DOCUMENT RESUME

ED 386 463 TM 023 799

AUTHOR Lomask, Michal S.; And Others
TITLE Large-Scale Science Performance Assessment in Connecticut: Challenges and Resolutions.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE 18 Apr 95
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Grade 10; High Schools; High School Students; Science Teachers; *Science Tests; *State Programs; Student Evaluation; Teacher Made Tests; *Test Construction; *Testing Programs; Test Use
IDENTIFIERS Common Core of Learning (Connecticut); *Connecticut; Connecticut Academic Performance Test; Large Scale Programs; Open Ended Questions; *Performance Based Evaluation

ABSTRACT

This paper examines several issues and challenges associated with performance-based assessment in science through experiences with two large-scale student assessment projects. The first is the Common Core of Learning (CCL) Science Assessment Project designed to explore the use of classroom-embedded performance assessment for the evaluation of students by their own teachers. The second is the Connecticut Academic Performance Testing (CAPT) program, designed as a statewide, on-demand assessment of tenth graders' knowledge of science. On-demand assessment is external to classroom evaluation; the teacher has almost no input into its design, administration, and scoring. During the 4 years of the CCL project (1989-1993), approximately 200 science teachers developed and used performance tasks in their classes, but the use of performance assessments did not spread to large numbers of teachers until the statewide CAPT came into use. A survey of 34 school science supervisors indicated that the impact of the CAPT is most apparent in the laboratory and on open-ended test questions. Two tables and seven figures illustrate these discussions. (Contains 17 references.) (SLD)

Connecticut State Department of Education
Box 2219
Hartford, CT 06145
203-566-6557

Large-Scale Science Performance Assessment in Connecticut: Challenges and Resolutions *

Michal S. Lomask
Joan Boykoff Baron
Jeffrey Greig

American Educational Research Association Annual Meeting
San Francisco, California, April 18, 1995
Part of symposium: "Learning, Instruction and Assessment in Science: Perspectives on this Ménage à Trois from Four Countries".

*The research and development conducted on the Common Core of Learning Assessment were supported, in part, by a grant from The National Science Foundation (SPA-8954692). This paper is a draft of a chapter being submitted to the *International Handbook of Science Education* (Editors: Barry J. Fraser and Kenneth Tobin).

≥

Large-Scale Science Performance Assessment in Connecticut: Challenges and Resolutions

Michal S. Lomask, Joan Boykoff Baron, Jeffrey Greig
Connecticut State Department of Education

Introduction

This paper examines several issues and challenges associated with performance-based assessment in science through the experiences of the authors with two large-scale student assessment projects. The first project is the Common Core of Learning (CCL) Science Assessment Project, designed to explore the use of *classroom-embedded* performance assessment for the purpose of student evaluation by their own teachers. The second is the Connecticut Academic Performance Testing (CAPT) program, designed as a *statewide, on-demand* assessment of tenth-grade students' knowledge of science.

What are the main differences between classroom-embedded and on-demand assessment? Classroom-embedded assessment, in general, is assessment that is an integral part of a teacher's regular classroom instruction. The teacher decides what, when and how to assess his/her students. The assessment is administered and scored by the teacher and its ultimate goal is to inform students and teachers about students' progress, for the purpose of *instructional feedback*. On-demand assessment, on the other hand, is external to classroom teaching. The individual teacher has almost no input into its design, administration and scoring. It is designed mainly for *accountability, monitoring and placement* purposes. Although classroom-embedded and on-demand assessment have different purposes, both assessments can use the same formats, such as written tests, performances, portfolios and exhibitions. Both assessment systems can provide students with opportunities for enhanced learning when *performance* tasks are part of the assessment, as they are in both the CCL and the CAPT programs.

The Common Core of Learning (CCL) Project: Classroom-Embedded Performance Assessment

In 1987, the Connecticut State Board of Education approved a new charter for education in Connecticut, summarized in a document called The Common Core of Learning. The CCL document set forth high levels of expectations for all K - 12 students in the Connecticut educational system. The document referenced content standards (e.g., the core content of earth, life and physical sciences), reasoning skills (e.g., inquiry, problem-solving and self evaluation) and habits of mind (e.g., persistence, intellectual curiosity and responsibility for one's learning). This broad spectrum of expectations created a need for a comparably broad assessment which would both stimulate and document progress toward the attainment of these goals. The challenges of building new assessments in the areas of secondary school science and mathematics were accepted by the CCL Assessment Project, established in 1989 and funded, in part, by the National Science Foundation.

Purpose of the CCL Assessment

Our main motivation for developing new assessment models was based on our interest in creating a medium through which students can document their learning and teachers can evaluate their students' progress. We envisioned assessments that provide students with opportunities to ask scientific questions, investigate the nature of scientific phenomena, and construct meaningful knowledge. Data collected through the CCL assessment were not used for individual comparisons or high-stakes state testing programs. Rather, the data were analyzed and studied for the purpose of building better assessment tasks, developing diagnostic scoring systems and learning about student performances under different conditions.

Content of the CCL Assessment

Based on earlier experiences with student performance assessment in Israel (Tamir, 1993) and Great Britain (Schofield et al, 1990) it was decided to focus the assessment on two major components:

Conceptual Integration tasks assess students' understanding and ability to **integrate the knowledge of science** concepts, theories and applications, after they have learned it in their science courses. Decisions about the specific content and levels of understanding for students in different stages of their science learning were guided

by the evolving science frameworks, developed by the American Association for the Advancement of Science (AAAS, 1989).

Inquiry Proficiency tasks assess students' ability to use prior knowledge and inquiry reasoning to **obtain new knowledge**. Knowledge of the specific content of the investigated topic undergirds the framework for any inquiry. Therefore, the assessment of inquiry proficiency cannot be devoid of conceptual understanding.

Format of the CCL Assessment Tasks

The CCL program included two types of performance assessment, one designed to assess conceptual integration and the second designed to assess inquiry proficiency. Conceptual integration was assessed through open-ended questions and scored by concept map analyses (Lomask, Baron & Greig, 1993). Inquiry proficiency was assessed through problem solving tasks (also called "event task", Hart 1994). All assessment tasks were designed to be administered and scored by teachers as part of their regular classroom teaching. Description of additional assessment models which were developed and explored during the life of the CCL project (1989-1993) are described in the monograph "Assessment As an Opportunity to Learn" (Baron, 1993).

The rest of this section will focus on the structure and function of the *problem-solving task* which served as the main format for our classroom-embedded assessment of inquiry proficiency. All CCL problem-solving tasks have the same structural shell and they contain the same five main parts, as described in Figure 1.

Figures 2 and 3 show one complete problem-solving task, *Yeast in Your Bread*, which was developed for a high school biology course. In this task, students work in groups to explore the effects of various factors on the metabolism rate of yeast. The task is grounded in the traditional biology curriculum, but it is designed to assess not only what students have learned but also how well can they use their knowledge to solve new problems. Unlike traditional lab activities, in this task students have to identify their own research questions, design relevant experiments, collect data and present valid explanations for peer review.

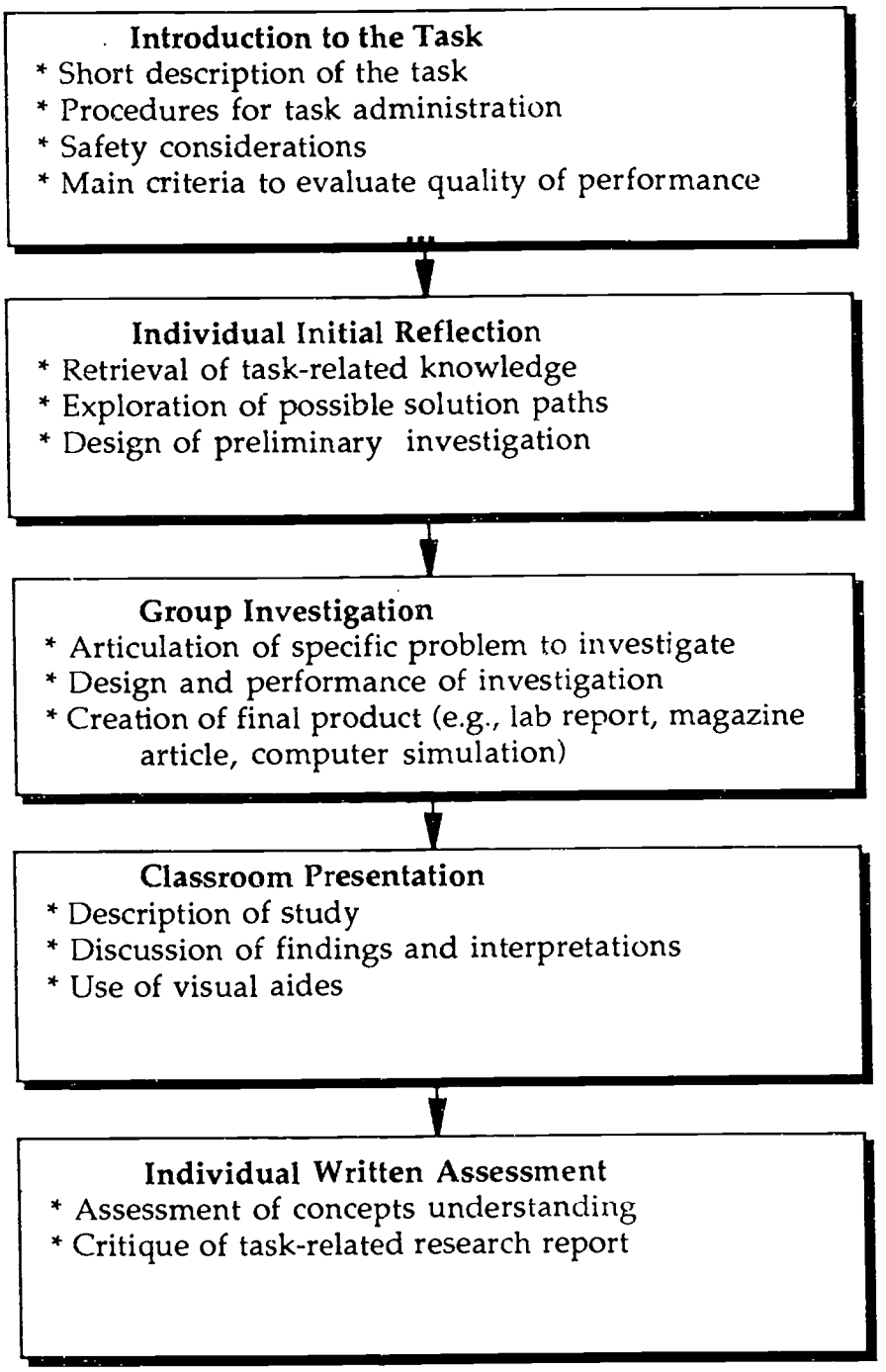


Figure 1: The Structure of the CCL Classroom-Embedded Problem-Solving Tasks

YEAST IN YOUR BREAD

Individual Initial Reflection

Suppose you decided to bake your own bread. If you were to look in a cookbook, it might tell you to add living yeast to the dough. You might ask yourself, "What are yeast? Why should I add them to my bread?" In this activity, you will have the opportunity to design and carry out an experiment to learn about yeast and the factors that affect their activity. By the end, you might understand why you have yeast in your bread.

- Knowing that yeast are living organisms, make a list of factors which might affect their rate of metabolism and predict the effect of each factor.

Group Investigation

- Choose one of the factors identified by your group and **design experiments** to study its effects on yeast activity. You will be provided with a culture of yeast in water. Adding about one spoon of sugar to 30 ml of culture will activate the yeast. Your design should be clear and complete enough so that someone else could easily repeat your experiment. Show your design to your teacher before actually carrying out your experiment.

- After getting approval from your teacher, **carry out** your experiment. **Record** all of the data your group collects in a clear and organized manner.

- What **conclusions** can be made from your experiment? **Explain** how you arrived at these conclusions and what additional experiments have to be done.

Group Presentation

- Prepare to **present** your investigation and findings to students in class. Try to **summarize** the problem, the investigation and the findings in a **clear** and interesting way. Add visuals (e.g., graphics, computer simulations) to **augment** your presentation.

Figure 2: Common Core of Learning (CCL) "Yeast in Your Bread" Assessment Task (Individual Initial Reflection and Group Investigation).

YEAST IN YOUR BREAD

Individual Assessment

1. A student in home economics is learning how to make bread, but is having difficulty getting the dough to rise. Knowing that you have just finished studying yeast activity, this student would like your help. Make a list of factors that might be affecting the activity of the yeast and explain to the student the effect of each factor on yeast activity.

2. The following report was written by a group of students working on the yeast task. Read the report and answer the questions that follow.

Our Group Report

We studied the effect of sugar on yeast activity. We made a suspension of yeast in water, and then poured it into six test tubes. We added a different amount of sugar to each test tube and let them stand for 30 minutes. Then we measured the volume of the suspension. Our results are shown below:

Amount of Sugar in Solution	Volume of Yeast Solution
1 ml	8 ml
2 ml	13 ml
3 ml	14 ml
4 ml	22 ml
5 ml	24 ml
6 ml	25 ml

Our conclusion: Sugar increases yeast activity. The more sugar you add to the yeast, the greater is its activity.

2. Draw a graph that represents the findings of the above group's study.
3. Does the above report give you enough information to replicate this experiment? If not, what would you need to known in order to perform the same experiment?
4. Do you think this group's conclusion is valid? Explain fully why or why not.

Figure 3: Common Core of Learning (CCL) "Yeast in Your Bread" Assessment Task (Individual Written Assessment).

Scoring and Evaluation of Student Performance on the CCL tasks

To assess the quality of work produced by students we used a dimensional scoring method. In this method, unlike holistic scoring, separate scores are assigned to separate dimensions of performance. The analysis and scoring of students' work on the CCL problem solving tasks were made along four dimensions of performance:

I. **Conceptual Understanding** - individual score, based on the Individual Written Assessment at the end of the activity (see Figure 3)

II. **Experimentation proficiency** - group score, based on the report of the Group Investigation (see Figure 2)

III. **Contribution to team work** - individual score, based on structured feedback from team members

IV. **Public presentation** - group score, based on the quality of each group's presentation, scored by structured feedback from the audience at the time of the presentation

The scoring guides for student performance were developed hand-in-hand with the performance tasks themselves and they articulated clearly the criteria for scoring student performance. The development of the scoring guides was informed by the analyses of science practical laboratory tests done in the first and second International Association for the Evaluation of Educational Achievement (Tamir, Doran & Chye, 1992). The guides are structured enough to allow reliable evaluation of students' work according to pre-set standards, but they are also flexible enough to allow for the consideration of different solution paths. The repeated use of a similar task format, dimensions of performance and scoring procedures, allowed students and teachers to gain familiarity with this form of assessment and created the basis for the evaluation of students' growth of understanding over time.

The involvement of human judgment in the evaluation of performance-based assessment is a source for measurement concerns. We found that the dimensional scoring method, when used by **trained** assessors, can produce reliable scores and detailed instructional feedback (Lomask, Baron and Carlyon, 1993). Table 1 reports the reliability of scores which were given to students who performed the CCL task "Exploring the MapleCopter". The data for the Conceptual Understanding dimension are based on the written work of 156 students. The Experimental Proficiency data are based on lab reports which were written by 52 groups of

students, 2-4 students in each group. The students' work was scored by two experienced science teachers who were trained to assess student performance in this specific task.

Table 1:
Estimation of Reliability of Scores of a CCL Problem-Solving Task, Based on Dimensional Scores Given to Students, by Two Independent Raters

	Variance Components			Generalizability Coefficients	
Dimension*	Students	Raters	Error	G1*	D**
Conceptual Understanding	10.213	0.000	1.943	0.84	0.91
Experimental Proficiency	5.882	0.000	1.421	0.80	0.89

* G1 is the estimated reliability for scores produced by one rater.

** D is the estimated reliability for average scores produced by two raters.

The dimensional scoring method required a high investment of time for rubric development, collection of benchmarks, assessor training and actual scoring. These data provide some encouraging evidence that dimensional scoring systems can produce reliable results if designers pay substantial attention to the details of the scoring process.

During the four years of the CCL project, approximately 200 science teachers who were part of the project developed and used performance tasks in their classes, but the use of performance assessment in science classrooms did not spread to large numbers of teachers in Connecticut. It was not until the statewide CAPT assessment that performance assessment became part of teachers' instructional practice across the state.

The Connecticut Academic Performance Test (CAPT): On-Demand, Statewide Performance Assessment

In 1993, the Connecticut State Department of Education (CSDE) implemented a statewide assessment of tenth-grade students' academic knowledge, called the Connecticut Academic Performance Testing (CAPT). The CAPT assesses approximately 30,000 public school students annually, in the areas of language arts, mathematics, and science. The CAPT is not a test of high school science content, but rather an assessment of cumulative science proficiency, addressing the content and skills students should have acquired while in grades K-10.

Purposes of the CAPT Assessment

The CAPT serves a variety of purposes. The first purpose is to establish high performance standards for all students. Students who meet or exceed the goal standards receive a certificate of mastery on their high school transcript. Students who do not meet the goals in one or more areas have the opportunity to voluntarily retake those parts of the test in subsequent years. The second purpose is to encourage schools to change their practices. Acknowledging the strong impact of state assessment on school curricula and teaching practices, the CAPT was designed to reflect local and national reform efforts. As such, the CAPT stresses conceptual understanding, skills in problem solving and the application of school-based knowledge to everyday problems. The third purpose is to provide accountability for Connecticut's education system, informing public policy makers about progress in student achievement toward a set of pre-established goal standards.

Content of the CAPT Science Assessment

The CAPT Science Framework is based on the idea that science is both a body of knowledge and a way of thinking about the world around us. Therefore the assessment emphasizes content, processes and applications in a balanced way. The CAPT science framework focuses on three aspects:

Understanding concepts and applications- addressing major concepts from the life, physical and earth/space sciences and their applications in everyday life.

Experimentation proficiency - addressing various experimentation skills, such as defining scientific problems, designing relevant and valid experiments, collecting data, finding patterns and drawing conclusions.

Scientifically-based decision making - addressing the use of scientific knowledge and logical reasoning to make informed decisions about science-related societal issues and communicate the rationale for their decisions.

Format of the CAPT Assessment

The CAPT science has two components which correspond to the first two parts of the science framework. The assessment of *concepts* and *applications* is accomplished through both written open-ended and multiple-choice items which are clustered around major theories and concepts of science. The assessment of scientific *experimentation proficiency* is accomplished through a lab performance task, followed by four task-related questions (see Figure 4 for a summary of the CAPT Science Assessment). In this part of the assessment, students design and carry out their own experiments to solve a given problem and then write about their results (see Figure 5). Students' work on the task is not collected and scored at the state level. Rather, teachers are encouraged to score the work of their own students and provide them with formative feedback. Figure 6 shows the lab follow-up questions, accompanied by distributions of student scores to each open-ended question.

A separate component, the CAPT Interdisciplinary Task, asks students to make *scientifically-based informed decisions.* In this task students are asked to read a variety of source materials (e.g., newspaper and magazine articles, editorials, political cartoons, graphs or charts) on a controversial issue and then write an extended piece in which they take and support a position on the issue

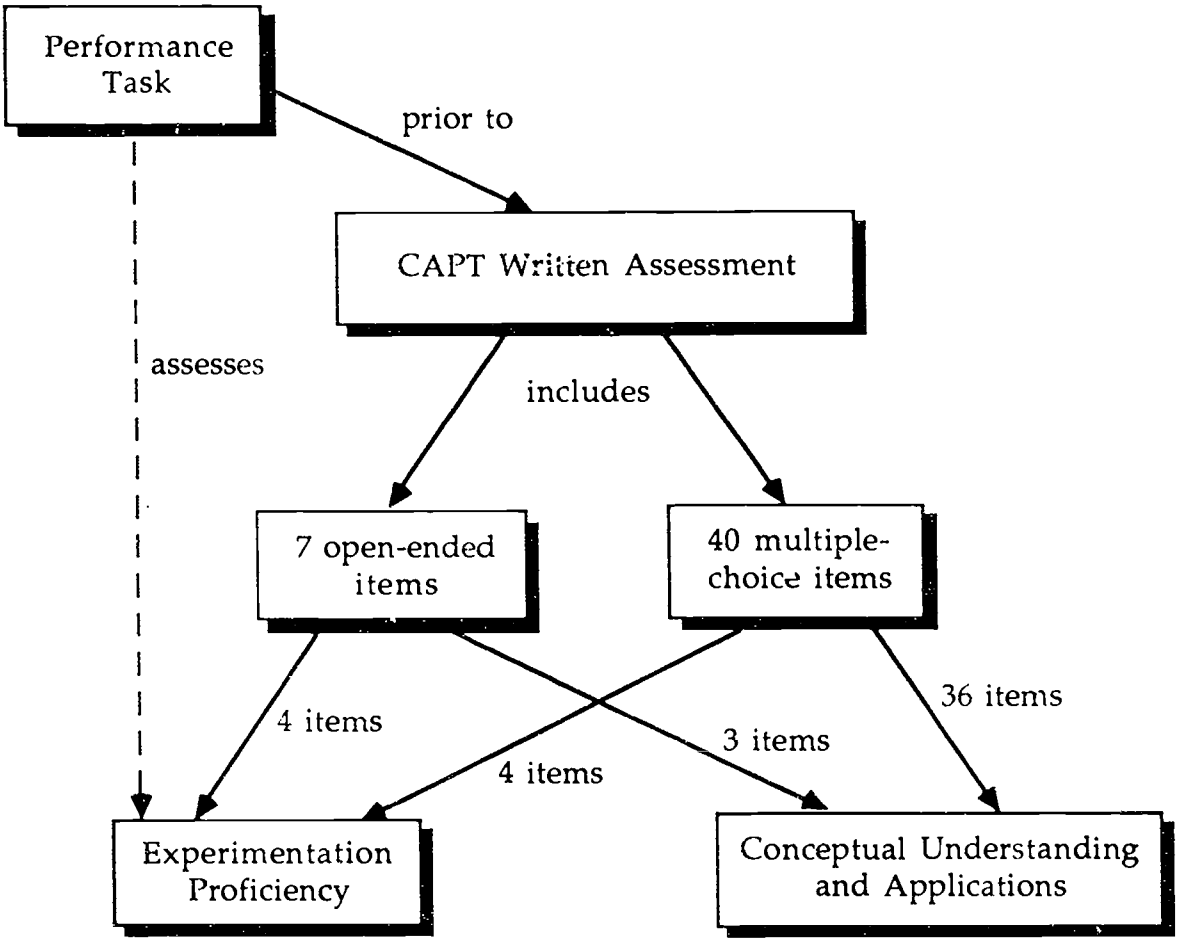


Figure 4: Structure and components of the Connecticut Academic Performance Test (CAPT) in science.

Scoring of the CAPT Science Assessment

The open-ended questions (including the four lab follow-up questions) are scored holistically on a four-point scale (0-3). The scoring process includes the development of task-specific scoring rubrics, training and qualifying of scorers and frequent calibration during scoring sessions. In this process scorers reached agreement on exact scores for 60% - 79% of the answers (depending on item) and agreement within one point score for 91% - 100% of the scored answers. The written part of the CAPT, on which students' scores are based, contains 40 multiple-choice items, each of them scored as 0 or 1. The CAPT science results provide three scores to each student: a score for conceptual understanding, a score for experimentation proficiency and a total score which is a combination of the two.

CAPT - Science 1994

SOILED AGAIN
Group Investigation

You will be investigating a problem related to acid rain. During this activity, you will work with a partner (or possibly two partners). However, you should keep your own individual lab notes because after you finish you will work independently to write a report about your investigation.

The Problem:

Acid rain refers to rain, snow or other precipitation with a pH below 5.6. In extreme cases, acid rain can have a pH as low as 2.0! Many lakes in the Northeastern United States, although often appearing crystal clear, have significant decreases in their number of fish and other life forms as a result of increasing acidity. You and your partner will design and conduct experiments to determine which earth material (sand, potting soil or limestone) or combination of earth materials best reduces the acidity of "acid rain." You will use a vinegar-and -water solution as a substitute for acid rain. You will investigate the problem by studying the percolation rate (the rate at which water seeps through a material) and neutralizing ability (the ability of a material to reduce the acidity of acids) of various earth materials.

Steps to follow:

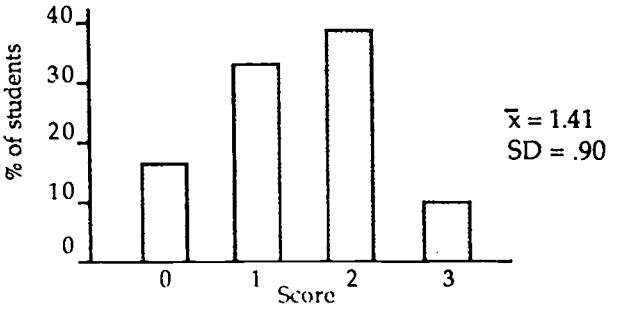
1. In your own words, state the problem you are going to investigate, and write your statement of the problem on the page provided.
2. Design one or more experiments to solve the problem. Describe your experimental designs on the page provided. Show your designs to your teacher before you begin your experiments. Remember that there are several different ways to investigate the problem.
3. After receiving approval from your teacher, work with your partner to carry out your experiments. Your teacher's approval does not necessarily mean that your teacher thinks your experiments are well designed. It simply means that in your teacher's judgment your experiments are neither dangerous nor likely to cause an unnecessary mess.
4. Use the vinegar solution as a substitute for acid rain. Use a Ph test strip to determine the acidity of the solution.
5. While conducting your experiments, take notes on your progress and record all observations and measurement data.

Figure 5: "Soiled Again" - Group Investigation.. Example from the Connecticut Academic Performance Testing (CAPT) in Science

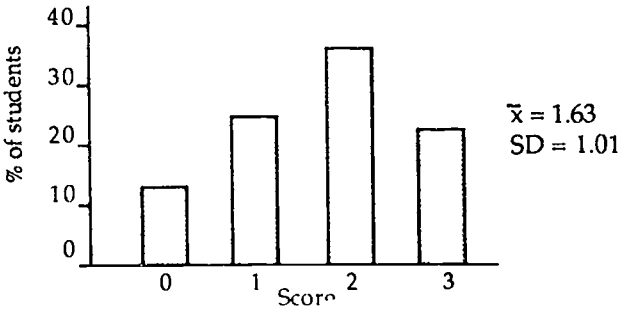
The results of the group's experiment are shown in the following table.

Earth materials	pH of "acid rain" before percolation	amount of "acid rain" percolated in 3 minutes	pH of percolated "acid rain"
sand	3.0	30 ml	3.5
potting soil	3.0	20 ml	3.5
crushed limestone	3.0	90 ml	5.0
all three earth materials	3.0	50 ml	5.5

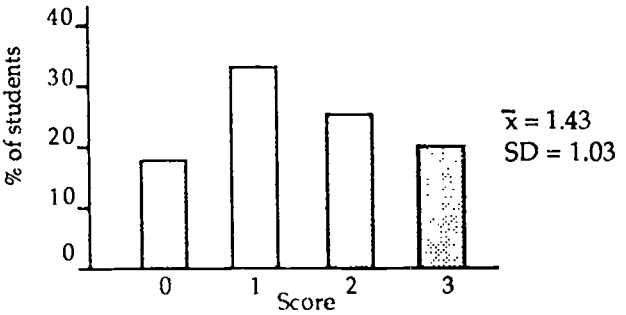
1. What is one problem that this group is investigating? State the problem in your own words.



2. What are the variables that need to be controlled in this experiment? Explain why it is important to control them.



3. Do you have enough information to replicate this group's experiment? If you think you do, tell what information you have. If you think you do not, tell what other information you would need.



4. The group concluded that sand and potting soil have the same ability to neutralize acidity because in each case the pH went from 3.0 to 3.5. Based on this group's experiment and results, do you think the group's conclusion is valid? Explain why or why not.

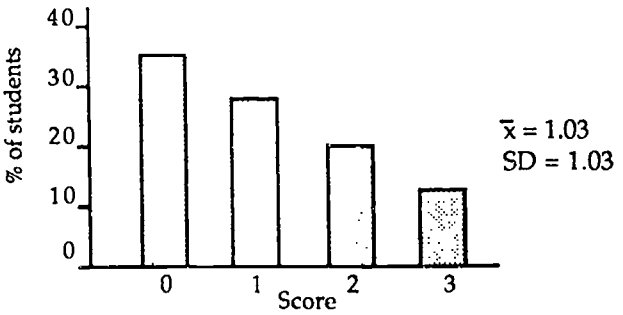


Figure 6: CAPT Experiment Follow-up Questions and Distribution of Students' Scores

Issues in Building Statewide On-Demand Performance Assessment

Written tests, using either multiple-choice or open-ended formats, have been part of the educational testing scene for a long time, and procedures for standardized and objective administration and evaluation are well established. Performance-based assessment, in which students' work on individual and group projects serves as the basis for the evaluation, is relatively new and its various components are still under study. To benefit from existing models, we decided to use the CCL problem-solving task format as the basis for the assessment of experimentation proficiency on the CAPT. We found that the use of CCL tasks (which originally were designed for low-stakes classroom-embedded assessment) in high-stakes, statewide assessment, presents many challenges to students, teachers and test developers. Some of these challenges and their resolutions in developing the CAPT program are described below:

Relevance of performance assessment content to school science curricula

Challenge - Connecticut has a long-standing tradition of local control of its education system. As a result, each school district has its own scope and sequence of science teaching and tenth-grade students across the state are exposed to different courses, curricula and textbooks. As part of the CAPT assessment, science teachers across the state are asked to administer the same performance task 2-4 weeks before the written test, regardless of the specific content they may be teaching at that time. For example, students might be asked to perform a physics task in the middle of a biology class. This has the potential to disrupt the normal course of instruction. The dilemma for us, the test developers, was how to develop assessment tasks that meet three conditions: 1) the task has authentic scientific content; 2) the content is accessible to most students, and 3) the activity can fit relatively seamlessly into a variety of different course contexts.

Resolution - Our resolution was to situate the performance task in a general science context, assuming that students' prior experience with general science (taught in most middle schools) will make the performance task both familiar and accessible. However we recognize that this solution may not fully eliminate the possible "out-of-context" nature of the CAPT performance task.

Adjustment for test administration time constraints

Challenge - The CCL tasks, which combined learning with assessment, were designed for 3-5 class periods, allowing enough time for students to reflect, converse and collaborate on the performance of these extended investigations. Due to testing time constraints, the CAPT performance task was limited to only two periods of class time (about 90 minutes). How can performance tasks be shortened, without stripping them of their meaning?

Resolution - Our resolution was to present students with more focused problems that can be investigated in a shorter time. In addition, the "floundering around" time, which is a typical characteristic of open-ended explorations was reduced by providing students with greater initial scaffolding for the task.

Scoring student performance

Challenge - Performance assessment produces extended student work, including, but not limited to experimental designs, tables of data, visual descriptions and oral presentations. In the CCL embedded assessment, teacher were responsible for the scoring and evaluation of their students' work and therefore had to go through extensive scoring training. To ensure reliable and standardized scoring of work produced by 30000 students, there is a need to train a large number of assessors in the process of evaluating complex performances and assigning consistent scores. How can we maintain a balance between the need for an efficient and reliable scoring system and the desire to encourage students to perform complex tasks (Baxter, Glaser, & Raghavan, 1993)?

Resolution - To alleviate the scoring load without reducing the meaning of the task itself, we decided to split the scoring process. Student performance during the lab activity, including their lab reports, was left to be scored by the class teachers, as in any other classroom-embedded assessment. The written individual assessment section of the task was transferred to the statewide on-demand assessment section and only this part was scored by state-trained assessors, thus reducing scoring time and costs.

Accommodating collaborative problem solving and individual accountability

Challenge - Our experience in using group problem-solving tasks in the CCL assessment reinforced our belief in the importance of students working together to design and carry out investigations, analyze and interpret data, and reach conclusions. However, the CAPT is a high stakes assessment and individual scores

are used to award certificates of mastery. How can we preserve collaborative group-problem solving while maintaining individual accountability?
Resolution - Again, the design of the CAPT assessment of experimentation proficiency attempts to balance these concerns. Students perform the laboratory activity during a five-week window prior to the administration of the written portion of the CAPT. Students work in small groups to design and carry out experiments to solve the given problem and draw conclusions based on their findings. Each student concludes this activity by writing his/her own lab report. Teachers are encouraged to score the lab report and provide each of the students with formative feedback. In this way students have enough opportunities to receive evaluation of their work prior to the written test, which might offset the group membership effect.

Striving for equity: standardization of laboratory materials and equipment
Challenge - Statewide assessment, designed for accountability and comparisons, requires standardization of test administration. The performance assessment, in addition, requires the availability of labs, equipment and materials. In the CCL embedded assessment, teachers had control over the tasks that their students performed and comparisons were not made across schools. How can we ensure that all students in large urban schools, small rural schools and mid-size suburban schools will perform the same task in comparable lab facilities and materials?
Resolution - Our resolution to this equity challenge was to have the State purchase and ship to all schools some of the equipment and materials needed for the performance of the specific task. The central purchasing and shipping of materials is costly but it contributes to basic equity among schools.

Discussion

The use of performance assessment on a large-scale testing program requires trade-offs. The science CAPT assesses approximately 30,000 students each year. The design of the assessment deals with the complex issue of alignment among curriculum, instruction, learning and assessment under fiscal and logistical constraints. The specific design of the science CAPT attempts to preserve some of the qualities of embedded assessment in a standardized statewide assessment. The use of follow-up questions does not allow for the direct measurement of students' ability to design and carry out their own scientific experiments. Rather, the ability of

students to apply scientific thinking to evaluate critically the work of others is assessed. Given the constraints of a large-scale assessment, a decision was made to assess the latter while still promoting the use of the former in classroom instruction. In a pilot study of the CAPT science program it was found that the scores of students who performed the lab task prior to the written test were significantly higher than those of students who had not performed the task (Greig, Wise and Lomask, 1994). This findings encourages us to keep the lab performance task as an integral part of the CAPT assessment, despite the logistical and financial burdens that it entails.

Recently, several measurement experts have raised concerns about the validity and reliability of large-scale performance assessment (Olson, 1995). In the remainder of this paper, we will discuss our work as it relates to these issues.

Consequences and Impact of Performance-based Assessment in Science

It is known that "what is on the test" shapes what teachers teach and what students learn. This is especially true for high stakes testing programs such as the CAPT. The use of performance assessment in which students are asked to show their knowledge and skills through "products" (e.g., models, writing, experiments, computer simulations) can help create models for good instruction and is strongly advocated by major initiatives to reform science education (The National Center for Improving Science Education see Raizen et.al., 1989, 1990 and National Research Council, 1995). The consequences of the assessment, both intended and unintended, on classroom practice must therefore be of major concern.

To learn about the initial impact of CAPT on science teaching throughout the state, we conducted a survey, asking 34 school science supervisors to express their level of agreement with eight statements about the impact of the CAPT in their schools. Table 1 shows the survey's statements and the respondents' degree of agreement with them.

Table 2: Reported Impact of CAPT on School Science

The extent to which the CAPT test impacted:	Percentage of respondents rating impact as 3.0 or greater on a 5-point scale*
The use of open-ended lab activities	81%
The use of open-ended test questions in science	79%
Professional development activities	74%
The way in which science is taught in your school	71%
The use of more integration within the science curriculum	60%
The adoption of new textbooks or resource materials	57%
More integration of science with other subjects	47%
The purchase of laboratory materials	39%

* Respondents were asked to rate the impact of CAPT on stated aspects, in the following way: 0 = "not at all"; 3 = "somewhat"; 5 = "a great deal".

The results of this survey show that after its first year of administration, the CAPT is having the most impact on the use of open-ended laboratory activities and open-ended test questions in science courses. This can be assumed to be the direct result of including a hands-on performance task and short written response questions within the structure of the CAPT. These results are encouraging. We hope that with subsequent future administrations of the CAPT, teachers will review the performance of their students and be engaged in school-wide discussions about ways to improve student learning.

The CAPT attempts to maintain a balance between the assessment of conceptual understanding and experimental proficiency. Anecdotal evidence indicates that some schools pay greater attention to the latter, shifting the focus of their science programs from "traditional content" to "exciting hands-on activities", instead of trying to strike a balance between the two within one coherent program. The danger in the rush toward performance assessment in science is that teachers might focus too narrowly on the task's activity and not enough on the conceptual structure underlying the task, thus creating a new type of fragmented, meaningless science program. Every effort should be made to prevent shallow integration of performance assessment into science instructional programs.

Measurement Issues

The stability of students' performance on different tasks is a major concern when using performance assessment. In studying hands-on science assessment activities, Shavelson, Baxter and Pine (1992) have reported high variability in student performance due to task content and format. This is problematic because the number of tasks that can be used on large-scale science assessment is generally very limited due to time and cost constraints Therefore, one of the greatest challenges in developing statewide performance assessment is creating performance tasks of equal difficulty to be used from year to year, so that growth in student achievement over time can be measured and documented. To study this issue, we compared the distribution of scores given to student work on five different CAPT performance tasks during a pilot study (see Figure 7). Results showed that the distributions of scores across different tasks was not comparable enough to be used as a stable measure of student progress from year to year. Comparisons of the average scores given to student work on different tests of lab follow-up questions indicated that this performance might be more stable and less influenced by the specific content of the task (the mean score and the standard deviation of the mean for the five performance tasks and the five sets of follow-up items respectively were M=1.15, SD=0.28; M=0.99, SD=0.15, respectively). Although the performance on the follow-up questions is relatively stable across lab tasks, it was found to be significantly affected by previous exposure to the actual task performance. Not surprisingly, on the follow-up items, students who performed the actual lab significantly outperformed those students who had not performed the lab. Furthermore, since students' CAPT scores are based on their performance on the follow-up questions rather than on their actual performance in the lab, established equating procedures can be applied to these items, particularly if these open-ended items are reported in combination with other items that assess scientific experimentation.

To ensure reliability of the scores given to student performance on the open-ended items, the CAPT focuses on the following:

- All student perform the same lab activity prior to the written test.
- The scoring criteria for the open-ended items are focused and clearly defined.
- Scorers go through extensive training and continuous calibration.

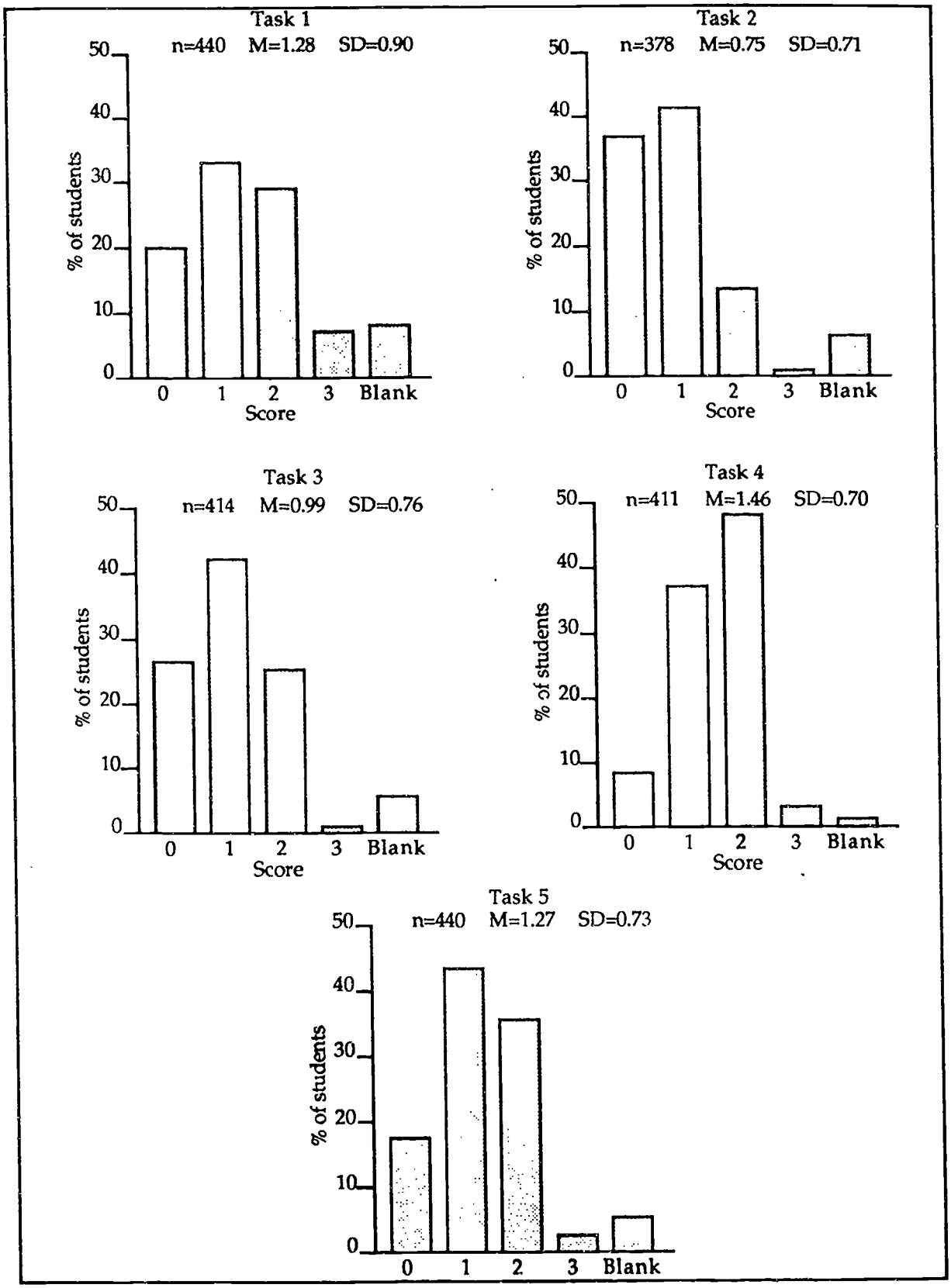


Figure 7: Frequency Distribution of Average Scores on Five pilot Performance Tasks

This system produced acceptable interrater agreement on scores (72% agreement on exact score, 95% agreement on adjacent scores) and increased our confidence in the feasibility of statewide performance assessment.

What can be done to maximize the validity of the assessment? First, the assessment framework has to match closely the curricular frameworks, so that students, teachers, parents and school administrators can see the coherence of the system and how assessment and teaching reflect each other. The current structure of the CAPT is aligned with the state curricular frameworks, and as such it has high face validity. To increase the validity of the assessment there is a need to include more points of reference to student actual work. This might be done by shifting from a single on-demand assessment to a portfolio collection of classroom-embedded assessments. Students' work on various tasks during the whole year can be collected and scored by teachers, with random audits by the State to ensure appropriate scoring procedures (Wolf and Baron, 1991). Portfolio-based assessment will eliminate the need to tightly equate tasks' difficulty and will contribute to a better alignment of curriculum, teaching and assessment. Portfolio-based assessment will require heavy investment in teacher staff development, including task development, task administration and performance scoring. This investment might be costly, but it will help build the capacity of schools to rethink their science programs and better prepare students and teachers to meet the challenges inherent in emerging national science standards.

References

American Association for the Advancement of Science (1989). *Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics and Technology*. Washington, DC: Author.

Baron, J. B., (Ed.) (1993). *Assessment As an Opportunity to Learn: The Connecticut Common Core of Learning Alternative Assessment of Secondary School Science and Mathematics*. Connecticut State Department of Education. Hartford, CT.

Baxter, G P., Glaser, R., & Raghavan, K. (1993). Analysis of Cognitive Demand in Selected Alternative Science Assessments. *Center for Research on Evaluation, Standards, and Student Testing*. University of California at Los Angeles, (October).

Connecticut State Board of Education (1987). *Common Core of Learning*. Hartford, CT: Author.

Greig, J., Wise, N., & Lomask, S. M. (1994). The Development of an Assessment of Scientific Experimentation Proficiency for Connecticut's Statewide Testing Program. *Annual Meeting of the American Educational Research Association*. New Orleans, LA, (April).

Hart, D. (1994). *Authentic Assessment: A Handbook for Educators*. Addison-Wesley, Menlo Park, CA.

Lomask, S. M., Baron, J. B., & Carlyon, E. L. (1993). Performance-Based Assessment As a Multiple Mirror: Reflections of Students in Science. *Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics*, Cornell University, Ithaca, NY, (August 1-4).

Lomask, M., Baron, J.B., & Greig, J. (1993). Assessing Conceptual Understanding in Science through the Use of Two- and Three- Dimensional Concept Maps. *Proceedings of the Third National Seminar on Misconceptions and*

Educational Strategies in Science and Mathematics, Cornell University, Ithaca, NY, (August 1-4).

National Research Council (1994). *National Science Education Standards.* National Academy Press. Washington, DC. (November).

Olson, L. (1995). The New Breed of Assessments Getting Scrutiny, *Education Week.* 1, 10-11. (March 22).

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis I. V. S., & Oakes, J. (1989). *Assessment in Elementary Science Education.* Washington, D.C. National Center for Improving Science Education.

Raizen, S.A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I. V. S. & Oakes, J. (1990). *Assessment in Science Education: The middle years.* Washington, D.C.: National Center for Improving Science Education.

Schofield, B. (Ed.), Bell, J., Black, P., Johnson, S., Murphy, P., Qualter, A., & Russel, T. (1990). *Science at Age 13: A Review of APU Survey Findings 1980-84.* Her Majesty's Stationery Office: London.

Shavelson, R. J. & Baxter, G. P. (1992). What We've Learned about Assessing Hands-on Science. *Educational Leadership,* 20-25 (May).

Tamir, P. (1993). *Student Assessment: Present and Future Trends.* Cinquième Rencontre Européenne de la Biologie. Barcelona, Spain (May 3-5).

Tamir, P., Doran, R. L, & Chye, Y. O. (1992). Practical Skills Testing in Science. *Studies in Educational Evaluation. 18,* . 263-275.

Wolf, D. & Baron, J. B. (1991). *A Realization of a National Performance-Based Assessment System.* Prepared for the Assessment Panel of the National Council for Educational Standards and Testing, Washington, DC (October 31).