DOCUMENT RESUME

ED 376 202 TM 022 311

AUTHOR Hamilton, Laura S.
TITLE Validating Hands-On Science Assessments through an Investigation of Response Processes.
SPONS AGENCY National Science Foundation, Washington, D.C.; Rand Corp., Santa Monica, CA. Inst. for Education and Training.
PUB DATE Apr 94
CONTRACT MDR-9154406
NOTE 45p.; 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 *Cognitive Processes; *Educational Assessment; *Elementary School Students; Grade 6; Intermediate Grades; Measurement Techniques; Protocol Analysis; Science Education; Student Attitudes; *Student Reaction; Test Construction; *Test Validity; Thinking Skills
IDENTIFIERS *Hands on Science; Large Scale Assessment; *Performance Based Evaluation

ABSTRACT

Many current efforts to develop large-scale science assessments involve hands-on tasks because of their presumed power to elicit and measure scientific reasoning skills. An analysis of the processes in which students engage while responding to such assessment is needed in order to discover the specific forms of reasoning that tasks elicit. This paper describes a framework that organizes the cognitive demands that assessment tasks place on students. The framework is applied to a set of science tasks completed by 20 sixth-grade students using think-aloud protocols, observations, and interviews. This procedure revealed several ways in which tasks required skills not anticipated by the test developers and provided a richer understanding of what successful performance entailed. Three tables and one figure present study findings. (Contains 32 references.) (Author/SLD)

Validating Hands-On Science Assessments Through an Investigation of Response Processes

Laura S. Hamilton

Stanford University

Paper presented at the 1994 Annual Meeting of the American Educational Research Association, New Orleans

This research was supported by National Science Foundation Grant MDR-9154406 and by the RAND Institute on Education and Training. The tasks were developed by researchers at RAND. I am indebted to Dr. Stephen Klein and Dr. Brian Stecher for the invaluable assistance and guidance they provided throughout the process of conducting this research. I would also like to thank Dr. Richard Snow for his comments on an earlier draft, and Andrea Buell for her assistance in collecting data.

Abstract

Many current efforts to develop large-scale science assessments involve hands-on tasks because of their presumed power to elicit and measure scientific reasoning skills. An analysis of the processes in which students engage while responding to such assessments is needed in order to discover the specific forms of reasoning that tasks elicit. This paper describes a framework that organizes the cognitive demands that assessment tasks place on students. The framework is applied to a set of science tasks using think-aloud protocols, observations, and interviews. This procedure revealed several ways in which tasks required skills not anticipated by the test developers, and provided a richer understanding of what successful performance entailed.

Validating Hands-on Science Assessments Through an Investigation of Response Processes

Current proposals for science education reform stress the need for assessments that measure scientific reasoning skills and competence in applying techniques used by scientists in their daily work. These efforts frequently involve the use of hands-on performance exercises because such tasks presumably measure students' use of scientific inquiry in ways that are more meaningful than traditional paper-and-pencil tests permit (Shavelson, Carey, & Webb, 1990). But do they? To be valid measures of scientific reasoning, scores on these tasks must reflect how well students engage in scientific processes, such as observation and inference (California State Department of Education, 1990). It is also important to ensure that skills that are not the focus of measurement do not influence performance to a great degree. The validation of performance tasks requires an investigation of the processes in which students engage while responding to the tasks, so that specific task demands can be identified. This paper describes a framework for studying these processes and illustrates the application of this framework to several hands-on science tasks.

Because measurement involves generalizing from a task to a construct or domain, it is often desirable to develop multiple tasks that assess the same constructs. One approach by which this can be accomplished is through the construction of a "task shell," a set of specifications that describes the format of the task, the skills to be tested, and the criteria for judging the adequacy of responses. Even when tasks are carefully constructed from well-specified task shells, however, there may be subtle differences in their characteristics that prevent them from assessing the same constructs. As with all performance measures, these characteristics may not be apparent through an inspection of scores or written responses. If tasks from a shell are to be used interchangeably, there needs to be evidence that they elicit the same processes and place similar cognitive demands on students.

The use of complex, performance-based tasks requires that traditional psychometric methods of validity assessment be supplemented with analyses from a cognitive psychological perspective (Snow & Lohman,

1989). In particular, statistical methods such as factor analysis or correlations with other measures sometimes fail to reveal sources of psychological similarities and differences among items and measures (Snow, 1993). Specific features of a task, such as the need to manipulate equipment or write extended essays, may result in an instrument that measures skills not anticipated by the test developer. Such construct-irrelevant variance (Messick, 1989) calls into question the meaning of scores awarded to students.

In addition to identifying sources of irrelevant variance, it is critical that test developers assure adequate coverage of the domain they wish to test. The validation of performance tests therefore must involve an investigation of the ways in which students demonstrate mastery of the relevant skills and understandings. Linn, Baker, and Dunbar (1991) advocate a broad conception of test validation, which includes this type of process analysis. This method of validation is especially important to consider when the domain being assessed is a process-rich one such as scientific inquiry.

The complexity of performance tasks makes interpretations of scores difficult. Mehrens (1992) notes that it is often impossible to make inferences from low scores on performance tasks because they do not indicate the source of the student's error. If a student fails to produce an adequate answer, it is often difficult to determine which of the many component skills is missing, or whether some task feature that is irrelevant to science is preventing the student from performing to the best of his or her ability (Messick, 1992). These critical features may not be evident from an inspection of the task or from analyses of students' written responses. An understanding of specific task demands and how these affect responses may shed light on how low as well as high scores are acquired.

An examination of the thought processes and strategies students use to formulate their answers can provide a more complete understanding of what hands-on tests actually measure. Several factors that are important to consider when studying the processes elicited by hands-on science tasks have been identified. The following discussion presents a framework that outlines these key constructs.

Framework for Studying Cognitive Demands of Performance Tasks

A review of the literature and an analysis of student responses to a set of hands-on tasks identified six broad categories of cognitive demands that tasks place on students. Depending upon the purpose of the assessment, these task requirements will be relatively more or less construct-relevant. However, their power to influence performance implies that each should be investigated as part of the validation process. Although the categories are not mutually exclusive, and many other categorizations could be offered, this way of organizing task demands proved useful when applied to the tasks used in this study, and it appeared to encompass the major components of hands-on science tasks in general. The framework is outlined in Table 1, and each of the six components is discussed below.

Insert Table 1 about here

Demands on Working Memory

Unwanted score variance may result from task demands that are developmentally too advanced for the student, and one way in which this can occur is when a task places unreasonable demands on the student's attentional resources, or working memory (Case, 1984). For example, the ability to attend to variables is often important in solving science tasks, and is a skill that develops gradually, partially as a result of an increase in working memory capacity. Because children have less working memory available for problem solving than do adults the requirement to keep track of numerous steps and concepts may hinder students' performance. Even when a student performs well on several subtests of a task, the integration of material from these components may require much of his or her limited attention and may result in less efficient performance than that exhibited on the separate subtests. In addition, capacity of working memory is not only influenced by developmental level but also by prior knowledge and experience (Case, 1985). The introduction of novel content places especially heavy demands on working memory and may result in suboptimal performance on

tasks that require reasoning skills (Chipman, 1986). Practice and instruction in skills necessary for task performance often reduce working memory requirements by producing an increase in students' ability to perform tasks quickly and with little demand on memory (Glaser, Lesgold, & Lajoie, 1985). Analyses of student responses to tasks can provide important information regarding the effects of experience on tasks' working memory requirements and on scores.

Use of Language and Communication

Most performance tasks require some reading, writing, or both, and a careful analysis of these language requirements is needed to interpret student performance. Because the ability to communicate is essential for participation in scientific communities, evidence of students' communication skills contributes to an understanding of their scientific proficiency. At the same time, language requirements may prevent students from demonstrating other important scientific capabilities. The introduction of new terms or of familiar terms used in novel ways may lead to misinterpretation and, consequently, poor task performance. Confusion about language may be most likely to arise when words have different meanings in everyday-world and scientific contexts (Linn & Songer, 1991). Studies of expert and novice performance (e.g., Larkin, McDermott, Simon, & Simon, 1980) have found that experts translate verbal statements into a language that facilitates problem solving. Children, lacking experience with problem solving activities, often look for a few key words to inform their solution attempts, even though this method tends to be unreliable (Greeno & Simon, 1988).

A related issue is the effect on performance of requiring written responses to questions. Students who are uncomfortable with written communication may be unable to demonstrate the extent of their scientific understanding. Baxter, Shavelson, Goldman, and Pine (1992) suggest that inexperienced students' inability to record scientific steps in a clear manner may lower interrater reliability for scorers of hands-on science notebooks compared to direct observations. The influence of writing requirements on task performance must be

investigated, especially when a specific form of writing, such as the recording of steps in a scientific experiment, is involved.

Metacognitive Skills

Metacognition refers to one's awareness of and control over his or her cognitive activities (Royer, Cisero, and Carlo, 1993). Complex performance tasks often require the application of metacognitive skills such as planning and monitoring one's actions. Students who are successful problem solvers generally set clear goals, reflect on their strategies, and adjust strategy use when appropriate (Campione, Brown, & Cc ell, 1988). They are also likely to consider all relevant aspects of a problem and to take time to plan their solution strategies (Sternberg, 1984). The ability to set goals and adapt methods to achieve them is often called strategic knowledge, which experts tend to display in greater amounts than novices (Greeno & Simon, 1988; Larkin, et al., 1980). Students are likely to demonstrate varying levels of competence in this area depending upon the content and procedures required by different tasks.

Efficiency in use of time is an example of a skill that contributes to performance, especially when tasks impose time limits. One's level of expertise in a given domain influences this efficiency. For example, excessive time spent looking back at previous parts of a task may indicate a lack of expertise (Anderson, 1987). The equipment provided to students also influences efficiency in use of time; equipment that is difficult to manipulate or that requires many parts to be identified or assembled may result in too little time spent attending to substantive aspects of the task. It is important to note that although students may have acquired competence in skills such as planning or efficiency of time use, these skills will be of little benefit unless students recognize the importance of applying them to the assessment si'uation.

Application of Prior Knowledge and Expectations

The knowledge and expectations students bring to the task situation influence their performance in ways that may or may not be

relevant to what the assessment was designed to measure. Assessment tasks are rarely "content-free" and therefore generally require the application of domain-specific knowledge. This is true even for tests that claim to assess reasoning skills. Prior knowledge and experience in a given context influence students' abilities to use higher-order processes such as problem solving. It is difficult to interpret results from tests of reasoning because students approach such tests with different levels of prior knowledge, which inevitably influences their responses (Chipman, 1986).

Although the above discussion suggests that possessing knowledge about a domain facilitates the application of reasoning skills, students often bring to the task expectations or experiences that may hinder problem solving efforts. As children grow and attempt to make sense of what goes on around them, they often develop misconceptions based on observations and generalizations (Levin, Siegler, & Druyan, 1990). Di Gennaro, Picciarelli, Schirinzi, and Bilancia (1992) refer to the knowledge acquired through everyday experiences and observations as "incidental science knowledge" and describe ways in which it can impede learning. For example, when confronted with findings that are contradictory to expectations, students may consider these as irrelevant to the situation at hand or as incorrect. Students frequently alter new, contradictory information to fit it into everyday conceptions rather than changing these conceptions to fit the new observations (Eylon & Linn, 1988). Students may even contort evidence at hand to make it consistent with expectations (Linn & Songer, 1991). One contradictory experience is therefore unlikely to alter intuitions if they are strong and are supported by everyday experience.

Siegler (1984) provides the example of an observed correlation between weight and falling fast, and notes that students are likely to use this information in formulating rules about the speed of falling objects. In general, students are more likely to recognize a causal effect where none was expected than to recognize the lack of a causal effect where one was expected (Kuhn, Schauble, & Garcia-Mila, 1992). Prior knowledge or expectations also tend to influence what hypotheses are offered and what evidence is sought (Klahr & Dunbar, 1988).

Acquisition of New Knowledge

A prerequisite skill for many performance tasks is the ability to assimilate new information, and the degree to which a student has developed this skill may influence performance substantially. Because many performance tasks are designed to test reasoning rather than content, often they introduce new concepts in order to assure that everyone approaches the task with the same content knowledge. For example, a task that requires the student to conduct a laboratory experiment might contain a section that describes each piece of equipment and its purpose. As discussed earlier, knowledge acquisition is often influenced by prior knowledge and misconceptions. Students tend not to integrate new information into general principles, especially if they believe the new information applies only to an isolated case (Linn & Songer, 1991). If many new concepts are introduced at once, students may have trouble distinguishing relevant from irrelevant information and thus may not know what aspects of the new material should be heeded (Di Gennaro, et al., 1992). In addition, students must understand that the task involves some learning; otherwise, they may view it as a collection of components that do not relate to each other, and may fail to apply what is learned in the introductory material to the remainder of the task. Such students will perform poorly even if they possess the necessary skills. Therefore the need to apply knowledge acquisition skills and the extent to which students are aware of this need should be investigated.

Use of Scientific Processes

Many hands-on tasks require students to apply scientific processes to newly encountered problems. Accurate interpretion of test performance requires an identification of the processes required and their relevance to the constructs being measured. Tasks that involve experiments generally require students to attend to variables, to interpret data, and to infer and predict from results. Often, students' performance may reveal failure to apply correct scientific procedures even when the required activities are developmentally appropriate. Students who lack experience with scientific investigations often

exhibit signs of misunderstanding of the proper way to conduct an experiment. Klahr and Dunbar (1988) note the prevalence of a "positive test strategy," which refers to a tendency to seek confirming evidence and to fail to look for disconfirming evidence. They also tend to accept all scientific findings as true rather than as fallible, and generally lack the scientist's understanding of the need for consistency in experiments (Eylon & Linn, 1988). This can be revealed through careless measurement techniques or through failure to hold variables constant. Because students are typically unaware of the variation that occurs in most phenomena (Shayer, 1986), or of the concept of measurement error, they may attribute excessive significance to small differences in results. The nature of the equipment used in experiments may enhance or hinder efforts to conduct accurate measurements. The extent to which performance tasks require students to apply specific processes may influence how dependent performance is on prior instruction.

Additional scientific processes are assessed through activities that do not involve experiments. Classification skills represent one such category of processes. Familiarity with the objects being classified influences the strategies students use to sort the objects, the quality of their solutions, and the extent to which perceptual or conceptual features are invoked (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Caracciolo, Moderato, & Perini, 1988). The manner in which the activity is described also affects responses. For example, Greene (1991) found that children are better able to construct representations of a hierarchical classification when text material is presented in a top-down rather than a bottom-up fashion; that is, when superordinate categories are described first. This suggests that the construction of a hierarchy may be facilitated by starting with superordinate categories, and that the text presented to students influences the approaches they take. It is difficult to separate process requirements of interest from requirements resulting from use of a content domain or a specific mode of presentation when interpreting test performance. Cognitive analyses of tasks can shed some light on this.

Summary

The six general categories of task demands discussed above provide a framework around which validity investigations may be structured. Methods that reveal the presence of these demands are needed to supplement traditional validity investigations. In the following sections, a methodology that was applied to a set of performance tasks is described, and examples of findings relevant to the framework are presented as evidence of the utility of this method for studying test validity.

Method

Subjects

Hands-on tasks were administered individually to 20 sixth-grade students. The student sample included students randomly selected from two classrooms in a school located in a large urban district, and children of employees at the institution where the research took place. The sample included seven girls and thirteen boys; students varied in terms of ethnicity, socioeconomic status, and previous experience with hands-on science. Two experimenters conducted the interviews and observations, and four of the students were observed by both experimenters simultaneously.

Procedures

The tasks used in this study were developed for a larger study of the feasibility of using hands-on tasks in large-scale assessment programs. Under standard conditions, they are administered in classrooms, and students' written responses are scored. In order to obtain a rich description of how students approached each task, we administered them individually and collected data on response processes using three methods: think-aloud protocols, structured observations, and post-test interviews. The think-aloud method has been used extensively by cognitive psychologists and is recommended because it results in little interference with task requirements and captures students' thoughts while they are still in short-term memory (Ericsson & Simon, 1984; Greeno & Simon, 1988). When applied to tests, it can

reveal instances in which students do not understand the instructions as intended, points at which the test assumes knowledge not previously specified, and ays in which misconceptions can lead to correct or incorrect answers (Norris, 1989). Students received the following instructions:

> Today you'll be asked to do some science explorations and to answer some questions about them. We're really interested in finding out what students are thinking about when they work on these activities. So while you're working, I'd like you to think out loud. That means you'll say in words everything that you're thinking and doing.

Students were then given an example of how to "think aloud" and asked if they understood what to do. The time limit for each task administered under standard conditions was 25 minutes. Students in this study were given one half hour to work on each task; this allowed time to think aloud but still imposed a clear time limit. Although students were permitted to work beyond the stated time limit as needed, most finished within the half hour.

While students worked on the tasks, observations of their activities were recorded. The observation sheets were highly structured but provided space for the recording of unusual approaches. The time spent on each step of the task was also recorded. Two investigators collected observations and protocols, and a measure of interobserver agreement was obtained by calculating the number of recorded events that matched on forms from four students who were observed by both experimenters. Agreement was 97%.

Students were interviewed briefly after each task, which allowed them to explain their reasoning in greater detail. These interviews were especially important for obtaining information from those students who were uncomfortable with the think-aloud procedure. The interviews also elicited students' opinions about the tasks, including which aspects of the task were difficult, confusing, and interesting.

Tasks

Two tasks from each of two shells were administered. Pendulum and Lever, developed out of one shell, require students to conduct an experiment, observe the effects of two independent variables on one dependent variable, and respond to inference questions based on their results. Table 2 shows the inference questions for both tasks. On Pendulum, students construct four pendulums using strings of two lengths and sets of washers of two different weights, and must determine which variable affects the time needed for the pendulum to swing 20 times. Students record their results on a data sheet and may refer to this to answer the inference questions. Similarly, on Lever, the length of the bar and the location of the fulcrum are varied and students conduct an experiment to determine which of these two variables affects the lever's lifting ability.

Insert Table 2 about here

Two other tasks were developed from a two-way cross-classification shell: Animals Classification (AC) and Materials Classification (MC). The first part of each task is a tutorial that provides examples of two-way cross-classifications using pictures of people; AC and MC have identical tutorial sections. On the second part of the task, the student is given eight animals (e.g., tiger, seal) or eight materials (e.g., pine cone, seaweed) and is asked to form a two-way cross-classification. Figure 1 illustrates a completed cross-classification table. The student chooses the variables by which to classify and is given a large placemat to use for sorting. He or she must fill in a table with letters corresponding to the objects in each box, and must label properties, groups, and boxes. The last question asks the student to remove a ninth object from an envelope and to classify it according to his or her own system.

Insert Figure 1 about here

Students completed tasks from a common shell in counterbalanced order, generally separated by two or three days. Because of time constraints, four of the twenty students were unable to take all four tasks.

Results for Inference Shell

The application of the methodology described above produced a great deal of useful information regarding this set of tasks, and suggested some general considerations for task developers. This section and the next present findings for each of the two shells. The specific ways in which the tasks demand skills from each category of the framework are discussed, with examples from student protocols. In addition, where appropriate, differences between tasks from a common shell are presented.

Because both inference tasks require the student to conduct an experiment and to answer questions based on the results obtained, they are potentially good sources of information regarding students' use of certain scientific processes as well as some of the other requirements discussed earlier. Seventeen students took both tasks; nine Pendulum first and eight Lever first. In addition, one took only Pendulum and two took only Lever.

Demands on Working Memory

Both Pendulum and Lever require students to attend to two variables and to determine which one influences the outcome of an experiment. According to Case (1985), this requirement is developmentally appropriate for sixth-graders. The data sheet permits students to record all results as soon as they are obtained, eliminating the need to retain results in memory. Therefore, these tasks appear to present few problems in terms of working memory requirements.

Student responses supported this hypothesis. In fact, many students were able to identify the important variable without referring to their data sheets. On Lever, for example, although only six students had complete and accurate data sheets, fourteen demonstrated an understanding that the location of the notch was the important variable

and that the length of the bar did not matter. Most of these students stated that they "remembered it from doing the experiment," indicating the effects were salient enough that retention was not a problem. In addition, all but two students on Lever and one on Pendulum mentioned both variables while verbalizing their thoughts during the inference questions. Thus, these task requirements appear to be appropriate for students at this developmental level.

Language and Communication

Although attempts are usually made during task development to keep vocabulary requirements minimal, most assessment tasks present some terms with which some students are unfamiliar. This proved to be the case with the tasks discussed here. Students were also asked to read written explanations and to follow sequential instructions for conducting the experiments, which made reading comprehension a necessary skill for successful performance. Because all answers were recorded in the test booklet, the tasks introduced a writing requirement as well.

Several students failed to follow the instructions to construct the complete set of pendulums and levers and to fill in the data sheets; for example, three students responded as though the list of pendulums (labeled A, B, C, and D) represented response options for a multiple-choice item and circled one pendulum. In addition, the need to record explanations in written form presented a challenge to many students; most were able to reveal their thoughts verbally much more clearly than in writing, providing better evidence of their understandings or misconceptions. For example, the written response of one student indicated that she had observed an effect opposite that which should have occurred, but her verbalizations revealed that her observations were correct and that she was confusing the concepts of distance and speed:

> (Written response) When the string is shorter it moves slower than when the string is longer.
> (Verbal response) When I worked it, the shorter one always made the washers go not as far as it did the long one. It doesn't go as far, so it's slower. And when you have the

longer one, it's longer so it goes farther, cause it doesn't matter the weight, just the length of the string.

As this example illustrates, written responses do not always capture students' reasoning about the effects they observe.

Pendulum appeared to present little difficulty in terms of vocabulary: Only two students mentioned vocabulary problems, and both involved only one word, "pendulum." On Lever, on the other hand, thirteen students expressed confusion over vocabulary. These instances involved the terms "notch" and "wooden stop," which serve as labels for parts of the bars and which are described both in text and in a diagram. This difficulty hindered performance: Eight students, when asked to measure from the end with the hook to the notch, measured to the wooden stop instead. The data sheets of these students reflected this inaccuracy. On the last question, which asked students to discuss the importance of the location of the wooden stop (the piece of wood that holds the washers, used to lift the levers, in place), seven students mistook either the triangular base or the notch for the wooden stop and were unable to answer the question correctly.

The requirement to record the fraction of the bar lifting weight seemed to be the primary source of difficulty in this task, particularly because it assumed an understanding of "notch." 8 students named this as the most confusing part of the task; only 7 students actually carried out this measurement accurately. The students who measured from the end of the bar to the wooden stop rather than to the notch expressed uncertainty when asked to decide whether the fraction was 1/2 or 1/4. For example, one student produced the following verbalizations:

> OK, measure the length from the end with the hook to the notch. The notch? This must be the notch (pointed to wooden stop). So that's 23.5. Fraction of the total length. What do they mean by that? Fraction, OK, total length of the bar. The end with the hook is here, and the end with the notch is here. One half or one quarter? It looks like more than one half and definitely more than one quarter. Should I put neither? So the total length is 24.5. So the fraction would

be 23.5 out of 24.5, so that's 23.5 of the bar. The fraction of the bar is 23.5.

Most students who faced this dilemma said they guessed on this item. Although all students reported having learned fractions in school, these results reveal the confusion that may arise when a term is used in an unfamiliar manner. Students who could not calculate the fraction had no accurate data sheets from which to answer the inference questions. Their performance on the remainder of the task was therefore hindered by this vocabulary requirement.

This discussion illustrates a fundamental difference between the two tasks from the inference shell; Lever contains terms that students find difficult to interpret and apply to their procedures, whereas Pendulum appears to lack this requirement. Because the two tasks require somewhat different skills, the extent to which they can be used interchangeably may be limited.

Metacognitive Skills

Extended tasks frequently require students to plan a solution attempt and to assess its effectiveness at various points in time. The inference tasks provided limited opportunities for examining students' planning strategies because step-by-step instructions were given. However, several instances in which students performed unnecessary actions or failed to complete steps in order were observed, and these frequently prevented students from completing the task within the designated time limit.

Inefficient use of time was observed repeatedly. Five students taking Pendulum and four taking Lever thought that they had to make the pendulum/lever that was described in the task introduction:

> (Read instructions silently, then picked up ruler and set on base.) I'm creating a lever, and I'm going to - it says to find out how the length of the bar and the location of the pivot point affect how much the lever can lift. So I'm going to try a little experiment.

This student spent four and a half minutes trying to balance the ruler using different fulcrum locations before turning to page 2 to begin the

actual task with the bars. Interestingly, four of the five students who did this on the first task also did it on the second, showing no sign of understanding that this was irrelevant to what the task required. Such experimentation might be viewed as an effective way to become familiar with the task content, but when a time limit is involved, it can hinder task performance. Another inefficient use of time was failing to complete the data sheet while collecting the data, which required students either to remember their results or to conduct the tests again. A final example of inefficiency was displayed by the students who attempted to test Lever E or Pendulum E despite its being taped to cardboard:

> OK, how many washers would it take to lift this one. So I'll set it on the base. But there's no place to put the washers. Well, over here I guess. Nope, won't fit. So maybe I can put the washers over here. But now where do I put the lead weight? I guess I can try to set it on here.

He spent several minutes trying various means of making the washers and weight stay on the bar while trying to balance it.

These examples illustrate how students may spend time in ways not anticipated by the task developer. When tasks impose time limits, such sources of inefficiency must be discovered and efforts made to reduce their occurrence. Many students fail to complete the Pendulum and Lever tasks within the given time limit when taking them under normal classroom conditions. Our observations revealed some common sources of inefficiency and how these may affect performance.

Application of Prior Knowledge and Expectations

Both inference tasks provided introductory material that described a pendulum or lever and informed the student of the purpose of the experiment. Although a student who had never worked with pendulums or levers could conceivably conduct the experiment and answer all inference questions accurately, verbalizations showed that most students had had some prior experiences that influenced their expectations and their responses.

The tendency to ignore results in favor of expectations or prior knowledge was exhibited frequently on Pendulum. Most of the students with inaccurate or incomplete data sheets mentioned using prior knowledge or general, everyday observations to answer the inference questions. These students generally did not refer to their data sheets. Expectations also influenced the performance of many students who conducted the experiment and recorded their results accurately. Of the twelve students who had reasonably complete and accurate data sheets, only four gave correct explanations. The others relied on expectations, which resulted in incorrect inference. One student who conducted the experiment and recorded all data accurately referred to the importance of weight in the inference questions. When questioned about his responses, he said:

> If it's lighter, when it goes up it takes a little longer. But when it's heavier it goes down faster. This experiment proves it, but I learned it already. It's like the law of gravity. If it's heavier it will fall faster.

The student turned to his data sheet and looked at his results, which he had not done while answering the inference questions. He expressed some confusion about the apparent discrepancy between the data sheet and his explanation, but dismissed it and decided that he was correct and must have recorded the data inaccurately.

A few students were more willing to allow the experimental data to take precedence over their expected results. For example, one student stated at the beginning, after testing one pendulum,

> I think that the more weight the pendulum the farther away it would swing at first cause once it's up there, the weight would press it down and keep it going for a while. Gravity pulls it down.

At the end of the task, after answering the inference questions correctly, he stated:

> If I hadn't have done the experiment I would have said the weight. But since I did do the experiments I knew that it was the length of the rope. So I was just proved wrong.

One student who performed well stated that he used school knowledge rather than his results to answer the questions:

> I already knew this stuff. Well, Galileo invented - well, found out that it doesn't matter how much weight it has, like cause he dropped things off buildings and they both hit the ground at the same time. So it wouldn't matter the weight.

When asked if he found the experiment useful, he stated that he already knew this and did not need to do the experiment. It is not clear from these observations what factors influence a student's tendency to rely on prior knowledge or expectations versus data, but it is evident that many students invoke expectations, possibly without realizing they are doing so.

This phenomenon was not observed on Lever; no students mentioned experiences or expectations that contradicted the scientific evidence. On the contrary, several students mentioned experiences with see-saws that they said helped them to realize that the location of the notch was important and that the length of the bar was not. It seems that students have more everyday experiences that support a correct explanation with Lever than with Pendulum, making Lever less susceptible to scientific misconceptions. This finding illustrates another way in which similar tasks differ in their demands: Many students taking Pendulum must overcome a powerful understanding of how the world works, and the results they obtain may be clouded by this understanding, whereas results obtained on Lever are supported by experience.

Acquisition of New Knowledge

Each inference task presents a brief introduction to the relevant apparatus. However, most of the information needed for answering the test questions is acquired from the experiment that the student conducts rather than from written descriptions, and this is discussed in the next section. The effects of ability to acquire new knowledge through text were more prominent with the classification tasks, and these will be discussed later.

Use of Scientific Processes

The inference tasks were designed to measure the student's ability to apply findings from an experiment to a set of inference questions. Pendulum asks students to record three measurements for each observation: time for pendulum to swing 20 times, length of string, and number of washers used. Likewise, students taking Lever collect measurements on weight needed to lift bar, length of bar, and fraction of bar lifting weight. Both tasks present students with a set of inference questions based on the data collected. To answer these questions, students must remember their results or interpret the data they recorded on the data sheet. Our observations revealed the extent to which students engaged in the desired scientific processes, and the ways in which the apparatus and the instructions given influenced the use of these processes.

The majority of students failed to attend to the accuracy and consistency of their measurements. Of the fourteen students who measured both strings on Pendulum, only four used a method that resulted in accurate measurement (that is, the string was aligned against the ruler in such a way that its actual length could be read). Most students, especially those who measured the strings without removing them from the hook, failed to line the string up with the ruler and therefore measured inaccurately. Students were consistent when they timed the pendulum's swing; only a few students failed to apply consistent criteria for when to start and stop the stopwatch. One potential source of difficulty, the complexity of using the stopwatch, did not appear to influence students' performance to a great extent. A few students expressed confusion over reading the time or pressing the buttons, but nearly all students who attempted to time each pendulum achieved fairly accurate results. Thus the objects that students used did not appear to prevent them from applying consistent measurement methods.

As described earlier, Lever asks students to record the number of washers needed to lift the weight, the length of the bar, and the fraction of the bar lifting weight. The first measurement was conducted accurately by 14 of the 19 students who took this task; the others

either added washers in bunches rather than one at a time, or did not count the washers used. Twelve students measured the bars using a method that led to consistent and accurate results: Ten picked up the ruler and measured the bar on the base, and the other two removed the bar and laid it against the ruler on the table. All of the students used one of these two methods and, unlike with Pendulum, neither method seemed to facilitate accurate measurement to a greater extent than the other. The students who did not measure the bars consistently either included the hook in some measurements and not in others (4 students) or failed to align the ruler to the bar (3 students). The third measurement, fraction of bar lifting weight, led to the most confusion and was discussed in the Language and Communication section.

These analyses reveal differences in the knowledge required to conduct accurate measurements on Pendulum and Lever. Some students who performed all measurements accurately on Pendulum, for example, did not realize that the washers on Lever had to be added one by one. Prior experiences with similar activities, along with specific features of the tasks, appear to influence the extent to which students demonstrate competence in their use of scientific processes.

In addition to measurement and data collection, the use of scientific inference is a central feature of both the Pendulum and Lever tasks. Evaluation of written responses indicates that many students collected accurate data but did not understand it or interpret it correctly. It is unclear from these responses whether students actually used their data sheets to answer the questions. Verbalizations and post-test questioning showed that many students failed to refer to their data sheets to answer the inference questions but that some of these students used the knowledge they had acquired from conducting the experiments.

On the first inference question on Pendulum, "Which two pendulums took the most time to swing 20 times?," only five students referred to their data sheets. This did not appear to predict understanding: Although four of these students answered that question correctly, only one maintained throughout the task that length of string was the important variable. When asked to explain their reasoning at the end,

only four students expressed correctly the relative importance of the two variables. Each of the remaining 14 students, even those who performed well on the inference questions, revealed some evidence of misconceptions about the effects of length and weight. For example, one student responded correctly that the pendulums with the long strings took longer to swing, but later said:

> But I think the weight of it matters more, because the string, there's no weight to it really. All it is is just string, it's not heavy at all. But the washer, it's pretty heavy, and it has a lot of weight to it.

When this student was asked if this was what his experiment showed, he answered that it was.

It is clear from these findings that many students have trouble interpreting their data. This difficulty may stem from several sources. First, due to slightly inaccurate timing methods, small differences in times existed for most students between pendulums with the same string length but different weights. One student concluded that weight must have an effect because of the time differences, even though these were .23 seconds and .07 seconds for the short and long strings, respectively. This student did not recognize the fact that his own inconsistency in starting and stopping the watch could be responsible for the difference. Lack of familiarity with the concept of measurement error is evident in such an interpretation.

Inaccurate data sheets were an additional, obvious impediment to correct inference. Four of the students did not notice the data sheet until after they had timed and measured the pendulums, and only one of these eventually filled it in accurately. Three other students recorded results from only one pendulum; these students asserted that the data sheet was designed for Pendulum A only. Again, a lack of experience with this type of scientific activity probably led to this confusion. As discussed earlier, most of the students with inaccurate or incomplete data sheets mentioned using prior knowledge or general observations to answer the inference questions but did not refer to their data sheets. A final source of error on the inference questions, and perhaps the most

pervasive, was the tendency to ignore results in favor of expectations or prior knowledge. This was discussed in a previous section.

The inference questions at the end of Lever were essentially identical to those in Pendulum. Interestingly, as discussed in a previous section, although only six students had complete and accurate data sheets, 14 demonstrated an understanding that the location of the notch was the important variable and that length did not matter. It may be that Lever is less susceptible to scientific misconceptions than is Pendulum, for the reasons discussed earlier. Another possible explanation relates to the nature of the measurements collected on the dependent variable. Because Lever involved discrete counts, students were likely to obtain identical measurements for the two levers with identical notch location. In contrast, it is highly unlikely that a student would obtain exactly the same measurement of time, a continuous variable, for any two pendulums. Students' data sheets for Lever, therefore, showed the lack of effect of the irrelevant variable more clearly than did the data sheets for Pendulum. Whatever the source of this difference, these analyses reveal that previous experiences with similar tasks, the nature of the physical apparatus, and the clarity of the instructions influence the extent to which students display correct or incorrect use of scientific processes.

Results for Classification Shell

The classification tasks, Animals and Materials, provided a good opportunity for investigating language requirements, acquisition of new knowledge, and the use of classification procedures. Because the tutorial sections were identical and the main tasks nearly identical, except for the objects classified, much of the discussion will refer to both tasks. Where applicable, differences between the tasks will be described. Sixteen students took both tasks (eight in each order), two took Materials only, and two Animals only.

Demands on Working Memory

Like the inference tasks, the classification tasks require students to attend to two variables. Unlike the Inference tasks,

neither classification task specifies what these variables are; instead, students must decide for themselves which variables to use. Activities such as these are generally developmentally appropriate, especially when familiar content is used. All students said they were familiar with many of the features of the animals and materials sorted. These tasks involve a hierarchical classification in that students must identify a variable that subsumes two levels (e.g., "big" and "small" are levels of the variable "size.") Previous research indicates that this is also developmentally appropriate for sixth-graders (e.g., Lowell, 1980).

These tasks require students to remember material learned in a tutorial section, which most likely places additional demands on working memory. The extent to which students made use of the tutorial section will be discussed in a later section. The time required for students to take the task when presented first or second provided evidence that students were retaining some of what they learned the first time, especially with regard to the tutorial. For the 16 students who completed both tasks, the average completion time for the tutorial was 11.90 minutes the first time, and 6.84 minutes the second time. Most of this difference is probably attributable to reduction in reading time; 14 students read all text material the first time, but nine of these skipped some or all the second time. The reduction in time observed for the main part of the task was not as large; students were unable to simply retrieve the answers they had produced previously because the objects were different. Nonetheless, a reduction in the number of times students turned back to the tutorial provides evidence that familiarity with the particular task format led to a reduction in the cognitive demands of the task and that students were able to retain what they had learned on one task and apply it to another.

Use of Language and Communication

The introductory section of this task introduces the concept of classification using the terms "groups," "properties," and "boxes," and these are the labels given to the blanks that students fill in. Although these are probably well-known words to most students, each has

a special meaning in this context. The tasks also require a fair amount of reading, especially in the tutorial section.

One of the primary sources of confusion for students seemed to be vocabulary. Although several examples were provided to illustrate how each term was used (e.g., "Gender is one property...Gender could be used to sort people into two groups; one group is males and the other is females."), students often expressed confusion over whether a term they were using should be called a group or a property. One student verbalized his confusion as follows:

> It says properties, but what's the properties of? These are from the sea and the ground, so I guess I should put properties 'sea' and 'ground.' But then the groups would be, what? Things from the sea - is that a group or a property? Property, I guess, so group would be rock and sand.

This student and others failed to acknowledge the need for a property to be a variable that subsumes two groups. Nine of the twelve students who were unable to cross-classify stated at some point during task completion that they were confused about the meaning of "group," "property," or both. None of the eight students who were able to cross-classify expressed such confusion. Seven students, when questioned after the task about the meaning of property, defined it as "something you own." This provides evidence that students may not have adjusted to the novel use of the term in these tasks, and that the more familiar definition may have interfered with interpretation.

Misunderstanding of the term "gender" sheds light on some incorrect responses to a seemingly simple question, "Do (people) A and B belong to the same gender group?" Several students said they were unfamiliar with this term. Although many were able to figure out what the question was asking, six students translated "gender" as "general." Of these, three answered the above question incorrectly, stating that the people "belong to the same general group." This is a small number but it indicates one way in which students may go astray when confronted with unfamiliar terms. These findings make it clear that vocabulary requirements must be addressed by task developers, even when they do not

initially appear to be excessive. It is important to discover the ways in which failure to understand terminology hinders solution attempts.

Metacognitive Skills

To complete the main part of the task, "Sorting Animals" or "Sorting Materials," students must create a two-way cross-classification with a set of eight objects and a blank two-by-two table. Because step-by-step instructions for completing the table are not given in this section, students must devise a plan for attacking the problem, and must adjust their solution strategies if their initial efforts ppear are unsuccessful. The tasks also impose a time limit, making efficient use of time essential for successful performance.

To investigate the use of planning, experimenters noted whether students sorted the objects immediately upon removal from the envelope or whether they expressed the formulation of a plan of attack. All eight of the students who formed a cross-classification set up the materials, looked, and then began either to fill in the table or to sort. Only two of the remaining twelve students used this strategy. The others either sorted the objects as soon as they were removed from the envelope or began filling in the table without looking at all of the objects. Understanding the value of planning appears to be an important skill for solving the classification tasks.

Observations of the order in which students approached these tasks provided evidence of the effects of particular strategies on the quality of solutions. The main classification tasks required students to record two properties, two groups corresponding to each property, four box labels, and the letters corresponding to the objects they placed in each box. We observed whether students chose their groups and properties before sorting the objects and naming the boxes, or vice versa. Of the eight students who created a cross-classification on one or both tasks, six labeled their properties and groups before sorting or naming boxes. Eleven of the twelve unsuccessful students sorted or labeled their boxes before attempting to fill in the property and group labels. It appears that success on this task is related to order of approach, and that considering variables by which to sort before sorting or naming specific

boxes contributes to success. Students' verbal reports lend support to this observation. Nine of the unsuccessful students who sorted or labeled boxes first expressed confusion when they attempted to fill in group and property labels. The verbalizations of a typical student follow:

> These are all ocean animals, so I'll put all the ocean animals together. And then farm animals. Then dog, that goes by himself, and tiger and elephant are both jungle animals so they're together. (Sorted animals and filled in box labels with animal names.) No, I don't know if they have the same - what to put down here, for groups. Property, they're all animals, so animals I guess. But I don't understand. What to put for groups, I mean. Think I'll skip it.

Later, when questioned about the source of her confusion, this student said she knew that her groups were right but did not understand the purpose of the labels. It is impossible to determine whether the order in which students approached the task contributed to their lack of success or whether a failure to understand some aspect of the task influenced both order of approach and success on the task. In the example above, it is clear that the student was not thinking in terms of cross-classifying but instead was attempting to solve the task by creating four separate groups. This student also showed signs of confusion over vocabulary, as discussed in the previous section, and of failure to apply what was learned in the tutorial, which will be discussed later.

Another focus of analysis was students' efficiency in using the limited time available for task completion. Most students showed some signs of inefficient use of time. As with some of the other tasks, and consistent with interview results, many students viewed the materials as objects to play with. For example, ten of the eighteen students who took Animals spent time attempting to make all eight animals stand up. Some of the animals were not designed to stand up easily; consequently, several minutes were used up by this effort. Some other inefficient uses of time were revealed, including a tendency exhibited by six students to fail to read instructions thoroughly the first time, which

forced them to reread in order to answer the questions. In general, however, the classification tasks seem to be less influenced by inefficient use of time than the inference tasks, perhaps because fewer materials were involved. An awareness of these sources of inefficiency may help to explain why some tasks are less frequently finished in the time allotted than others, and may suggest ways to mitigate this problem.

Application of Existing Knowledge and Expectations

Because instructions describing how to form a cross-classification were provided, prior experience with this kind of activity was not required for successful performance. Students did need to use their knowledge about the animals and materials they sorted in order to decide how to group them, but because they could use any groups they wanted, no specific knowledge was required.

Despite the instructions, there was evidence that students applied their own notions of how to classify to their solution attempts:

> It says groups, so I'll put them together the way they go, like ocean things and tree things, and then animal things and then ground things. We do this in school all the time, like put all the animals into groups like the ones that go in the ocean or in the woods.

Students such as this one seemed to be using previous experience with grouping objects to inform their solutions. Most students stated that they had learned some facts in school that helped them to do these tasks, such as where various animals live or which ones are mammals. Several mentioned other sources of knowledge such as visits to the beach or to Sea World. It appears that these tasks elicit scientific knowledge about animals and materials but that they may also be influenced by prior knowledge concerning classification procedures, and that this may hinder performance.

Acquisition of New Knowledge

Because this task requires a cross-classification, which may be unfamiliar to many students, it is important to discover the extent to

which students rely on and learn from the tutorial section. This section introduces the new concepts and vocabulary, and provides examples of tables for students to complete. Successful performance on the task requires the application of what is learned in this section to the main task.

Experimenters recorded when and how often students referred to this section, and also questioned students about their impressions concerning its usefulness. Seven of the eight students who formed a cross-classification turned back to the tutorial section during the solution attempt, as did five of the twelve unsuccessful students. All of these students stated that they were looking back to find out how to label groups and properties. These numbers indicate that referring to the tutorial was neither necessary nor sufficient for success on the tasks. Even so, there does seem to be a relationship between use of the tutorial and success at forming a cross-classification. Students were also less likely to refer to the tutorial when taking the second task, indicating that they retained what they learned on the first.

Perhaps more informative than whether or not students referred are their perceptions of whether the tutorial was helpful. During post-test interviews, students were asked, "Was doing the activity with the people helpful to you when you did the part with the animals (materials)?" Although 12 of the 20 students stated that it was helpful (including all eight successful students), their reasons varied. Successful students' comments focused on the tutorial as an example of how to approach the task, or referred to ways in which it defined the task:

> At first I was just going to separate them into two groups, one from the sea and one from the forest. But I figured that wasn't what it was asking me to do. I knew I had to find at least some connection, like something that would be from the sea that was also from the forest and something that wasn't from anything. I knew that cause that's what happened in all the examples. It wanted me to complete the charts the way they did in the examples.

Many students, including four who did not form a cross-classification, simply mentioned that it helped them to figure out what to write for groups and properties. Students who stated that it was not helpful

typically said that it was much easier than the main task or that it was too different. Examples of student comments follow:

> Not really helpful. I mean, it was a lot easier with the people cause they did some of the properties already for you, so you didn't have to decide all by yourself.

> No, cause it was different with the materials. You had to make your own groups and there were four different groups, not like over here where they just had two, male and female or cap and no cap. So it was completely different.

> No cause animals are a lot different from people.

Thus, many students failed to make the necessary connections between the tutorial and the main task, and seemed to consider them as two separate components. This may reflect students' prior experiences with tests that are composed of separate pieces, none of which is designed to teach a concept. Successful performance on these tasks requires both knowledge acquisition skills and the understanding of the importance of appying these skills to the testing situation.

Another possible explanation for students' failure to use the tutorial material is the fact that the tables in this section of the test could be completed by applying different strategies than those required for the main task. One of the tutorial tables had all of the labels filled in and asked students to place the people cards in the appropriate boxes. This table could be completed by referring only to the box labels and comparing them with the pictures; for example, Box 1 was labeled "male-cap;" students could look at all of the cards and find the males who were wearing caps. In fact, eight of the fifteen students who completed this table mentioned only the box labels when deciding how to sort the people cards. Only three students actually sorted the cards; the others simply looked at each card and decided which box was most appropriate. On the second table in the tutorial section, students were given a partially labeled table that had the letters corresponding

to the people written in the boxes. They were asked to finish filling in the labels and were given lists of properties, groups, and boxes to use. The labels that were already recorded allowed students who understood the concept to fill in the rest of the table without ever looking at the cards. Seven of the twenty students used this approach successfully; of these, only three were successful on the main task. Evidently, an understanding of the cross-classification concept as it was used in the tutorial was not sufficient to produce successful performance on the main task. A possible explanation may be that the tutorial never asked the students to look at the pictures and think of their own way of classifying them, and consequently students were unprepared for this task when it arose. The classification tasks require an ability to apply what was learned in a tutorial section to a novel situation and to adjust solution strategies appropriately.

Use of Scientific Processes

The primary activity in which students are involved is the construction of a two-way cross-classification. This requirement places a greater demand on students than what is typically expected of them in school classification activities, which often involve sorting objects into distinct groups. The tasks also require students to apply knowledge about the animals and materials classified, but the source of this knowledge may vary depending upon how students perceive the task demands.

Students who grouped objects without forming a cross-classification showed evidence of a tendency to categorize at different levels based on familiarity or on surface features. For example, several students put the sea animals and the birds in their own groups but did not put the mammals together. The criteria that most of these students used for categorizing were generally based on experiences such as visits to the zoo. The task instructions do not specify criteria for categorizing, and most students did not attend to the scientific properties of the objects but instead relied on surface similarities. It is likely that different results would have been obtained if students had been told to apply scientific criteria.

A related issue is students' perceptions of the classification tasks as tests of science. The extent to which students viewed these as activities relevant to science sheds some light on their understanding of what science actually involves and on their classroom experiences with scientific procedures. Twelve students stated that the tasks were good tests of science. All of these students mentioned either their relevance to science class or the fact that they were hands-on. Students who did not believe the tasks to be good tests of their scientific understanding listed the following reasons:

> Science? No, cause I think this would relate more to mathematics. The objects, yes, they're in science, but the classifying part, that's not science. We do that stuff in math.

> No, this isn't really like science, just a little bit. Because most science things they tell you about electricity and stuff like that. Not like this stuff.

> No, cause they say the same things they would say if your teacher was teaching this. They say the same thing, so they could just teach it the class. They don't really need a test like this.

These comments illustrate that students may not view the particular activities involved, including the tutorial section, as relevant to science assessment. These perceptions may influence some aspects of performance, such as the extent to which the tutorial is used or the source of knowledge that is applied, and therefore can serve as supplementary validity evidence. It is interesting to note that the three students who viewed the tasks as reflecting mathematical knowledge moreso than scientific concepts were successful on both tasks. Perhaps the kinds of skills students apply to task solution reflect their perceptions of the subject matter being tested. This points to the importance of collecting data relating to the meaningfulness of assessment tasks and the extent to which students view tasks as subject-

relevant, because these factors may influence the approaches students use to solve the tasks.

As discussed earlier, the two classification tasks were identical in their requirements except for the objects that students classified. Thus any differences in the cognitive requirements of the two tasks are attributable to these objects. Most students used the same approach on both tasks, but struggled more with Materials. Students spontaneously expressed a greater number of attributes for Animals than for Materials, and when questioned at the end about whether they had thought of any properties besides the ones they used, more students mentioned additional properties for the animals than for the materials. Greater familiarity with animals might explain this; eleven students reported having studied animals in school, whereas only five said they had studied the objects classified in the Materials task. Even though all students were familiar with all of the objects used in both tasks, the animals appeared to lend themselves to classification better, perhaps because students have more experience thinking about how animals are grouped and classified. Even when tasks in a shell differ along only one dimension, the types of knowledge and reasoning students are required to apply, and the difficulties students experience in formulating their solutions, may affect performance to different degrees.

Discussion

The analysis of process requirements described here illustrates various ways in which tasks may unexpectedly require students to apply skills that are irrelevant to what the task is designed to measure, and reveals some ways in which misconceptions and previous experiences influence task performance. The framework and data collection methods allowed us to structure our investigations of these critical features of task validity. Aspects of performance revealed by this study, such as the effect of misconceptions on task performance, may be considered construct-relevant or irrelevant, depending upon the intended purpose of the assessment. In either case, it is important that such features be revealed so that task developers and users understand what their tests

are measuring. Table 3 provides a list of the framework categories and some relevant questions that can be used to structure validity investigations. These questions are worded in a general manner but can be adapted to fit various science assessment situations.

Insert Table 3 about here

Several limitations of the validation method proposed here should be noted. Perhaps the most apparent limitation is that it is time-consuming and labor-intensive. Clearly, feasibility and cost would prevent test developers from interviewing large and representative samples of students, making it unlikely that all of the many ways in which task demands could influence performance would be identified. However, even a small number of interviews can serve as a means of confirming or disconfirming the developer's hypotheses as to what tasks require students to do.

An additional limitation is the possibility that the method itself may alter student performance. The requirement to take a test individually while talking aloud with an interviewer nearby creates a substantially different context than when the same tests are administered in a classroom. Many factors, such as the desire to make a good impression, may influence the approaches students take and the quality of their solutions. Despite these dangers, it is likely that the demands a task places on students do not change to a large extent and that this method of validation can provide useful information about assessment tasks.

A related concern is the validity of the think-aloud method as a means of data collection. It has been argued that individuals do not report accurately when asked to describe their thought processes (for a discussion of these arguments, see Ericcson & Simon, 1984). This may be especially true for children who are uncomfortable with the procedure. However, when students are asked to report only what they are currently thinking and doing and are not required to interpret or explain during task performance, the information acquired is likely to be an accurate

record of thought processes and approaches (again, see Ericcson & Simon, 1984). The success of the think-aloud method in other settings makes it a natural one to be applied to test validation, and when used in combination with observations and interviews, it can be a valuable tool.

Despite these limitations, the study reported here provides evidence of the usefulness of think-aloud protocols, structured observations, and interviews for test validation. These methods revealed numerous ways in which task demands such as vocabulary interpretation influenced performance. They also provided a clearer picture of what successful task performance involved; for example, the use of planning and the order of approach on the classification tasks. This kind of rich description of student performance is essential if we are to assert that our tasks elicit the use of specific reasoning processes. These descriptions may also provide clues as to which aspects of tasks are influenced by prior instruction.

The value of this method is especially apparent when tasks from a common shell are compared. This study revealed that even when tasks are developed to measure the same constructs, specific features may cause them to place slightly different demands on students. These features may not be apparent through inspection of the task or through psychometric data collected after administration.

The framework used in this study imposes a structure on the information collected through protocols, observations, and interviews. Although the tasks used in this study were pilot tested as a means of assessing their administrative feasibility, it was not until this structured methodology was applied that some of their problems as well as their strengths became apparent. By organizing the questions we ask about tasks in terms of the six categories described earlier, we can better understand the specific skills and sources of knowledge that tasks require students to apply. During early phases of task development, such information can suggest ways in which tasks might be revised to serve their intended purposes more effectively. Once large-scale data have been collected, information acquired from this method can provide explanations for results obtained from psychometric analyses. Its primary contribution, however, is its power to aid test

developers in their efforts to specify what skills are needed, what constructs are being measured by their tests, and how scores should be interpreted. Prudent use of performance assessments demands that test developers gather several types of validity evidence. The methods described in this paper can be effectively used to supplement traditional validation procedures and to shed light on the meaning of scores obtained from performance assessments.

References

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. Psychological Review, 94, 192-210.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1-17.

California State Department of Education (1990). Science Framework for California Public Schools. Sacramento: Author.

Campione, J. C., Brown, A. L., & Connell, M. L. (1988). Metacognition: On the importance of understanding what you are doing. In R. I. Charles & E. A. Silver (Eds.), The teaching and assessing of mathematical problem solving (pp. 93-114). Reston, VA: Erlbaum.

Caracciolo, E., Moderato, P., & Perini, S. (1988). Analysis of some concrete-operational tasks from an interbehavioral standpoint. Journal of Experimental Child Psychology, 46, 391-405.

Case, R. (1984). The process of stage transition: A neo-Piagetian view. In R. Sternberg (Ed.), Mechanisms of cognitive development (pp. 19-44). New York: Freeman.

Case, R. (1985). Intellectual development: Birth to adulthood. Orlando: Academic Press.

Chipman, S. (1986). What is meant by "higher-order cognitive skills" (ED 279 668). ERIC Report.

Di Gennaro, M., Picciarelli, V., Schirinzi, D., & Bilancia, L. (1992). Incidental science knowledge in fifth grade children: A study of its relationship with cognitive development and cognitive style. Research in Science and Technological Education, 10, 117-126.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Eylon, B., & Linn, M. C. (1988). Learning and instruction: An examination of four research perspectives in science education. Review of Educational Research, 58, 251-301.

Glaser, R., Lesgold, A., & Lajoie, S. (1985). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. Glover, J. C. Conoley, & J. C. Witt (Eds.), The influence of cognitive

psychology on testing and measurement (pp. 41-85). Hillsdale, NJ: Erlbaum.

Greene, T. R. (1991). Text manipulations influence children's understanding of class inclusion hierarchies. Journal of Experimental Child Psychology, 52, 354-374.

Greeno, J. G., & Simon, H. A. (1988). Problem solving and reasoning. In R. C. Atkinson (Ed.), Steven's handbook of experimental psychology (pp. 589-672). New York: Wiley.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.

Kuhn, D., Schauble, L, & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. Cognition and Instruction, 9, 285-327.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. Science, 208, 1335-1342.

Levin, I., Siegler, R. S., & Druyan, S. (1990). Misconceptions about motion: Development and training effects. Child Development, 61, 1544-1557.

Linn, M. C., & Songer, N. B. (1991). Cognitive and conceptual change in adolescence. American Journal of Education, 99, 379-417.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 3-9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13-104). New York: Macmillan.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments (RR-92-39). Princeton, NJ: Educational Testing Service.

Norris, S. P. (1989). Can we test validly for critical thinking? Educational Researcher, 18(9), 21-26.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382-439.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. Review of Educational Research, 63, 201-243.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. Phi Delta Kappan, 71, 692-697.

Shayer, M. (1986). Data processing and science investigation in schools. Research papers in education, 1, 237-253.

Siegler, R. S. (1984). Mechanisms of cognitive growth: Variation and selection. In R. J. Sternberg (Ed.), Mechanisms of cognitive development (pp. 141-162). New York: Freeman.

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (pp. 263-331). New York: Macmillan.

Sternberg, R. J. (1984). Mechanisms of cognitive development: A componential approach. In R. J. Sternberg (Ed.), Mechanisms of cognitive development (pp. 163-186). New York: Freeman.

Table 1

Framework for Organizing Demands of Hands-on Science Tasks

1. Demands on working memory
2. Use of language and communication
3. Metacognitive skills
4. Application of prior knowledge and expectations
5. Acquisition of new knowledge
6. Use of scientific processes

Table 2

Inference Questions

PENDULUM:	LEVER:
1. Which two pendulums took the most time to swing 20 times?	1. Which two levers needed the most washers to lift the weight?
2. Dale says the weight of the pendulum has the biggest effect on how fast it swings. Pat says the length of the string is more important. Who is right? Explain your answer.	2. Chris says the length of a bar has the biggest effect on its ability to lift objects. Jody says the location of the notch is more important. Who is right? Explain your answer.
3. Look at Pendulum E on the cardboard. How much time would it take Pendulum E to swing 20 times?	3. Look at Bar E on the cardboard. How many washers will it take to lift the weight with this bar?

Table 3

Questions to Guide Validity Investigations

DEMANDS ON WORKING MEMORY

What is the developmental level required by the task?
Are students required to keep track of numerous steps or concepts?
What is the effect of experience with similar activities?
What is the effect of prior knowledge or familiarity with task content?

USE OF LANGUAGE AND COMMUNICATION

How extensive are reading requirements?
Is new vocabulary introduced?
Are familiar terms used in unfamiliar ways?
Do writing requirements influence expression of subject-matter understanding?
Is a special kind of writing required (e.g., the recording of steps in a scientific experiment)?

METACOGNITIVE SKILL DEMANDS

How does each of the following influence performance?
- Planning
- Monitoring
- Goal setting
- Adjustment of strategy use
- Efficiency in use of time

Are students aware of the need to apply these skills?

APPLICATION OF PRIOR KNOWLEDGE AND EXPECTATIONS

To what extent are "reasoning" items influenced by prior knowledge?
Is the content area one with which students are likely to have had prior experience?
How do expectations and misconceptions influence responses (e.g., procedures used or hypotheses put forth)?
Are scientific facts contradicted by intuitions?

ACQUISITION OF NEW KNOWLEDGE

Does successful task performance require learning new concepts?
Is integration of new information into existing knowledge required?
How extensive are these requirements?
Are students aware of the need to acquire new information during task completion?

USE OF SCIENTIFIC PROCESSES

What scientific processes does the task require?
Does equipment facilitate the use of correct scientific processes?
Do incorrect procedures always lead to low scores?
Does prior instruction influence the use of scientific processes?

Figure 1

Two-Way Cross-Classification Table