DOCUMENT RESUME

ED 404 343 TM 025 265

AUTHOR	Yepes-Baraya, Mario
TITLE	A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment.
INSTITUTION	National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Apr 96
CONTRACT	R999J40001
NOTE	42p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cognitive Processes; Construct Validity; Difficulty Level; Grade 8; Interviews; Junior High Schools; *Junior High School Students; National Surveys; Problem Solving; *Protocol Analysis; Science Process Skills; *Science Tests; Student Motivation; Test Anxiety; Test Items; *Test Validity
IDENTIFIERS	*National Assessment of Educational Progress; Subject Content Knowledge

ABSTRACT

The cognitive processes students use in doing the 1996 science assessment of the National Assessment of Educational Progress (NAEP) were studied using two booklets from the 1993 NAEP science field test. Blocks of items from these booklets, a hands-on task block and either a conceptual/problem solving block or a theme block, were administered to 16 eighth graders who varied in proficiency in science as measured by the Metropolitan Achievement Tests. Students were interviewed after the test, and they were informed that the purpose of the study was to understand their thought processes as they answered the test items. Participants were offered a gift certificate incentive to complete the test. The combination of think-aloud protocols and concurrent interviews following the assessment was an effective way to explore participants' thought processes. Evidence suggests that the assessment is tapping the constructs it was designed to assess, namely science knowledge structure, reasoning, and hypotheses formulation and testing. In addition to lack of opportunity to learn, other main reasons for item difficulty were lack of factual knowledge, lack of conceptual understanding, and lack of knowledge of principles. The effects of motivation and test anxiety on performance were also apparent. (Contains 2 tables, 11 figures, and 17 references.) (SLD)

A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment

Mario Yepes-Baraya

Educational Testing Service
Princeton, NJ 08541

Paper presented as part of the Symposium *School-Based Special Studies of a Large-Scale Performance Assessment: Results and Implications for Assessment and Instruction* at the annual meeting of the National Council on Measurement in Education, New York City, April 1996.

The work reported herein was partially supported under the National Assessment of Educational Progress (Grant No. R999J40001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment

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

Overview and Purpose of the Study

The study described in this paper is part of a research effort to improve our understanding of the cognitive processes students use in doing the 1996 NAEP science assessment. This initiative is important because for the first time the assessment includes a variety of innovative item types and tasks designed to tap students' higher-order thinking skills. The study was conducted in 1995 and it involved two booklets from the 1993 NAEP science field test. Each booklet had three different blocks of items: a conceptual/problem-solving block, a theme block, and a performance-task block. The same blocks had been used in a prior study to identify item attributes related to item performance (Yepes-Baraya & Allen, in press; Allen, Park, Liang, & Thayer, April 1995; Park & Allen, 1994). The purpose of the present study was threefold: 1) to provide validation information about the item attributes previously identified, 2) to identify the cognitive processes used by respondents as they worked through the assessment, and 3) to pilot the assessment of cognitive components beyond those in the 1993 NAEP science framework, to include metacognitive skills and motivation. The NAEP blocks were administered to a group of sixteen eighth-grade students, that had varied proficiency in science, who were subsequently asked to participate in a study that included a think aloud and a concurrent interview. The results obtained provided information about the validity of many of the item attributes identified earlier. The results also have clear implications for enhancing the validity of the assessment, as well as for improving instructional practices.

Conceptual Basis for the Study

Three main sources were used in the design of the present study: 1) the NAEP science framework, 2) item attributes related to item performance identified by a group of ETS researchers in a previous study, and 3) a cognitive model of problem solving. Each of these sources is briefly described below.

The NAEP Science Framework

As summarized by O'Sullivan (1995), the NAEP science framework is based on a two-fold view: that scientific knowledge should be organized in a structure that connects discrete pieces of information in a meaningful way and that science proficiency depends on a student's ability to know and integrate facts into larger concepts and themes using the tools, procedures, and reasoning processes of science. As an outgrowth of this view, the framework calls for NAEP's science assessment to include:

- performance-based tasks that probe students' abilities to use materials to make observations, perform investigations, evaluate experimental results, and apply problem-solving skills (no less than 30 percent of the assessment time), and

- constructed-response and multiple-choice questions that explore students' abilities to explain, integrate, apply, reason, plan, design, evaluate, and communicate (no more than 70 percent of the assessment time)

The core of the science framework consists of a three-by-three matrix that describes three major fields of science: earth, physical, and life; and three elements of knowing and doing science: conceptual understanding, scientific investigation, and practical reasoning. In addition to these main dimensions, the framework includes two additional categories that describe science - the nature of science (which includes technology) and the organizing themes of science (models, systems, and patterns of change). The framework can be summarized as shown in Figure 1.

Item Attributes Related to Item Performance

In a study conducted between 1993 and 1995, a group of ETS researchers identified a set of 36 item attributes related to item performance using the same booklets as those used in the present study (Yepes-Baraya & Allen, in press; Allen, Park, Liang, & Thayer, April 1995; Park and Allen, 1994). A brief description of the attributes identified is provided in Figure 2. Five clusters emerged from the 36 attributes identified: 1) content knowledge, 2) reasoning and explaining, 3) hypothesis formulation and testing, 4) processing figural information, and 5) item format and reading difficulty. Content knowledge pertains to items for which knowledge of facts, concepts (including science vocabulary), principles, or procedures can be used to answer an item. Reasoning and explaining refers to items requiring some form of deductive or inductive reasoning to answer the item. Items in the third cluster 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 may facilitate or impede answering the item. Attributes in clusters 1-3 are performance attributes, while attributes in clusters 4-5 are item attributes. Performance attributes refer to the content knowledge and cognitive skills required to answer the item, whereas item attributes refer to surface characteristics of the item that may, nevertheless, affect item performance (Tatsuoka, 1994).

A Cognitive Model of Problem Solving

A cognitive model which takes into account the main components of problem solving ability (Sugrue, Fall 1995)--in particular those that can be modified by instruction-- was modified to include the attributes related to item performance already identified. As shown in Figure 3, the model has three main components: knowledge and information processing, metacognitive skills, and motivation. Knowledge and information processing includes the five main clusters identified in the item attributes study. The first three clusters--content knowledge, reasoning and explaining, and hypotheses testing and formulation--are measures of knowledge structure (Sugrue, August 1994), i.e., the degree to which an individual's knowledge reflects the conceptual organization and ways of thinking of a discipline. The remaining two clusters--processing figural information, and item format and reading difficulty--are measures of an individual's perceptual skills and ability to interpret complex verbal and nonverbal cues. The second component, metacognitive skills, refers to the self-regulatory skills of planning and monitoring, including the ability to solve a problem or complete a task within a specified time. The third component of the model is motivation, the ability and willingness of the individual to engage in a task or problem situation and persist until its completion. Motivation involves measures of the individual's perceived ability to do the task, the perceived task difficulty, and the perceived attraction to the task. Additionally, a correlate measure of test anxiety was included.

Methodology

The participants in the study were sixteen eighth-grade students in a Central New Jersey suburban middle school. They were selected based on their overall performance in the Metropolitan Achievement Tests (MAT). Five of the participants were above average performers (H = High; the percentile range for these students was 89-99), nine were average (M = Middle; the percentile range for these students was 55-67), and two were below average performers (L = Low; the percentile range for these students was 30-40). Because of time constraints, participants did not do all three blocks as in the NAEP science assessment. Instead, they completed two blocks of items: a hands-on task and either a conceptual/problem solving block or a theme block. As shown in Figure 4, depending on the types of blocks to be completed, participants were assigned to one of four subgroups. The time allowed for the completion of each block was 30 minutes, the same time as for the NAEP science assessment.

Prior to the test administration, the participants were told that the focus of the study was "to better understand their thought processes as they answered the items in the assessment, NOT to determine if they got the right answer or not." Participants, however, were instructed to "work to the best of their ability." They were promised a gift certificate, to be redeemed at a local science store, if they completed the test and a one-hour interview after the test. Following completion of the two blocks of items, individual interviews were scheduled over a period of several weeks. Additionally, the participants' science instructors were interviewed to determine how much of the content in the assessment had been covered in class and to determine the instructors' teaching and assessment practices.

Prior to the interview, participants were reminded that the focus of the study was "to better understand their thought processes as they answered the items in the assessment, NOT to determine if they got the right answer or not." It was made clear that their performance in the assessment would not be discussed at any time during or after the interview, and they would not be identified by name when discussing the results of the study with their teachers or anyone else. At the beginning of the interview, with the responses to each item covered, the investigator asked the participant to "read each item aloud as if you were taking the test for the first time, and talk aloud to yourself as you think about and produce the answer, as if I were not here." After the participant produced what the investigator considered a complete answer, the investigator asked the participant to state how sure he/she was of the response given, and coded the participant's response according to the perceived-ability scale described in Figure 5. As with other scales used in this study, the perceived-ability scale was not shown to the participant. Instead, the choices represented by the points in the scale were verbally presented to the participant after the completion of each item, and the participant's choice was registered. The reason for following this procedure was to eliminate from the communication with the participant any references to quantitative scoring and/or evaluation measures.

Once the participant's perceived ability to answer an item was determined, the investigator shared with the participant the participant's original answer to the item. Depending on the participants's answer to the item during the think aloud, his/her perceived ability, and the original answer, the investigator had the opportunity to probe for additional information. There were at least four types of situations that merited probing deeper. The first type involved a participant who was able to quickly select one of the options when answering a multiple-choice item, without providing much information as to the basis for his/her selection. In this situation, the investigator asked the participant to explain why each of the other options was not chosen. The second situation occurred when a participant was not "very sure" of the answer given. Whenever this happened, the participant was asked to explain why he/she was not sure. The third type involved discrepancies between a participant's answer during the think aloud and the original answer. In this situation, participants were asked to explain the discrepancies. The fourth situation occurred when additional data was needed to validate one or several of the item attributes previously identified.

The procedures described above were intended to: 1) put the participant at ease during the interview and 2) minimize disturbances of the participant's cognitive processes, the effect of memory errors, and distortions due to interpretation by the participant (van Someren, Barnard, & Sandberg, 1994).

Following completion of the think aloud, the interview focused on assessing the other components of the participant's motivation (perceived task difficulty and perceived task attraction), by following the procedures and using the scales described in Figure 5. All interviews were audiorecorded and selectively transcribed.

Results

Information About the Validity of the Item Attributes

The data gathered in the process of conducting the think aloud and concurrent interviews provided validation evidence for most of the item attributes previously identified.

As stated earlier, the attributes were originally identified by a team of researchers that included science experts, cognitive scientists, and psychometricians. Although the science field test was pilot tested with several groups of students, no systematic effort had been made until now to elicit information from students regarding their thought processes as they worked on the assessment. A given attribute was considered validated for a particular item in the assessment if data could be gathered from a participant's response to that item during the think aloud and concurrent interview to make inferences about the presence of the cognitive processes defined by the attribute. For example, one of the items was a constructed response item containing a diagram of a pond ecosystem that asked students to identify those animals in the pond that are omnivores, and to explain the reason for their selection. It was hypothesized before the study that this item was characterized, among others, by knowledge of facts and knowledge of concepts (Attributes 1 and 3, respectively--see Figure 2). An item is said to be characterized by Attribute 1, if knowledge of facts can be used to answer the item. Similarly, an item is said to be characterized by Attribute 3, if knowledge of concepts can be used to answer the item. In order to get full credit, at least two omnivores had to be identified and an appropriate explanation provided. Four participants' answers are presented below:

A - "The heron, turtle, fish, and frog because they eat meat and plants."

B - "The turtle because it eats water lilies."

C - "The heron and the snake because both eat plants and animals."

D - "Only the bacteria are omnivores because it (sic) breaks down animals but it can also kill plants. The other animals, I suppose, are carnivores because they don't eat plants."

During the interview B explained that omnivores "eat plants." In light of this definition, B's answer is logical. It is incorrect, however, because B's concept of "omnivore" is erroneous. D's answer, on the other hand, suggests that D's concept of "omnivore" is correct, but D's knowledge of facts about what the animals in the pond eat is inaccurate. From the evidence gathered from these four participants' responses, it can be inferred that the item is indeed characterized by Attributes 1 and 3. A full discussion of the attribute validation process is beyond the scope of this report, but it will be the subject of a forthcoming report (Yepes-Baraya, in preparation). However, some observations and generalizations are discussed below.

1. Any given item in the study is described simultaneously by a number of attributes, ranging from 8 to 16, of the 36 attributes identified.

2. Both types of attributes, performance attributes (clusters 1-3) and item attributes (clusters 4-5) were validated.

3. For the most part, those attributes deemed to be critical to answering an item correctly are performance attributes. This would suggest that, overall, the assessment is tapping those constructs it was designed to assess thus dominating the effect of item attributes that, in some cases, may be sources of construct-irrelevant variance (Messick, December 1994).

4. The examination of difficult items, i.e., items that all or all but one of the participants were unable to answer correctly, was particularly enlightening. The main reasons for item difficulty were: lack of factual knowledge, lack of conceptual understanding, lack of knowledge of principles, lack of opportunity to learn, reading difficulty, and running out of time.

5. An example of lack of factual knowledge contributing to item difficulty was a multiple choice item requiring the identification of the two most abundant gases in the Earth's atmosphere. Although participants were able to reason convincingly during the think alouds about their choices and had plausible hypotheses, they lacked the factual knowledge to answer the item correctly.

6. An example of lack of knowledge of principles was a multiple choice item that required the explanation of the variation of air pressure with altitude. It was clear from the think aloud that participants did not have an appropriate model to answer the item correctly and that their reasoning was based on subjective and inappropriate notions of air pressure.

7. There were two sets of items for which lack of opportunity to learn contributed to item difficulty. The first set was part of an ecosystem theme block. According to the life science instructor, most of that material had not been covered at the time the participants took the test that was part of the present study. The second set were earth science items. According to both the life science and the physical science instructors, although students are exposed to some earth science content from grades 2-6, earth science at the level covered in the test is not studied until the ninth grade.

8. In a few instances, reading difficulty was identified as one of the factors contributing to item difficulty. Because the items used in the present study were part of a field test, those items found to be difficult to understand were modified or eliminated.

9. Some participants did not answer the last few items in a block. From the think alouds, it became clear that this happened in two instances: 1) when the block was speeded, and 2) when the participants lacked planning and monitoring skills--as was the case with one of the hands-on tasks.

10. As reported in the literature, (Braswell & Kupin, 1993), the distinction between multiple-choice and constructed-response items is not as clear-cut as it might appear on the surface. The think alouds made it apparent that while certain multiple-choice items encourage the mental construction of a response before the options are considered, other multiple-choice items require that each option be read before selecting the most appropriate one.

Overall Performance and the Assessment of Metacognitive Skills and Motivation

Table 1 on page 32 summarizes participants' standard scores for the assessment as well as their scores on the scales used to measure metacognitive skills and motivation. The first column on the left ranks participants by their overall standard score in the NAEP science blocks administered (presented in the fourth column). The second column ranks participants by overall performance in the Metropolitan Achievement Test, as described at the beginning of the methodology section. Column three lists the block assignments for each participant. Table 2 presents results by block assignment and booklet. Additional information is provided in Figures 6-11. The following observations and findings are noteworthy:

11. As shown in Table 1, participants' ranking in the study (by overall standard score, a weighted average of their block and task scores) is highly correlated with their overall performance on the MAT (H = high, M = middle, and L = low). The only notable exception was a high performer in the MAT who ranked in the bottom half (eleventh) in the study.

12. Table 1 and Figure 6 shows that the mean standard task score was higher than the mean standard block score. However, the variation in task scores was twice the variation in block scores. As shown in Figure 11, the correlation between task scores and block

scores is positive and moderate. However, the correlation might be stronger were it not for the variation introduced by block 21 H (see 13 and 15 below)

13. Table 2 shows the mean scores by block assignment and booklet. The highest mean scores were those of the two hands-on tasks (21 H = 62.7 and 26 H = 61.5), followed by one of the conceptual/problem solving blocks (26 C = 61.1), the two theme blocks (26 T = 56.0 and 21 T = 54.7), and the other conceptual/problem solving block (21 C = 42.9). Variation in block scores was greatest for one of the hands-on tasks (21 H = 36.7), followed by the other task and the two conceptual/problem solving blocks (26 H = 17.4, 21 C = 17.8, 26 C = 15.5), and by the two theme blocks (21 T = 9.73, 26 T = 4.35). The difference in mean scores and score variation between booklet 21 (54.8, SD = 19.6) and booklet 26 (59.4, SD = 12.5) can be explained in terms of relative greater difficulty of block 21 C and the greater score variation in block 21 H.

14. The greater relative difficulty of block 21 C can be explained in terms of 5 items (out of a total of 13) that most participants could not answer correctly. Three of the items were earth science items, the content of which had not been covered in class. Of the remaining two items, one was a physical science/life science item characterized by an emphasis on knowledge of facts and the other one was a physical science item characterized by an emphasis on hands-on activities or demonstrations requiring special equipment that, in this case, was not available to participants in science class.

15. The greater relative variability in block 21 H can be explained in terms of planning. Figure 6 shows six participants who had similar block scores (between 40 and 60). However three of these participants (6, 7, and 2) scored very high on the task and the other

three (15, 13, and 12) scored very low. Those who did poorly on the task, also scored low in planning and were unable to develop an appropriate strategy to complete the task on time.

16. The correlation between block scores and perceived ability is positive and moderately strong, as shown in Figures 7 and 11. However, for a given block score, there is considerable variation in perceived ability: some participants saw themselves more able than others. It is not clear why this happened.

17. The correlation between block scores and perceived block difficulty (PBD) is positive and weak, while the correlation between task scores and perceived task difficulty (PTD) is positive and moderately strong, as shown in Figures 8, 9 and 11. This difference may be partially explained in terms of different item types (and item attributes) in the hands-on tasks. It might be easier to assess an item's difficulty when working with items involving process skills, such as manipulation of equipment and materials, observation, and recording of data, than when working with items involving more abstract processes, such as reasoning with content or the formulation of hypotheses.

18. As expected, the correlation between the overall score and the perceived attraction to science (PAS) is positive and moderately strong, as shown in Figures 10 and 11.

19. The last column of Table 2 shows participant responses to the question, "Between the multiple-choice and constructed-response items in the assessment, which did you prefer and why?" Overall, those participants who did better in the assessment tended to prefer constructed-response items. The reasons given were "you put exactly what you mean," "I like expressing myself," and "you have more leeway." Some of these participants found that multiple-choice items can be "confusing" or "tricky." Conversely, those who expressed a preference for multiple-choice items stated that these items are "easier," "take less time to

answer," "require no writing," or "allow for the elimination of one or more options," while constructed-response items "are more difficult" and "really make you think." Research on test performance and test anxiety (Schmitt & Crocker, 1981; Crocker & Schmitt, 1987) reported by Snow (1993) indicates that highly anxious students find that multiple-choice items help them maintain attention to the task, while having to construct a response can lead to self-doubt and disrupt their thinking. Non-anxious students, on the other hand, are not adversely affected by having to construct a response. These findings may help to explain the findings in this study.

Conclusions

The present study was designed to further our understanding of the cognitive processes students use in doing the 1996 NAEP science assessment. Although admittedly the small sample size precludes making broad generalizations, the study provides in-depth observations that simultaneously extend the significance of the NAEP data, constrain the inferences that can be made from NAEP results, and suggest new avenues to enhance the usefulness of NAEP.

The combination of the think alouds and concurrent interviews following the administration of the assessment proved to be an effective way to better understand students' cognitive processes as they worked on the NAEP science field test. The administration of the test in conditions similar to those of the NAEP science assessment was deemed important in order to obtain an objective record of student performance, free of any potential disruption introduced by the think aloud and concurrent interview. The think alouds and concurrent interviews provided a window into participants' thinking with science content.

The interviews with the science instructors produced important information about the context for science learning for the participants in the study.

Evidence gathered in the process of validating the item attributes suggests that the assessment is tapping, for the most part, those constructs it was designed to assess, namely science knowledge structure, reasoning, and hypotheses formulation and testing.

A large-scale assessment is designed to provide a snapshot of what students (mostly as groups, and not as individuals) know and can do in a given subject at a certain point in time. Large-scale assessments are not designed to provide an in-depth and detailed picture about the individual participants in the context of their learning environments. The use of the think alouds and concurrent interviews with the participants in the study, as well as the interviews with their science instructors provided evidence to make inferences about the performance of the individual participants. As Mislevy (1995) points out, opportunity to learn is one of the few strong and consistent predictors of student performance in a large-scale assessment. This point was confirmed in the present study with the earth science items and also with the items that were part of the hands-on tasks.

In addition to lack of opportunity to learn, other main reasons for item difficulty associated with performance attributes were: lack of factual knowledge, lack of conceptual understanding, and lack of knowledge of principles. Future research needs to identify variables in the learning environment (and in instructional approaches) that foster or inhibit the attainment of these constructs.

The use of think alouds as research instruments suggests the possibility of innovative item formats and task formats for future assessments. Items that combine the best features of multiple-choice and constructed-response items would appear to have the potential for enhanced test validity without increasing scoring costs.

The effect of metacognitive skills, in particular planning and monitoring skills, on item performance was made evident in the present study. Although lack of opportunity to learn may be one reason for poor metacognitive skills, further research is needed to explain the large differences in metacognitive performance observed in the present study.

The effect of motivation and test anxiety on test performance was also apparent. As expected, performance on the assessment improved with increased perceived ability and perceived attraction to science, and decreased with increased perceived difficulty. Similarly, students who expressed a preference for constructed-response items (an indirect measure of average or low test anxiety) performed better than those who expressed a preference for multiple-choice items (an indirect measure of high test anxiety). However, further research is needed to refine the measures used in the present study.

Future studies should also explore how learning environments--formal and informal--impact performance on the assessment in terms of the variables employed in the present study, including metacognitive skills and motivation.

Acknowledgements

The author gratefully acknowledges Nancy L. Allen of the Division of Psychometrics and Statistics at ETS for continued support in this project and thoughtful comments during the preparation of this report. Many thanks also to Susan J. Miller for assistance with the SPSS graphs and to Jean V. Yepes for editorial assistance. 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.

Bennett, R.E., & Ward, W.C. (Eds.) (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R.E. Bennett, & W. C. Ward (Eds.) *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 167-182). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of anxiety. *Journal of Experimental Education, 55,* 201-205.

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

Mislevy, R.J. (1995). *On inferential issues arising in the California Learning Assessment System*. (MS # 94-02). Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

O'Sullivan, C. (1995). *The 1993 NAEP science field test: Hands-on tasks and test specifications*. 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.

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.

Schmitt, A.P. & Crocker, L. (1981, April). *Improving examinee performance on multiple-choice tests*. Paper presented at the American Educational Research Association, Los Angeles.

Snow, R. (1993). Construct validity and constructed-response tests. In R.E. Bennett, & W. C. Ward (Eds.) *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Sugrue, B. (August 1994). *Specifications for the design of problem-solving assessments in science*. (CSE Technical Report 387). Los Angeles: National Center for Research on Evaluation, Standards, and Student-Testing (CRESST)/University of California, Los Angeles.

Tatsuoka, K.K. (1994). Personal Communication.

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.

Yepes-Baraya, M. (in preparation). *Modeling student's thinking on a science assessment*. Princeton, NJ: Educational Testing Service.

Yepes-Baraya, M. & Allen, N.L. (In press). *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.

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.

Figure 1. NAEP Science Framework

	Fields of Science		
Cognitive Domains	Earth	Physical	Life
Conceptual Understanding			
Scientific Investigation			
Practical Reasoning			
	Nature of Science		
	Themes Models, Systems, Patterns of Change		

Figure 2. 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. Can knowledge of relationships be used to answer the item?
6. Does item have science vocabulary that must be understood to answer item?
7. Does item require info. that could have been gained through practical experience

Reasoning and explaining

8. Can reasoning from general concept/principle/law to specific conclusion be used?
9. Is tracing cause-effect from one component to another in system needed to answer item?
10. Can formal inductive reasoning be used to answer item?
11. Can application of concept/principle/idea be used to answer item?
12. Can thinking with models/analogies be used to answer item?
13. Does item require that a response be given and the response be justified?

Hypothesis formulation and testing

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

Processing figural information

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

Item format and reading difficulty

29. Is item a 4-category constructed-response item?
30. Is item a short constructed-response item?
31. Does item stem have one or more intratext referentials (e.g., it, this, these)?
32. Does item stem have one or more clauses with fronted structures?
33. Must response meet all conditions specified in stem?
34. Does item have hypotheticals/exceptions/negations that make item complex?
35. Can item be solved by choosing the odd option out?
36. Does item require only info. in item itself (incl. procedural knowledge)? [Not (s)]

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

Figure 3. A Cognitive Model for Problem Solving[1]

Knowledge and information processing skills

1. Content knowledge
2. Reasoning and explaining
3. Hypothesis formulation and testing
4. Processing figural information
5. Item format and reading difficulty

Metacognitive skills

1. Planning
2. Monitoring

Motivation

1. Perceived ability to do task
2. Perceived task difficulty
3. Perceived attraction to task

[1] Adapted from Sugrue (1995).

Figure 4. Block Assignments for the Study

Group 21 C (Conceptual/PS) Conceptual/PS Block (21 C) Hands-on Task (21 H)	**Group 26 C (Conceptual/PS)** Conceptual/PS Block (26 C) Hands-on Task (26 H)
Group 21 T (Theme) Theme Block (21 T) Hands-on Task (21 H)	**Group 26 T (Theme)** Theme Block (26 T) Hands-on Task (26 H)

Figure 5. Measures for the Assessment of Metacognitive Skills and Motivation

Metacognitive Skills

Planning and monitoring, including the ability to solve a problem or complete a task within a specified period of time, are metacognitive skills. For each hands-on task, a combined measure of these two skills was developed to include evidence that each participant: 1) worked systematically and avoided unnecessary repetition, 2) completed all steps required, and 3) completed the task. One point was added for each of these elements. The scale thus ranged from 0 to 3, with a score of 3 representing strong planning and monitoring skills. Additionally, information was elicited to better understand participants' performance on the task.

Motivation

Three variables were considered in the assessment of motivation: perceived ability, perceived task difficulty, and perceived task attraction.

Perceived ability (PA) is a measure of each participant's level of confidence that the response given to each item in a conceptual/problem solving block or theme block is correct. After completing each item, participants were asked, "How sure are you that your response is correct?" The choices were: 3 = very sure (nearly 100% sure), 2 = pretty sure (70-80% sure), and 1 = not sure (50% sure or less). A mean rating was derived by adding the individual item ratings and dividing by the total number of items. The higher the rating, the greater the participant's perceived ability.

Perceived block/task difficulty (PBD/PTD) is a measure of each participant's perceived difficulty of the block or task as a whole. After completing the think aloud, participants were asked, "How easy or difficult was the block/task?" The options were: 1 = easy, 2 = moderate, 3 = difficult. A separate rating was obtained for the block and the task. The higher the rating, the higher the perceived difficulty of the block or task.

Perceived attraction to science (PAS) is a composite measure of each participant's perceived attraction to the subject of the assessment, to science as a school subject, and to science as a possible career or occupation. The composite scale ranged from 1 = low attraction to 2 = neutral or moderate attraction to 3 = high attraction.

Test Anxiety

A correlate measure of test anxiety used was participants' preference for constructed-response items over multiple-choice items. Research on test performance and test anxiety (Schmitt & Crocker, 1981; Crocker & Schmitt, 1987) reported by Snow (1993) indicates that highly anxious students prefer multiple-choice items because they help them maintain attention to the task, while having to construct a response can disrupt their thinking. Non-anxious students, on the other hand, are not adversely affected by having to construct a response.

Figure 6

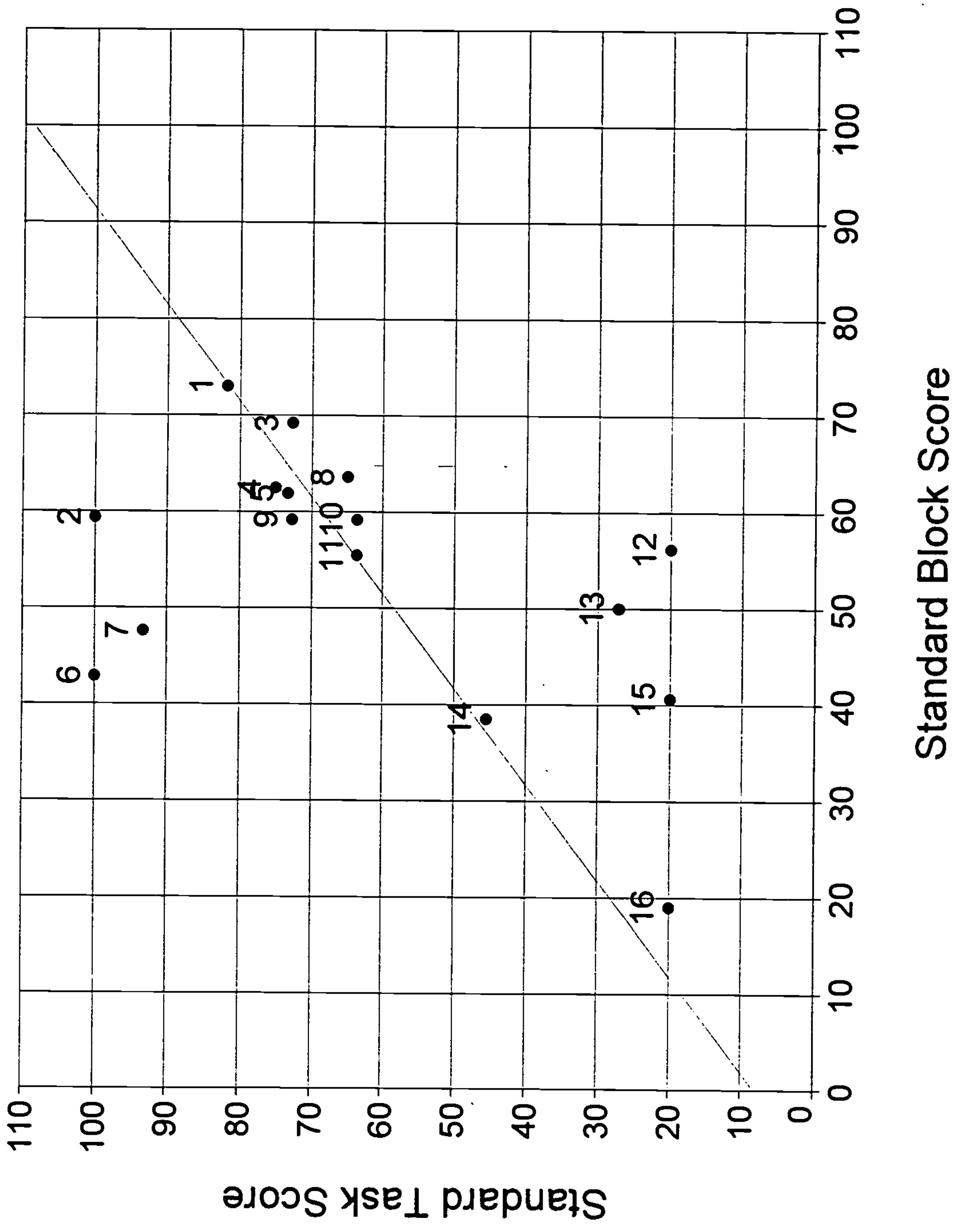
Standard Task Score
0
10
20
30
40
50
60
70
80
90
100
110
Standard Block Score
0
10
20
30
40
50
60
70
80
90
100
110
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 7

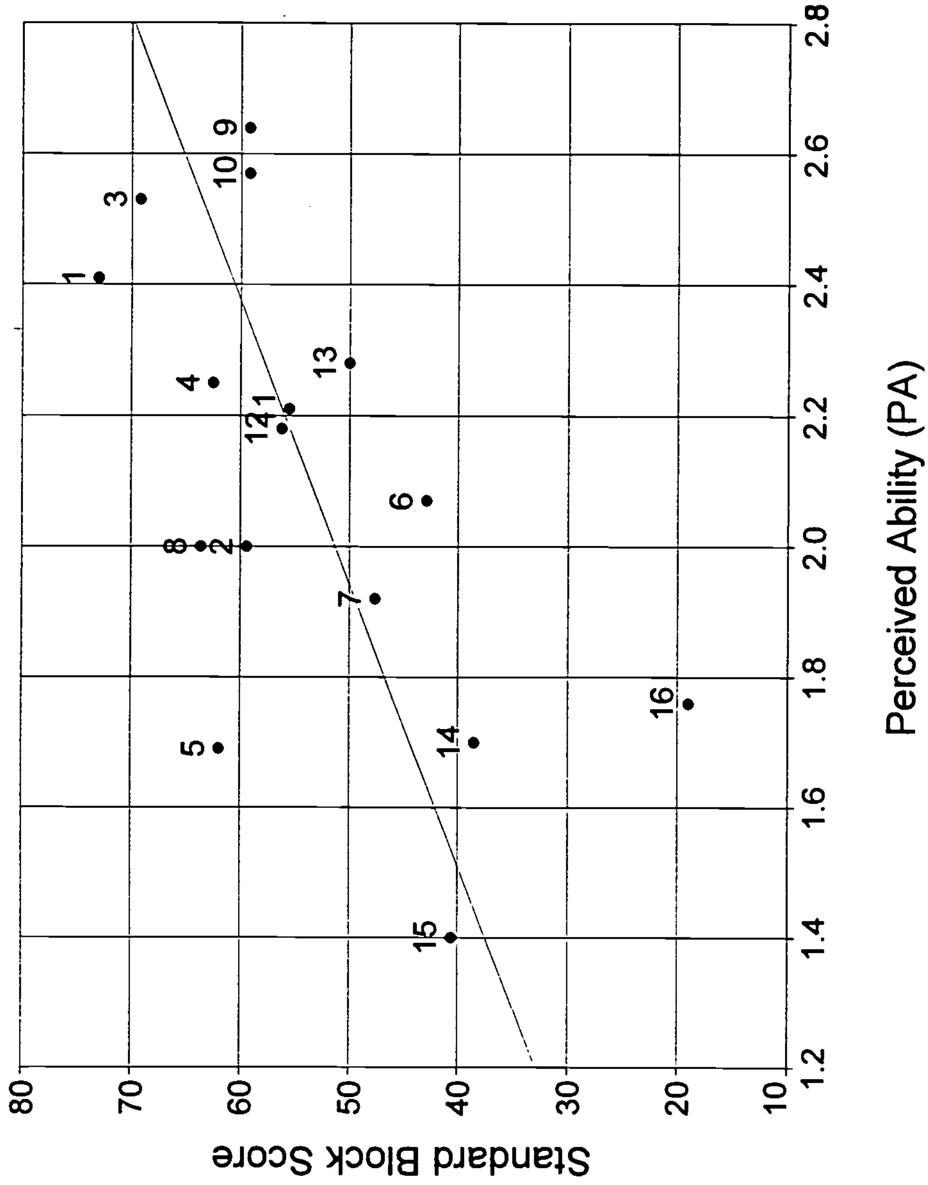
Standard Block Score
80
70
60
50
40
30
20
10
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Perceived Ability (PA)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 8

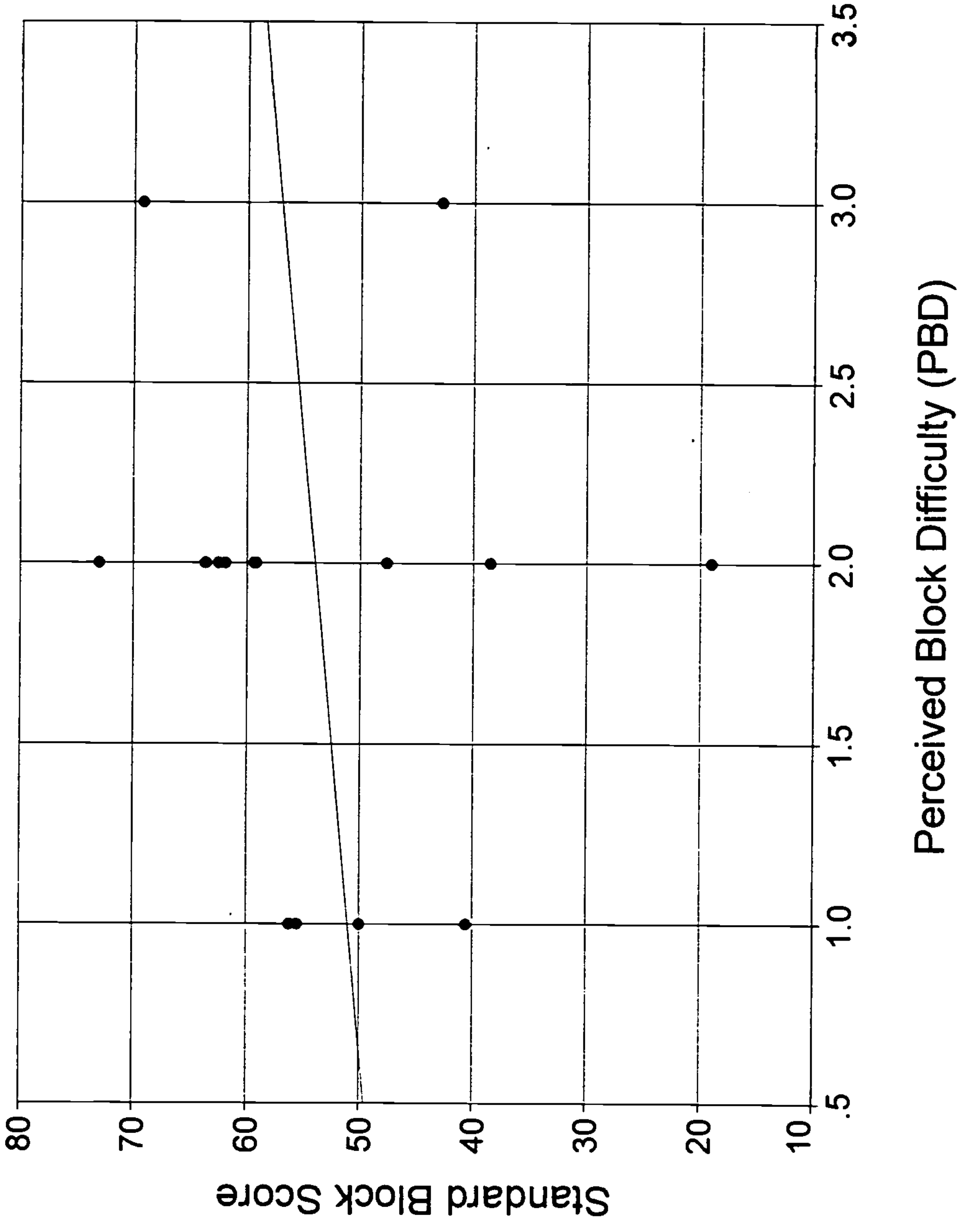
Standard Block Score
80
70
60
50
40
30
20
10
.5
1.0
1.5
2.0
2.5
3.0
3.5
Perceived Block Difficulty (PBD)

Figure 9

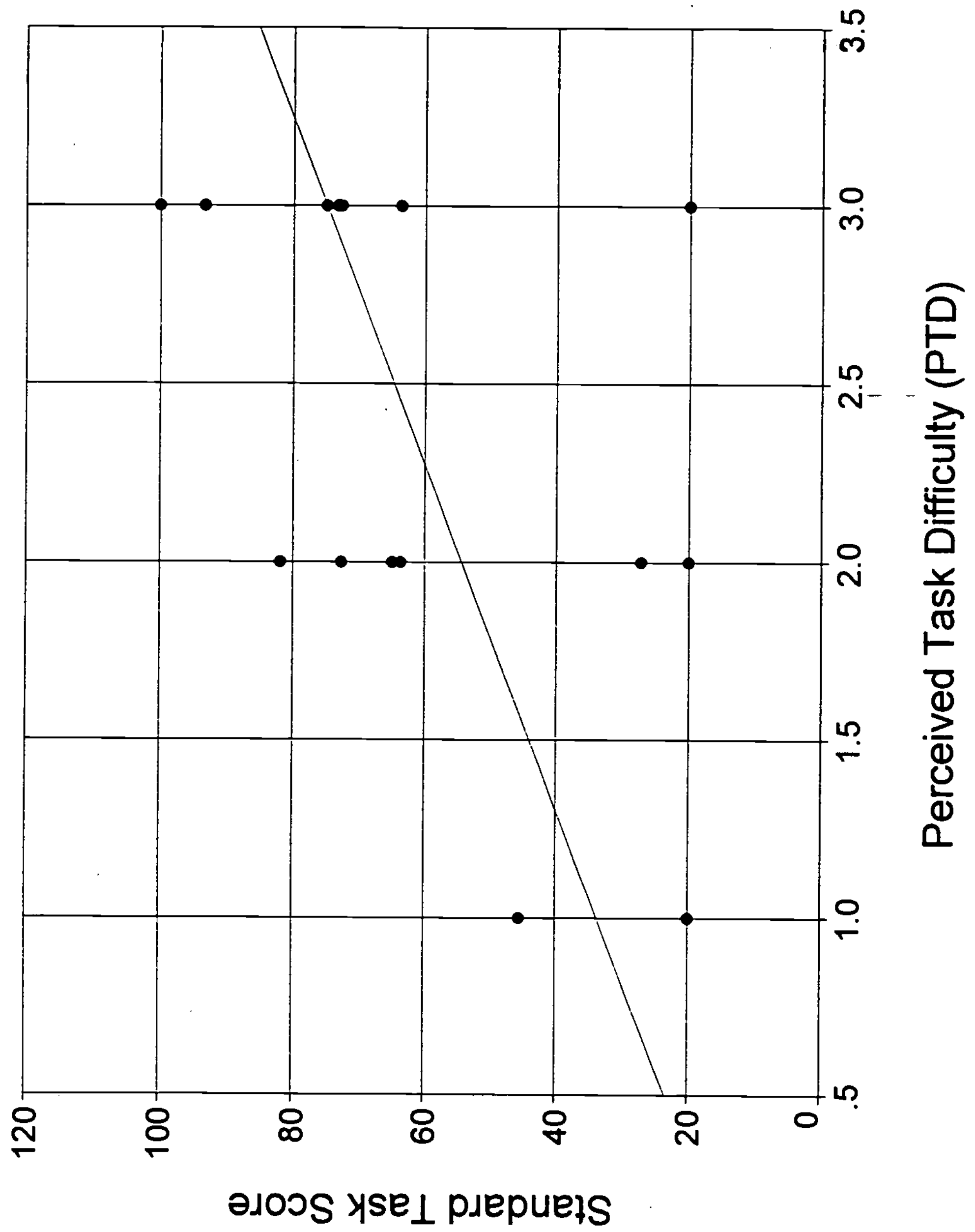
Standard Task Score
120
100
80
60
40
20
0
.5
1.0
1.5
2.0
2.5
3.0
3.5
Perceived Task Difficulty (PTD)

Figure 10

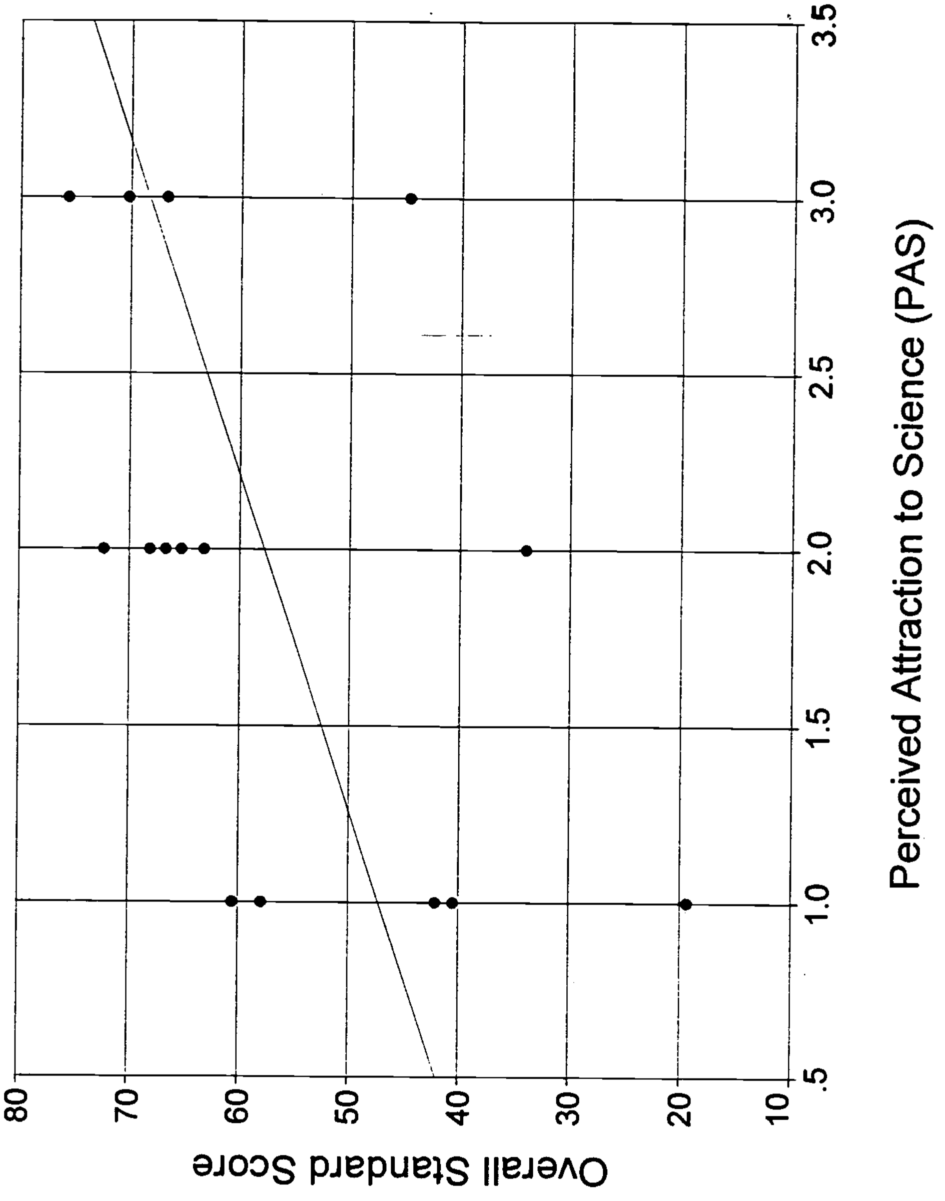
Overall Standard Score
80
70
60
50
40
30
20
10
.5
1.0
1.5
2.0
2.5
3.0
3.5
Perceived Attraction to Science (PAS)

Figure 11. Correlations

Standard Task Score and Standard Block Score

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Other	Pearson's R	.489	.189	2.100	.054[a]
	Spearman Correlation	.451	.243	1.892	.079[a]
N of Valid Cases		16			

a. Based on normal approximation

Standard Block Score and Perceived Ability (PA)

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Other	Pearson's R	.601	.112	2.817	.014[a]
	Spearman Correlation	.501	.211	2.164	.048[a]
N of Valid Cases		16			

a. Based on normal approximation

Standard Block Score and Perceived Block Difficulty (PBD)

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Other	Pearson's R	.136	.199	.515	.615[a]
	Spearman Correlation	.280	.237	1.093	.293[a]
N of Valid Cases		16			

a. Based on normal approximation

Figure 11. Correlations (Continued)

Standard Task Score and Perceived Task Difficulty (PTD)

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Other	Pearson's R	.534	.183	2.360	.033[a]
	Spearman Correlation	.541	.190	2.406	.031[a]
N of Valid Cases		16			

a. Based on normal approximation

Overall Score and Perceived Attraction to Science (PAS)

		Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Other	Pearson's R	.505	.172	2.192	.046[a]
	Spearman Correlation	.615	.170	2.922	.011[a]
N of Valid Cases		16			

a. Based on normal approximation

Table 1. Results by Cognitive Component

Rank in Study (by Overall Score)	Rank in MTA	Block Assignments	Overall Standard Score (0-100)	Block Standard Score (0-100)	Task Standard Score (0-100)	Planning & Monitoring (0-1-2-3)	Perceived Ability (PA) (1-2-3)	Perceived Block Difficulty (PBD) (1-2-3)	Perceived Task Difficulty (PTD) (1-2-3)	Perceived Attraction to Science (PAS) (1-2-3)	MC vs. CR Preference
1	M	26C	75.7	73.0	81.8	3	2.41	2	2	3	CR
2	H	21T	72.3	59.4	100.0	3	2.00	2	3	2	MC
3	H	26C	70.2	69.2	72.7	2	2.53	3	2	3	CR
4	H	21T	68.1	62.5	75.0	1	2.25	2	3	2	MC
5	M	21C	66.7	61.9	73.3	2	1.69	2	3	3	CR
5 (6)	H	21C	66.7	42.9	100.0	2	2.07	3	3	2	MC
5 (7)	M	21C	66.7	47.6	93.3	3	1.92	2	3	2	CR
8	M	26C	65.3	63.6	64.9	3	2.00	2	2	2	CR
9	M	26T	63.2	59.2	72.7	2	2.64	2	3	2	MC
10	M	26T	60.5	59.2	63.6	3	2.57	2	3	1	MC
11	H	26T	57.9	55.5	63.6	3	2.21	1	2	1	MC
12	M	21T	44.7	56.2	20.0	0	2.18	1	2	3	CR
13	L	26T	42.1	50.0	27.2	0	2.28	1	2	1	MC
14	M	26C	40.5	38.5	45.5	2	1.70	2	1	1	MC
15	M	21T	34.0	40.6	20.0	0	1.40	1	3	2	MC
16	L	21C	19.4	19.0	20.0	0	1.76	2	1	1	MC
	5H, 9M, 2L	Mean = SD =	57.1 16.1	53.6 13.5	62.1 27.8		2.10 0.35	1.87 0.61	2.38 0.72	1.94 0.77	6 CR 10 MC

Table 2. Results by Block Assignment and Booklet

	Block 21 C	**Block 21 T**	**Block 21 H**	**Booklet 21**
Mean =	42.9	54.7	62.7	54.8
SD =	17.8	9.73	36.7	19.6
N =	4	4	8	8
	Block 26 C	**Block 26 T**	**Block 26 H**	**Booklet 26**
Mean =	61.1	56.0	61.5	59.4
SD =	15.5	4.35	17.4	12.5
N =	4	4	8	8

TM025265

NCME Annual Meeting, April 9-11, 1996

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: A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment

Author(s): Mario Yepes-Baraya, Ph.D.

Corporate Source: Educational Testing Service

Publication Date: April 1996

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.

☒ **Sample sticker to be affixed to document**

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"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 □

or here

Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

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:	Position: Associate Research Scientist
Printed Name: Mario Yepes-Baraya	Organization: Educational Testing Service
Address: Educational Testing Service Mail Stop: 08-R Rosedale Road Princeton, NJ 08648	Telephone Number: (609) 734-5357 Date: April 8, 1996

OVER