0.07777
Att1	0.03201	0.77608	-0.50723	-0.05262	0.01960	0.14714	-0.12422	-0.00147
Att21	0.08482	0.75092	-0.35923	0.02796	0.40247	-0.06197	-0.03157	0.01420
Att24	0.27612	0.72565	-0.03624	-0.30158	0.03583	-0.27786	0.05788	0.13052
Att19	-0.01160	-0.65268	-0.07067	-0.08765	0.20614	-0.12221	-0.02047	0.03410
Att 5	0.19292	0.62359	-0.28894	-0.08323	0.02354	0.00279	-0.31038	0.22924
SI	0.56959	-0.60815	0.42506	-0.00168	0.08361	-0.14418	-0.20686	-0.04229
Att 22	0.29550	0.55260	-0.27305	-0.11754	0.35237	-0.08769	0.06811	0.35450
Att 15	-0.05079	-0.30436	0.76992	0.02833	-0.03565	-0.07644	-0.26154	-0.01526
Att 10	-0.19784	-0.15445	0.72415	0.13397	0.16034	0.13735	0.11707	0.00477
Att 9	0.19568	-0.14736	0.71203	-0.02733	0.25427	0.16314	0.33880	0.04716
PS	-0.08742	-0.41406	0.70201	-0.10703	0.08923	-0.30915	-0.06406	-0.13550
Att 38	-0.01837	-0.06478	0.45553	0.10322	-0.16443	-0.05710	-0.00495	-0.12293
ES	-0.07004	-0.00165	-0.01639	-0.80221	-0.30571	0.20694	-0.05666	0.21854
PC	-0.12510	0.16765	-0.04781	-0.72535	0.22358	0.14168	0.02645	-0.15655
Att 8	-0.44082	-0.23096	0.20291	0.67425	0.27358	0.17348	-0.05284	0.13390
LS	0.16392	0.31739	-0.55634	0.66191	0.14552	0.17622	0.04732	-0.01386
SYSTEMS	-0.37312	0.35401	-0.40355	0.63793	0.19161	0.16237	0.16369	0.04565
Att 7	0.16783	-0.27640	0.06315	0.62147	0.26733	0.16475	-0.52780	0.15339
Att 4	-0.10938	0.04195	-0.20362	-0.61279	0.20487	0.03508	-0.03860	-0.18262
Att 23	-0.02935	0.13096	-0.00864	-0.16435	-0.80499	-0.01079	-0.18231	0.02131
Att 39	-0.08022	-0.06599	0.12403	0.12509	-0.70733	0.03896	0.25482	-0.09109
Att 20	0.10212	0.18604	0.18236	0.01847	0.70436	-0.28829	-0.29040	0.02380
Att 33	-0.05057	0.02205	0.01610	-0.06542	-0.68877	-0.05179	0.48249	-0.11541
Att 14	-0.09192	-0.13171	-0.23671	-0.22784	-0.60385	0.11797	0.34191	0.45524
PR	-0.09589	0.30322	-0.17320	0.15866	0.07604	-0.64181	-0.06091	0.07735
Att 37	-0.05886	0.06452	-0.01972	0.19717	0.03814	0.63346	-0.02964	-0.11052
CU	-0.44844	0.33896	-0.26366	-0.10896	-0.12858	0.57731	0.22952	-0.01561
Att 36	-0.22902	0.06160	0.21214	0.41090	0.24292	-0.56434	-0.26836	-0.02772
Att 11	-0.34345	0.26317	-0.18156	-0.23324	-0.16292	0.54063	-0.21975	0.24805
Att 35	0.53125	0.10202	0.09371	0.09282	0.12094	0.53568	0.01945	-0.13754
Att31	-0.27294	0.34920	0.28312	0.29536	0.24140	-0.48642	-0.13002	0.11704
Att 18	-0.03773	-0.20986	-0.07407	0.17702	-0.26408	0.00383	0.78241	-0.00181
Att 40	-0.01433	-0.12588	-0.07488	-0.10628	-0.17242	0.03058	0.72692	0.39256
Att 6	-0.11333	0.16157	0.33344	-0.02450	-0.06986	0.20033	0.63250	-0.28394
Att 26	-0.13815	0.26089	-0.24095	0.15659	0.11711	-0.24014	-0.04289	0.78836
Att 13	-0.15011	0.33340	-0.12754	0.14225	0.17470	-0.26030	0.06206	0.77435
Att17	-0.18585	0.16692	-0.30485	-0.21323	0.23412	-0.24248	0.07125	-0.56062

Table 5
Contingency Tables for the Items Associated with the Hands-on Performance Task

Attribute	1		5		24		35		36		37		38	
don't have|have	0	1	0	1	0	1	0	1	0	1	0	1	0	1
not Hands-on	1	28	7	22	11	18	24	5	15	14	27	2	29	0
Hands-on	7	1	6	2	7	1	7	1	2	6	8	0	7	1
Phi	-.841**		-.439**		-.408*		-.053		.221		-.126		.317	
Cramer's V	.841**		.439**		.408*		.503		.221		.126		.317	

Framework Category	PS		SI		CU	
not in|in	0	1	0	1	0	1
not Hands-on	26	3	26	3	8	27
Hands-on	0	8	1	7	7	1
Phi	.808**		.715**		-.502**	
Cramer's V	.808**		.715**		.502**	

Table 6
Contingency Tables for Extended Constructed Response (ECR) Items

Attribute	1		5		16		31	
don't have | have	0	1	0	1	0	1	0	1
not ECR	7	24	12	19	30	1	17	14
ECR	1	5	1	5	4	2	5	1
Phi	.053		.170		.407*		-.214	
Cramer's V	.053		.170		.407*		.214	

Framework Category	LS		PR	
not in | in	0	1	0	1
not ECR	15	16	26	5
ECR	1	5	6	0
Phi	.236		-.174	
Cramer's V	.236		.174	

Table 7
Contingency Tables for the Items in the First Third of Each Block (Early in Block)

Attribute	1		5		16		31		35		36	
don't have|have	0	1	0	1	0	1	0	1	0	1	0	1
Late in block	8	16	8	16	21	3	12	12	21	3	9	15
Early in block	0	13	5	8	13	0	10	3	10	3	8	5
Phi	.387		-.051		-.219		-.262		.137		-.230	
Cramer's V	.387		.051		.219		.262		.137		.230	

Framework Category	SI		CU		PR	
not in|in	0	1	0	1	0	1
Late in block	14	10	15	9	19	5
Early in block	13	0	0	13	13	0
Phi	-.448**		.608**		-.291	
Cramer's V	.448**		.608**		.291	

Table 8
Regression Analysis Predicting Item Mean Score from the Item Attributes

Attribute	B	SE B	Beta	T	Sig T
ATT40	.951962	.158226	.920293	6.016	.0002
ATT11	-.283986	.054393	-.596090	-5.221	.0005
ATT35	-1.370365	.141254	-2.159484	-9.701	.0000
ATT17	.774343	.104341	1.220245	7.421	.0000
ATT38	-.825307	.155719	-.572169	-5.300	.0005
ATT23	-1.446252	.287076	-1.002658	-5.038	.0007
ATT19	.410657	.082879	.396995	4.955	.0008
ATT10	.581841	.084154	1.221289	6.914	.0001
ATT4	-.474586	.113996	-.329021	-4.163	.0024
ATT6	.229701	.074622	.384618	3.078	.0132
ATT25	-.768542	.114882	-1.020262	-6.690	.0001
ATT16	-.977670	.128027	-1.140906	-7.636	.0000
ATT13	.426213	.068943	.750115	6.182	.0002
ATT37	.285421	.099583	.275926	2.866	.0186
ATT31	-.165735	.043608	-.347879	-3.801	.0042
ATT39	.382390	.089130	.507634	4.290	.0020
ATT22	.218748	.063552	.466059	3.442	.0074
ATT14	-.408542	.103090	-.597110	-3.963	.0033
ATT15	.275598	.070412	.434300	3.914	.0035
ATT33	.805743	.177607	.940273	4.537	.0014
ATT36	-.632043	.087845	-1.346616	-7.195	.0001
ATT8	.652309	.085135	1.274659	7.662	.0000
ATT30	.667468	.120177	.778912	5.554	.0004
ATT18	-3.603793	.472540	-2.498438	-7.626	.0000
ATT5	.294258	.080875	.600570	3.638	.0054
ATT28	2.366349	.255014	2.287627	9.279	.0000
ATT20	-1.071834	.216516	-1.422893	-4.950	.0008
(Constant)	.881724	.130511		6.756	.0001

R Square	.98207
Adjusted R Square	.92828
Significance of F	.0000

Table 9
Regression Analysis Predicting Item Mean Score from the Framework Categories

Category	B	SE B	Beta	T	Sig T
PR	-.225432	.103710	-.329484	-2.174	.0372
PC	.373765	.182020	.361331	2.053	.0483
NS	-.262609	.129018	-.306456	-2.035	.0502
ES	-.313906	.113861	-.494668	-2.757	.0096
(Constant)	.542141	.043825		12.371	.0000

R Square	.30170
Adjusted R Square	.21441
Significance of F	.0186

Figure 1
Information Available for this Study

Source of data:	students	test developers/ framework committee	science experts/ cognitive scientists
Data available about items:	item difficulty other item statistics	framework categories	item attributes
Data available about students:	item responses estimated thetas	none	none

Figure 2
The Framework for the 1993 NAEP Science Field Test

		FIELD OF SCIENCE		
		Earth	Physical	Life
KNOWING & DOING	Conceptual Understanding			
	Scientific Investigation			
	Practical Reasoning			
		Nature of Science		
		Themes		

Figure 3
Plot of the Eigenvalues of the Principal Component Analysis of the Item Attributes

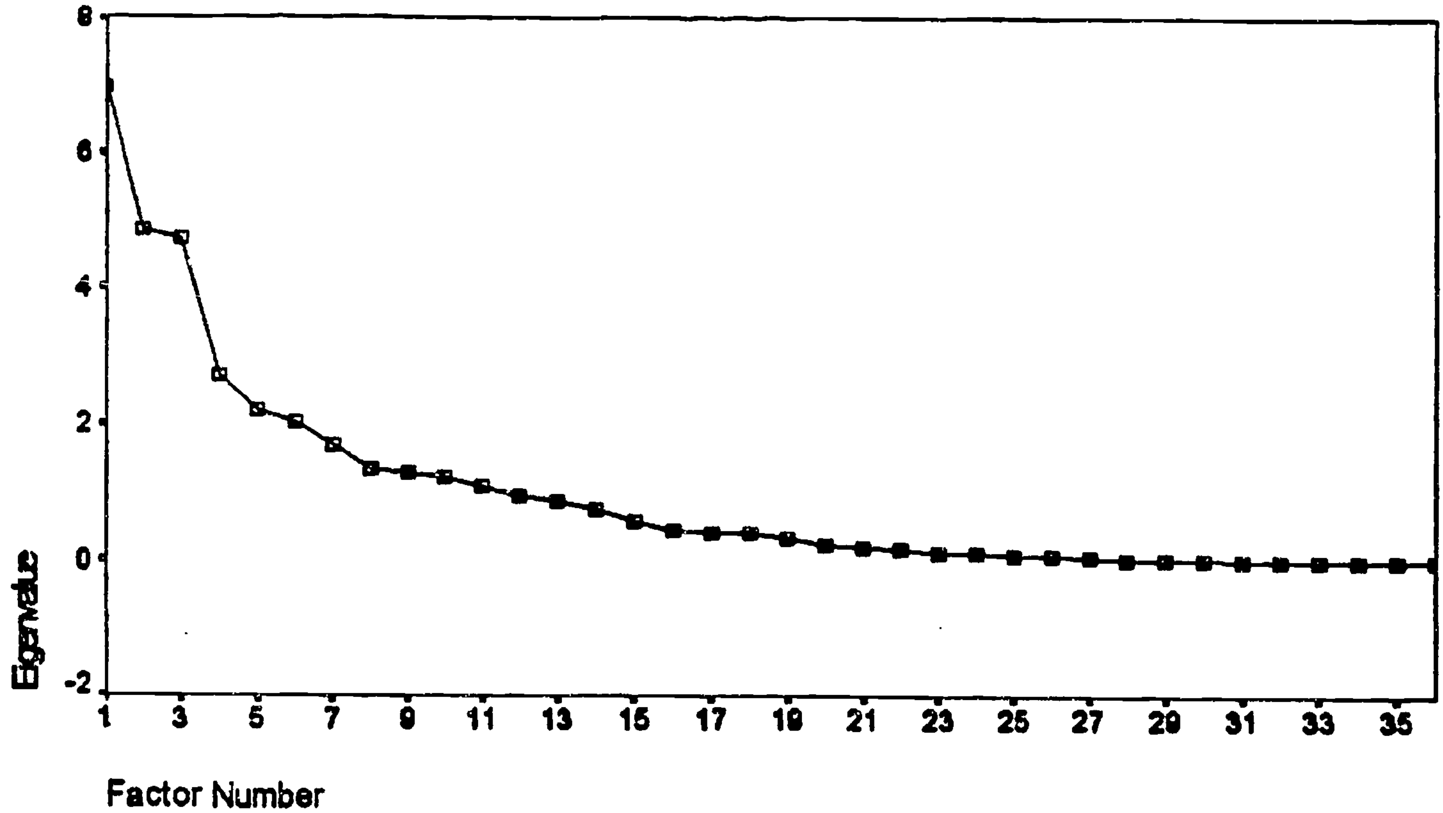

Figure 4
Plot of the Eigenvalues of the Principal Component Analysis
of the Framework Categories

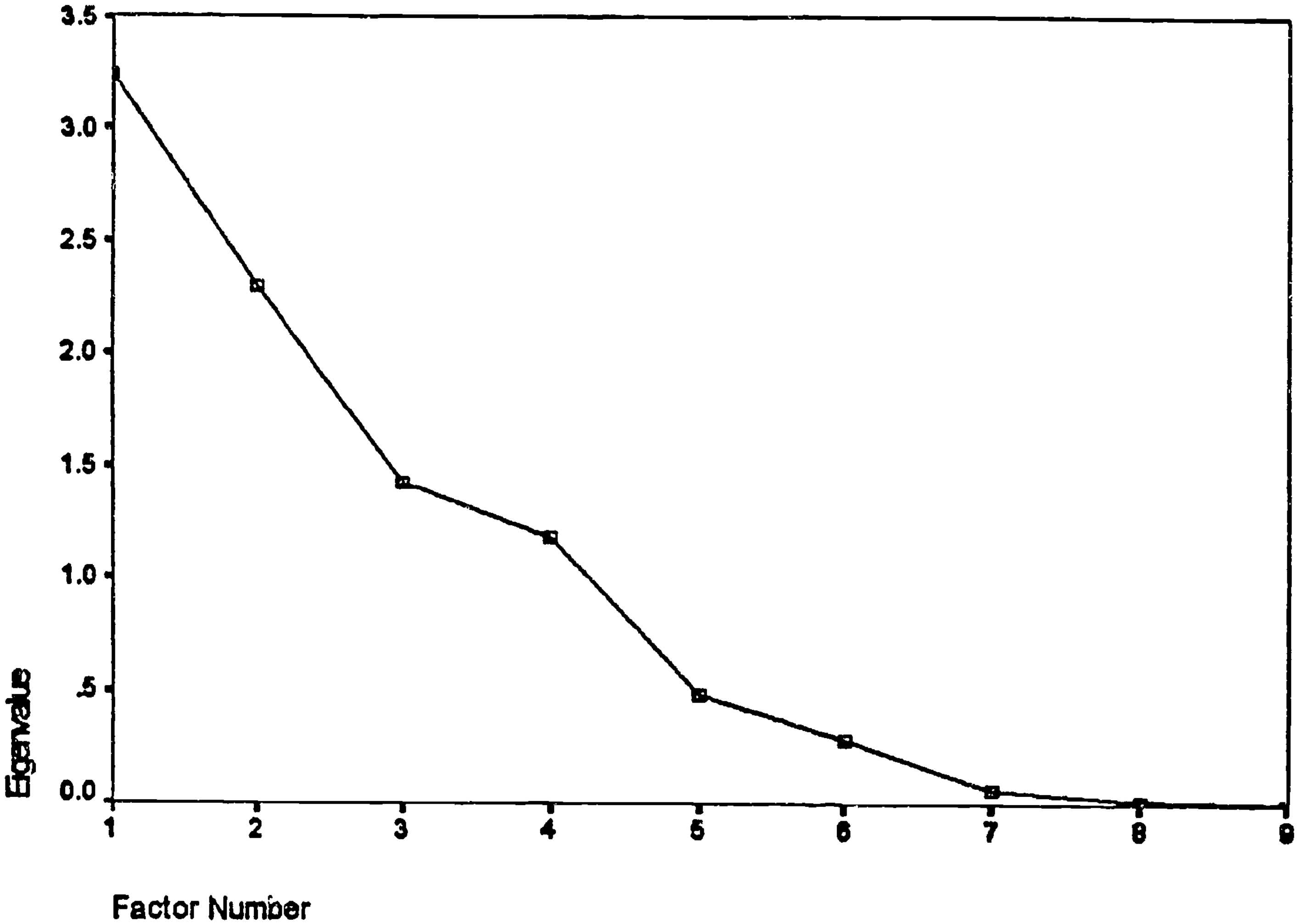

Figure 5
Plot of the Eigenvalues of the Principal Component Analysis of the Item Attributes and Framework Categories