DOCUMENT RESUME

ED 349 816 FL 020 635

AUTHOR Baron, Joan Boykoff
TITLE SEA Usage of Alternative Assessment: The Connecticut Experience.
PUB DATE Aug 92
NOTE 63p.; In: Focus on Evaluation and Measurement. Volumes 1 and 2. Proceedings of the National Research Symposium on Limited English Proficient Student Issues (2nd, Washington, DC, September 4-6, 1991); see FL 020 630.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Ability; Elementary Secondary Education; Limited English Speaking; *Program Descriptions; *State Programs; *Student Evaluation; *Testing
IDENTIFIERS Common Core of Learning (Connecticut); *Connecticut Assessment of Educational Progress; Connecticut Mastery Testing Program

ABSTRACT

This paper focuses on the use of alternative assessments at the state level with a focus on the Connecticut experience. Judging from the size of audiences attending sessions on alternative assessments at national conferences and the numbers of articles appearing on performance assessment in recent educational journals, it is obvious that there is a growing interest in the subject among state departments of education and local school districts. The first part of the paper describes the Connecticut Assessment of Educational Progress (CAEP) program, which, between 1980-1987, used performance assessments to assess what students know and can do in art, music, business and office education, English language arts, science, foreign language, drafting, graphic arts, and small engines. Sample exercises and their scoring rubrics are presented and described. The second part describes the Connecticut Mastery Testing program, and the third describes the Connecticut Core of Learning Assessment Program in math and science. The fourth part of the paper summarizes and sets forth some of the prerequisites for the effective use of performance-based assessments to determine what students know and can do. The final section acknowledges some of the paradoxes inherent in using performance-based assessments with students of limited English proficiency. Two responses to the paper, one by Mary Jean Habermann, the other by Richard A. Figueroa, are appended. (VWL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SEA Usage of Alternative Assessment: The Connecticut Experience

Joan Boykoff Baron[1]
Connecticut State
Department of Education

ED349816

This paper is about the use of alternative assessments at the state level with a focus on the Connecticut experience. The topic is a timely one. Judging from the size of audiences attending sessions on alternative assessments at national conferences and the numbers of articles appearing on performance assessment in recent educational journals, it is fair to say that there is a growing interest in this subject among state departments of education and local school districts. Current efforts in this country in states such as Arizona, California, Connecticut, Kentucky, Maryland, and Vermont are paralleled by efforts in other countries. Recent developments in Australia, Great Britain, and the Netherlands (Raizen et al., 1990) provide evidence of an international quest for new forms of assessments which simultaneously will better serve students, teachers, and policy makers. Students will be able to self-monitor their own progress; teachers will be able to make more informed decisions about their students' levels of understanding, and policy makers can have access to accountability data that more closely mirror the skills and applications valued by society.

This new interest in performance assessment stems from both a push and a pull. The push comes from the growing dissatisfaction with this nation's over-reliance on multiple-choice tests (Baron, 1990b; Shepard, 1989; Wiggins, 1989). Many find multiple-choice tests inadequate for assessing higher order thinking skills, deep understanding of content, complex problem solving, communication, and collaboration. Others suggest that they are having a deleterious effect on instruction by encouraging teachers to fragment their curriculum and teach isolated bits and pieces that do not hang together conceptually or tell a coherent story. The pull comes from the ecological and systemic validity of performance assessment (Frederiksen and Collins, 1989). Many educators believe that performance-based assessments more closely represent the kinds of activities that we want our students to be able to undertake as members of society and that practicing for the assessment improves these valued skills and understandings.

FL 020 635

BEST COPY AVAILABLE

Defining Performance Assessment

Over the past decade the term performance assessment has been used to describe many different types of tasks. At the simplest level, a performance assessment can mean a short open-ended written task requiring a student to produce a few sentences. At its most complex level, it can mean a group task in which students work for several days or weeks to design, carry out and report on an investigation on a complex loosely structured problem or even on a problem selected and framed by the students. This paper, by tracing the work in Connecticut over the last decade, reflects the full range of possibilities from the use of a calculator to solve a series of mathematics tasks to a several-day science task in which a group of students work together to design, carry out, and orally report on the results of a series of experiments.

The Potential of Performance-Based Assessment for Improving Education

In this paper, I will focus on the potential of performance-based assessment to make a meaningful contribution to the education of our nation's students. I am operating from the assumption that we as a nation are not currently satisfied with what our nation's students know and can do. Recent reports from both the National Assessment of Educational Progress (NAEP) and International Comparative Assessments (ICA) have been far from reassuring. Most Americans, beginning with our president and governors, believe that we are a nation at risk and are calling for dramatic school reform. In this paper, we will explore the possibilities inherent in using performance-based assessment as one potential lever for changing a complex educational system. There are five aspects to the contribution that revitalizing student assessment can make to the school reform effort.

Clarifying Our Goals and Values

The first requirement is that, when designing performance tasks, it is critical to begin with a clear idea of what we value. In the spirit of AMERICA 2000 (U.S. Department of Education, 1991) and other systemic school reform efforts, I am making the assumption that we are starting with a blank slate and setting out to create assessments based not on what is currently being taught or what is currently in the curriculum but, rather, on what we hope that our students will know and be able to do to function effectively in society. Simply stated, we need to develop assessments based upon what should be happening rather than what is happening. Toward this end, there

is strong consensus among educators in all disciplines that what we value today are students who have a deep understanding of content and can use higher order thinking skills to solve complex and often loosely structured problems. We also put a high premium on students' ability to communicate and collaborate effectively with others. These values are shared universally -- by educators in mathematics, science, the arts and humanities, as well as policy makers, representatives of the business community, and the general public.

Providing Richer Opportunities to Assess What We Value

The second contribution of performance assessment is that it can provide much richer opportunities to assess what we value. Today, based on work in cognitive psychology, task designers are striving to provide interesting real-world contexts to serve as situations for students to integrate their knowledge of content with their knowledge of processes and procedures (Brown, Collins, & Duguid, 1989; Resnick, 1988; Wertsch, 1985). This is by no means easy to accomplish because for so many years they have been kept separate. We are also attempting to incorporate communication skills into our new assessments, calling upon students to report their findings both orally and in writing. This represents a departure from past practice in which we have tended to measure communication skills separately. Finally, despite very little experience in assessing students working together in groups, we are attempting to provide rich contexts in which groups of students can fruitfully solve complex, interesting, and important problems.

Describing Quality Performance

The third contribution of performance assessment is that it permits us to develop a language for describing quality performance. When we develop the scoring guides for teachers and students to use in evaluating students' work, we are developing a multi-faceted description of quality. We are describing the dimensions or characteristics that accompany effective performance and finding examples of students' work across the full range of quality. This can be extremely enlightening for both students and teachers. Therefore, it is important that students' work be scored and interpreted by both the students and their own teachers. In this way, students learn to self-assess their own work and to reflect upon the extent to which they are becoming more effective writers, scientists, or artists. And teachers become more secure in their judgments of the quality of their students' work that has significant ramifications for their work in assessment, curriculum, and instruction.

Setting Standards

The fourth contribution is about standard setting. Using the descriptive criteria established for judging the quality of students' performance, we can set agreed-upon levels of satisfactory and outstanding work. Here, we are asking, "How much is good enough to warrant being labeled as adequate or exemplary?" Many educators today are familiar with how this is done in judging writing samples where teachers participate in short training programs in order to be able to recognize reliably the difference between a 3 and a 4 paper. Once teachers have learned what the attributes of quality work are and have had the opportunity to examine examples of students' work at various levels of quality, they can learn to apply these criteria to new student samples. Under these conditions, different scorers will make consistent (i.e., reliable) judgments about the same student's work. Our experience in Connecticut in scoring students' work on state assessments in a variety of subject areas is that teachers find this process energizing and empowering. For many of them, this represents the first time that they have a forum in which to articulate their own standards of quality. Unfortunately, most teachers today use scoring practices based upon tacit standards that are not shared with their students or their colleagues.

Changing Educational Conversations

The fifth and perhaps most important contribution of performance assessment is that it can dramatically alter the nature of the conversations taking place in classrooms and in the broader educational community. It influences the way teachers talk to students and the way teachers talk to one another. It influences the way students look at their own work and reflect upon its quality. When students internalize a definition of what quality means and can learn to recognize it, they have developed a very valuable critical ability. They can talk with their parents and their teachers about the quality of their work and take steps to acquire the knowledge and skills required to improve it.

Once the descriptive language and the standards are in place, similar conversations can occur between teachers and parents, between administrators and teachers, and between policy makers and members of the general public. In our current mania for "total test" scores and normative comparisons, we have begun to lose our grasp on what quality work means and how we might recognize it. It is argued here that through performance-based assessment, we can take steps to regain our understanding of quality and move toward its realization. Furthermore, it is essential to recognize that being able to describe quality work can assist us in both monitoring student progress and developing a richer array of indicators of school effec-

tiveness. It means that we will be looking at multi-faceted manifestations of student achievement and aggregating judgments on richer and more integrated examples of students' work.

Similarities and Differences between Assessment and Instructional Tasks

There is a growing number of educators around the world who believe that there is little difference between an effective performance assessment task and an effective curriculum or learning task. Burstall (1990) calls the recent British assessment tasks "bits of curriculum." Wolf (1988) refers to the Arts PROPEL assessment tasks as "episodes of learning." I have called for "blurring the edges among assessment, curriculum, and instruction (Baron, 1990b). We view assessment tasks as learning opportunities which, at their best, are explicitly designed to foster students' understandings and skills while undergoing the assessment. This is particularly true when tasks are designed for groups of students to work together to both formulate and solve real-world problems. This should not be construed to mean that we recommend assessment tasks as initial exposures to the understandings and skills being assessed. Rather, assessment tasks are seen as integrative culminating tasks in which students deepen their understandings and synthesize many separate pieces of the curriculum.

Despite the similarities between assessment and instructional tasks, there are a few important differences. Specifically, in assessment tasks as compared with instructional tasks, the role of the teacher is less intrusive. Teachers should be willing to allow their students to flounder; they shouldn't feel the need to rush in to help their students when they don't know how to solve a problem. In addition, when using performance tasks as assessment, it is important to include a set of clear criteria for judging students' performance. Thus, the notion of "teaching to the test" becomes a desirable activity when the tests are seen as an integral part of the curriculum. If we succeed in defining the "shoulds" as described above, then the assessments would serve simultaneously to articulate and embody the goals and objectives of a course of study.

Overview of Performance Assessment in Connecticut

The next five sections of this paper describe Connecticut's attempts over the past decade to develop assessments which use meaningful performance tasks to determine what students know and can do. In all cases, results from the assessments were aggregated and reported to both state-level policy makers and school-based educa-

tors. Each group received data at an appropriate level of specificity. That is, teachers received data suitable for programmatic improvement and policy makers received accountability data suitable for determining how well educational programs were working. The examples come from three assessment programs -- two that are designed to sample a small percentage of students in order to generalize to the rest are Connecticut Assessment of Educational Progress (CAEP) and Connecticut Common Core of Learning program (CCL) and the third that tests every student in grades 4, 6, and 8 in order to identify what students might be in need of remedial assistance [Connecticut Mastery Testing (CMT)].

The first part of the paper describes the CAEP program which, between 1980 and 1987, used performance assessments to assess what students know and can do in art and music, business and office education, English language arts, science, foreign language, drafting, graphic arts, and small engines. Sample exercises and their scoring rubrics are presented and described.

The second part of the paper describes the CMT program which, since 1985, has included the use of calculators for mathematical problem solving in grade 8 and the use of writing samples and note-taking exercises in grades 4, 6, and 8.

The third and longest section of the paper describes the Connecticut Common Core of Learning Assessment Program in Mathematics and Science. Together, teachers and curriculum specialists from several states developed and tried out performance-based assessment tasks often lasting several days. This component of the project is composed of complex sustained tasks in which groups of students work together to design and carry out mathematical and scientific investigations. These are administered and scored by the students' own classroom teachers who participate voluntarily and receive special training. During the 1990-91 school year, a second component was added. This consists of a set of open-ended written exercises which assesses students' conceptual understandings of "big ideas" in science and mathematics. Sample tasks and scoring systems are provided from both components of the project as well as a summary of the components of effective performance tasks.

The fourth part of the paper summarizes and sets forth some of the prerequisites for the effective use of performance-based assessments to determine what students know and can do. The final section of this paper will acknowledge some of the paradoxes inherent in using performance-based assessments with students of limited English proficiency.

Performance Assessment in The Connecticut Assessment of Educational Progress (CAEP) Program between 1980 and 1987

In the 1980s, the CAEP program conducted assessments in eleven subject areas to determine how well students statewide were performing. The emphasis was on program evaluation and not on what individual students knew and were able to do. The CAEP assessment allowed us to ease into performance assessment gradually. In a low stakes testing environment, we began with short, individual on-demand exercises which were scored by external assessors who either observed the student during the task or scored students' work later at a neutral scoring site. These assessments are organized chronologically and summarized in Figure 1, which indicates what grades were tested, how long each performance task required, and when the scoring took place. In all cases, other than those in the vocational educational areas, only a small number of randomly selected students participated in the assessment.

Figure 1
Performance Testing in the Connecticut Assessment of Educational Progress Program, 1980-87

Subject	Year	Grades Tested	Performance Task	Whole Sample or Subsample	Administration Time	When scored? (After self-administered testing or during other-administered testing)
Art	1980-81	4, 8, 11	Draw a room wall and draw a table with people around it	Subsample period	1 class	After
Music	1980-81	4, 8, 11	Sing "America" and complete a musical phrase	Subsample	A few minutes	During
Business and Office Education	1983-84					
*Accounting		12	Make journal entries and complete a payroll record	Whole	1 class period	After
Genera Office		12	Timed typing	Whole	1 class period	After

continued

Figure 1 (Continued)

Subject	Year	Grades Tested	Performance Task	Whole Sample or Subsample)	Administration Time	When scored? (After self-administered testing or during other-administered testing)
Secretary		12	Type and compose part of a letter	Whole period	1 class	After
			Take short-hand	Whole	Part of a class period	After
English Language Arts	1983-84	4, 8, 11	Write 2 essays	Subsample	1 class period	After
			Take a dictated spelling and word usage exercise	Subsample	Part of a class period	
			Revise errors in focus, organization, support and mechanics	Subsample	1 class period	After
			Take notes from a taped lecture	Subsample	Part of a class period	After
Science	1984-85	4, 8, 11	Use scientific apparatus: weigh, measure, focus *microscope*, etc.	Subsample	1 class period	During
		8,11	Design and conduct an experiment	Subsample	1 class period	During
Foreign Language	1986-87	9-12	Write a letter	Whole	1 class period	After
French German Italian Spanish			Speak to an interviewer	Subsample	1 class period	During
Industrial Arts and Technology Education	1986-87					
•Drafting		12	Produce a series of drawings	Subsample	3 1/4 hours	During
Graphic Arts		12	Produce a brochure	Subsample	5 1/2 hours	During
Small Engines		12	Service and repair small engines	Subsample	3 1/4 hours	During

For information, contact Joan Boykoff Baron, Connecticut State Department of Education, P.O. Box 2219, Room 342, Hartford, CT 06145. (566-3847)

In the sections which follow, several of these assessments will be described in greater detail, and examples of tasks will be provided.

Art and Music: 1980-81

Our first attempt at using performance assessment was facilitated by the NAEP program that had assessed art and music using performance tasks almost a decade earlier. Our CAEP assessment used four NAEP tasks and their accompanying scoring criteria and standards. In art, students were asked to make two drawings -- one of their bedroom wall and one of a table with people seated around it (Connecticut State Department of Education, 1982).

In music, students were asked to sing "America" and complete a musical phrase. The drawings were scored after the assessment was complete; the musical performances were scored during the assessment. Using performance assessment in the arts felt natural for teachers and was a comfortable starting point for our work.

Business and Office Education: 1983-84

All twelfth grade students who completed a two-year sequence in general office, secretarial, or accounting courses participated in this assessment (totalling approximately 4,000 students). In addition to a Business Knowledge multiple-choice test, the students were asked to complete a series of tasks which corresponded to the entry-level tasks that these students would be expected to perform in the workplace when they graduated from high school within a few months of the tests. The secretarial students were asked to transcribe letters from dictation and produce a letter using appropriate letter format and composition (see Table 1). The general office students took a timed typing test which was scored on both speed and accuracy (see Table 2). The accounting students were asked to make a series of journal entries which were scored on a variety of criteria related to the correctness of the balance and the titles (see Table 3). All of the papers were scored at a central scoring site by trained Connecticut Business and Office teachers. The performance standards were established by using a combination of several widely used standard-setting procedures which involved judgments by committees of experts from both the business and education communities as well as teachers' ratings of student competence (Connecticut State Department of Education, 1985).

Table 1
Results on Letter-Typing Exercise
(Secretary Text)

Scoring Category	Findings
Format	
Vertical Spacing	82% satisfactory (51% excellent; 31% acceptable - could be improved)
Margins (left-right)	66% satisfactory (26% excellent, 40% acceptable)
Date/Closing (spacing, placement)	83% satisfactory
Paragraphing Format	90% satisfactory (87% excellent7 3% acceptable)
Typing	
Typing/Proofing/Correcting	37% satisfactory (18% no typographical errors. 9% errors corrected adequately)
Hyphenation	80% correct hyphenation or no hyphens used
Spacing after Punctuation	63% spacing correct throughout letter
Omissions/Alteration of Text	84% satisfactory (69% no text changes. 15% acceptable changes)
Composition	
Content	65% all information given
Readability	62% satisfactory (16% highly readable, 46% adequate readability)
Spelling, Grammar, Punctuation	19% no errors

Table 2
Timed Writing/Typing Results
(General Office Test)

Gross Words per Minute	% of Students	Errors per 5 Minutes	% of Students
0-18	6.3	0-3	15.8
19-28	14.0	4-7	23.8
29-38	25.1	8-14	37.6
39-48	34.5	15-21	13.5
49-58	15.9	more than 21	9.3
more than 58	4.2		

NOTE: Standards of acceptable performance were set at 39 gross words per minute and 7.5 errors per 5 minutes, as indicated by the dashed lines above.

Table 3
Results on Journalizing Performance Exercises (Accounting Test)

Task	Percent Correct	Common Errors
Entry to close Salaries Expense Account on Dec 31 in General Journal	11%	9% incorrect figures unbalanced (whether titles correct or not) 8% incorrect account titles
Entry in General Journal to record payment of payroll and payroll taxes	11%	19% incorrect account titles
Entries in appropriate journals, given cash balance, credit memo, check payment info pertaining to a particular account.		
A. Cash Balance - Cash Receipts Journal	44%	
B. Credit Memo - General Journal	7%	8% correct w/o "credit memo" explanation 9% wrong journal
C. Cash Receipt - Cash Receipts Journal	53%	8% included sales discount
Entry for Cash Payment to Creditor- Cash Payments Journal	15%	48% ignored discount
Entry for Cash Payment of Federal Taxes - Cash Payments Journal	17%	9% wrong account title

English Language Arts: 1983-84

This assessment contained multiple-choice sections in several aspects of English Language Arts including literature, listening and note-taking skills, and writing. Using a procedure called matrix sampling, different students took different parts of the assessment. However, no attempt was made to equate the different parts of the assessment because no use was to be made of the scores of individual students. Some students were asked to write two essays -- one narrative and one persuasive. Each essay was scored on more than a dozen traits ranging from the quantity and quality of supporting details to more mechanical aspects of students' writing. Others participated in a revising test in which students were asked to read and correct another student's error-laden essay. Some students were asked to provide the supporting arguments for an essay in which the beginning and end were provided. Still others were asked to provide the beginning and end of an essay in which the middle was provided. Finally, some students took a dictation test in which they heard common homonyms used in context (e.g., to, too, two; their, there, they're). Using a sample of only a few thousand students at a grade level, this assessment gave us a very thorough picture of the writing

skills of Connecticut students. These understandings could not have been obtained through multiple-choice tests. Furthermore, using standards and expectations suggested by a statewide advisory committee, it gave us a very consistent picture of students' shortcomings in producing adequate supporting details in their writing as assessed in a variety of approaches (Connecticut State Department of Education, 1985).

Science: 1984-85

This assessment included a hands-on component in which pairs of students were randomly selected to accompany a specially trained external administrator to a small room in the school. There, one member of the pair was assessed on his or her ability to use various types of scientific apparatus (e.g., scales, thermometers, microscopes, balance beams, miniscus). The other student was assessed on his or her ability to design and carry out an experiment (i.e., the Survival Task) which had been developed for the Assessment Performance Unit (APU) in Great Britain. In designing and carrying out the experiment, the students were scored by an external evaluator who watched each student working alone. The evaluator looked at how carefully the student controlled for each variable and how well the results of the experiment could be trusted. (See Figure 2 for a description of the task, the scoring elements, and the data.) Using standards and expectations suggested by our advisory committee, the results were very disappointing: Whereas approximately two-thirds of the students in both grades 8 and 11 controlled for each variable individually, only one third of the students carried out an experiment whose results could be trusted (Connecticut State Department of Education, 1986). These data proved to be very valuable to us when planning for the Common Core of Learning Assessment five years later. It reinforced the importance of having students design as well as carry out investigations, something which has been getting short shrift in most science classrooms in our nation (Baron, 1990a).

Figure 2

Statement of the Problem - The Survival Task*

Imagine you are stranded on a mountainside in cold, dry, windy weather. You can choose a jacket made from one of the two fabrics in front of you. This is what you have to find out:

Which fabric would keep you warmer?

You can use any of the things in front of you. Choose whatever you need to answer the question.

You can:
- use a can instead of a person
- put warm water inside to make it more life-like
- make it a "jacket" from the material

Make a clear record of your results and conclusions so that someone else can understand what you have found out.

It would be nice to find the answer to the problem, but how you do it is important. Your answer must be a reliable one that I can trust, so please work in a careful and scientific way.

*This task was adapted from a task developed by the Assessment Performance Unit in Great Britain.

Results of the Connecticut Assessment of Education Progress in Science 1984-85

Control - Can (both size and material)

Grade 11	Grade 8	
69	64	controlled
22	21	not controlled
5	15	irrelevant considering approach
5		no response

Control - Fabric (size and fastening)

Grade 11	Grade 8	
65	64	controlled
31	34	not controlled
4	2	no response

Control - Water (initial temperature)

Grade 11	Grade 8	
75	62	controlled
16	23	not controlled
4	15	irrelevant considering approach
5		no response

Control - Water (volume)

Grade 11	Grade 8	
69	57	controlled
23	27	not controlled
4	17	irrelevant considering approach
5		no response

Control - Measurement Intervals/Temperature Drop

Grade 11	Grade 8	
63	53	controlled
28	23	not controlled
4	22	irrelevant considering approach
5	1	no response

Control - Temperature Measurements

Grade 11	Grade 8	
90	69	all measurements within 2 degrees of test administor's readings
3	5	all except one or two measurements within 2 degrees of test administrator's readings
0	5	irrelevant considering approach
7	21	no response

Control - Measurement Schedule

Grade 11	Grade 8	
65	52	permits detection of temperature change
29	43	does not permit detection of temperature change
7	5	no response

Control - Recording of Data

Grade 11	Grade 8	
65	58	data organized and recorded clearly enough to permit appropriate interpretation
30	41	data not organized and recorded clearly enough...
5	1	no response

Control - Water (initial temperature)

Grade 11	Grade 8	
75	62	controlled
16	23	not controlled
4	15	irrelevant considering approach
5		no response

Control- Conclusion

Grade 11	Grade 8	
57	51	conclusion consistent with data
12	13	conclusion not consistent with data
25	35	conclusion not possible because of design or execution
6		no response

Control - Overall Evaluation of Experiment

Grade 11	Grade 8	
39	23	design and execution such that one could "trust" conclusion
33	37	design and execution have minor problems with could create some doubt about conclusion
23	39	design and execution such that one should have no faith in the conclusion at all
5		no response

Modern Foreign Languages: French, German, Italian, Spanish: 1986-87

Our assessment in modern foreign languages consisted of items in culture, reading, listening, speaking, and writing. Communicative proficiency was highly valued by the advisory committee and it determined to develop an assessment based on the ACTFL Guidelines which represented a scale of communicative proficiency ranging from Novice to Advanced. (The quality standards were built into the ACTFL scale itself.) The reading test used authentic materials from advertisements, menus, and newspaper articles. The listening test used tape recorded conversations and weather reports. The speaking test required an oral interview lasting up to one half hour in which a specially trained Connecticut teacher who participated in a week-long ACTFL training program interviewed students one at a time. The writing assessment consisted of a letter written to a student who would be visiting next year (see Figure 3). This assessment task was specially designed to give all participating high school students (those who had completed three or more years of a modern foreign language) a chance to write something. The letter began by asking for a description of members of the student's family and the rooms in his or her house -- both of which are generally learned very early in the study of foreign language. The present tense was called for at the beginning of the letter and the past and future tenses were required later in the letter. From this one developmentally constructed task we learned a lot about the student's level of written proficiency. Students' essays were scored as Novice, Intermediate, Intermediate High, or Advanced using the scoring rubrics displayed in Figure 3. Two specially trained Connecticut foreign language teachers scored each student's essay and the level of exact agreement was over 90 percent.

Figure 3
Connecticut Assessment of Educational Progress (CAEP) - Foreign Language Writing Test

Directions: Now that your family has been accepted to host an exchange student in the INTERPALS PROGRAM, write a letter in **Spanish** welcoming the exchange student from Cordoba who is coming to live with you. The student's name is Mercedes Sanchez Aparicio.

In your letter, write about

- your family and the house in which you live
- your school and daily activities
- your interests and hobbies
- something interesting that has happened in your school or community recently

Figure 3 (Continued)

Also, ask Mercedes for any information you would like to know about her.

WRITE YOUR LETTER IN YOUR ANSWER BOOKLET.

(RUBLICS FOR SCORING)

O	Blank paper, paper entirely in English or dialectal language.
N (Novice)	Use of high-frequency words, memorized phrases and formulaic sentences on familiar topics. Little or no creativity with the language beyond memorized patterns.
I (Intermediate)	Recombinations of learned vocabulary and structures into simple sentences. Language may be inadequate to express anything other than the most elementary ideas. Choppy sentences with frequent limited vocabulary and syntactical resources. Sentences will be high end. Often reads very much like a direct translation from English.
IH (Intermediate High)	Can write creative sentences, sometimes fairly complex ones, but not consistently. Structural forms reflecting time, tense or aspect are attempted, but the result is not always successful. An ability to describe and narrate in paragraphs is emerging, but the use of basic cohesive elements indicating transition is inconsistent. Vocabulary and structural resources allow the student to paraphase at times. Papers will often read like an academic exercise.
A (Advanced)	Able to join sentences in simple discourse on familar topics. Has sufficient writing vocabulary to express self simply with some circumlocation although the language may not be idiomatic. Good control of the most frequently-used syntactic structures (e.g., common word order patterns, coordination subordination). Writing may reflect some native-language interference, but there is a sense that the student is comfortable with the target language and can go beyond the academic task.

Drafting, Graphic Arts, and Small Engines: 1986-87

High school students who had completed a two-year sequence in drafting, graphic arts or small engines participated in this assessment. Each student took a multiple-choice test of background knowledge and a sample of students in each area was selected to participate in a performance assessment. As above, in the Business and Office assessment, these tasks represented job-entry level skills that students would be expected to have obtained before being employed. In the Drafting test, students spent more than three hours drawing a series of orthographic projections; in graphic arts, students spent more than five hours producing a brochure, and in small engines, the students spent over three hours servicing and repairing a series of small engines. Every task was scored by a trained observer from business and industry who accompanied the student throughout the time and assessed the quality of the student's product and process. In the Drafting example provided in Figure 4, the quality of the product was assessed on its accuracy; its appearance (e.g., smudges, incomplete erasures, tears, and rips); its alignment of views, including correct views, including correct projection, view selection, and view position, and its completeness and correctness with attention to missing or misrepresented lines, and the size and shape. The quality of the process was judged on its technique, including the use of instruments, the fastening and problem-solving approaches, and the construction method; the layout, including view position, spacing, and projection; the lines, with attention to density, width, and character; and the geometrics, with attention to parallelism, perpendicularity, concentricity, tangencies, and angularity. This assessment represented a major step forward in articulating the scoring criteria that are often used tacitly in assessments of this type where an expert in the field holistically assesses the quality of a student's drawing. On Figure 4, for each scoring scale, there is an asterisk next to level B. Using a combination of standard-setting approaches with teachers and representatives from industry, level B was determined to be the expected level of performance for a student entering the workplace immediately after graduation from high school (Connecticut State Department of Education, 1988).

Figure 4
Drafting Job One—Orthographic Projection

Process

	A	*B	C	D	E
1. Technique (Approach)	0-2 Errors	3-4 Errors	5-6 Errors	7-8 Errors	More Than 8 Errors

Check:
_ Use of Instruments
_ Fastening Paper to Board
_ Problem-Solving. Approach
_ Construction Methods

Figure 4 (Continued)

2. Layout (Final)	A	*B	C	D	E
	0-1 Errors	2-3 Errors	4-5 Errors	6-7 Errors	More Than 7 Errors
	Check: _ View Position _ View Spacing _ View Projection				
3. Lines	A	*B	C	D	E
	0-2 Errors	3-4 Errors	5-6 Errors	7-8 Errors	More Than 8 Errors
	Check: _ Density _ Width _ Character (straightness, intersections, tangency, consistency)				
4. Geometrics	A	*B	C	D	E
	0-2 Errors	3-4 Errors	5-6 Errors	7-8 Errors	More Than 8 Errors
	Check: _ Parallelism _ Perpendicularity _ Concentricity _ Tangencies _ Angularity				
5. Accuracy	A	*B	C	D	E
	95% of Lines	90% of Lines	80% of Lines	70% of Lines	<70% of Lines
6. Appearance	A	*B	C	D	E
	95% Free	90% Free	80% Free	70% Free	<70% Free
	Check: _ Smudges _ Incomplete Erasures _ Tears/Rips				
7. Alignment of Views	A	*B	C	D	E
	Conforms to Master	1Criterion Incorrect	2 Criteria Incorrect	3 Criteria Incorrect	>3Criteria Incorrect
	Check: _ Correct Projection _ View Selection _ View Position				
8. Completeness & Correctness	A	*B	C	D	E
	0-2 Errors	3-4 Errors	5-6 Errors	7-8 Errors	More Than 8 Errors
	Check: _ Missing Lines _ Misrepresentation of Lines _ Construction (not true size/shape)				

* Indicates Entry-Level Job Standard

BEST COPY AVAILABLE

Performance Assessment Tasks on the Connecticut Mastery Test in Mathematics, Reading, and Language Arts, including Writing: 1985-1991.

In 1985, Connecticut moved from a proficiency test which had been taken by every student in Grade 9 to a mastery test taken by each student in Grades 4, 6, and 8. The large majority of the test uses a multiple-choice format. However, there are three performance tasks. First, as in the ninth grade test, every student produced a writing sample which was holistically scored by two specially trained Connecticut teachers at a central scoring location. If a student fell below the standard (set by the State Board of Education at a 4 on an 8-point scale), the paper would be analytically scored by a third reader on a series of four dimensions (support, focus, organization, and mechanics). Students also participated in a note-taking exercise based on the prototype developed for the CAEP program in which they took notes from a tape-recorded lecture and then used those notes later in the test to answer a series of questions. The final set of performance tasks occurs in the eighth grade mathematics assessment that contains one part on which students use calculators to solve complex multi-step problems. Because this is a higher-stakes assessment than CAEP, teachers report that they are providing more opportunities for their students than they would be providing without the assessment -- opportunities to do more writing, take notes, and use calculators. Returning to a point made earlier, if these are skills that are highly valued by society, using appropriate performance assessments can serve an important function.

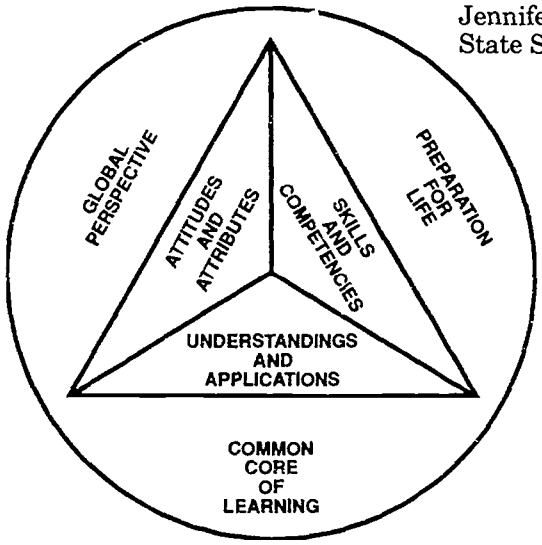
The Connecticut Common Core of Learning Assessment Program in Science and Mathematics: 1990 to Present

In 1986, Connecticut's Commissioner of Education Gerald N. Tirozzi convened a blue-ribbon committee to determine what Connecticut students should know and be able to do after completing high school. The results of their deliberations are provided in Figure 5, which summarizes the attributes and attitudes, skills and competencies, and understandings and applications that they deemed appropriate. The Common Core of Learning document (Connecticut State Board of Education, 1987) was adopted by the State Board of Education.

Figure 5

Connecticut's Common Core of Learning is organized under three major headings with subheadings that reflect significant groups of skills, knowledge and attitudes:

Illustration designed by
Jennifer C. Goldberg
State Student Advisory Council



Attributes and Attitudes

Self-Concept
Motivation and Persistence
Responsibility and Self-Reliance
Intellectual curiosity
Interpersonal Relations
Sense of Community
Moral and Ethical values

Skills and Competencies

Reading
Writing
Speaking, Listening and Viewing
Quantitative Skills
Reasoning and Problem Solving
Learning Skills

Understandings and Applications

The Arts
Careers and Vocations
Cultures and Languages
History and Social Sciences
Literature
Mathematics
Physical Development and Health
Science and Technology

The Connecticut Common Core of Learning Assessment Project's overall objective is to develop performance-based assessment tasks for high school students in mathematics and science that can be used by both teachers and educational policy makers to determine what students know and can do. The content and processes included in our assessment tasks are modeled on the recommendations of mathematicians and scientists, mathematics and science educators, and representatives from business and industry. The structure of the tasks has been strongly influenced by psychological theory and research in the areas of cognition, motivation, learning and instruction. Two documents which shaped our earliest thinking in the project were The Curriculum and Evaluation Standards for School Mathematics, National Council of Teachers of Mathematics (NCTM), 1989, and Science for all Americans, (American Association for the Advancement of Science (AAAS), 1989. The first document stresses the importance of mathematics as problem solving, communication, connection making, and collaboration, and relates content to these broader purposes. The AAAS document describes the major conceptual understandings that underlie our view of the natural world as well as the appropriate attitudes and dispositions associated with science. Both documents support the view of education producing active and engaged students who are able to formulate problems, plan investigations, collect and analyze their own data, and communicate their findings effectively in writing and orally. They both envision students who are able to solve problems effectively by themselves and in groups. Connecticut's Common Core of Learning document fully supports this view of learning and assessment (Baron et al. 1989).

Some Departures from Earlier Assessment Programs.

By 1990, we felt ready to extend our performance-based assessments in several ways. First, we supplemented our on-demand tasks with embedded tasks. This approach allowed teachers to exercise choice in a number of important ways. Teachers could choose which assessment task to use and when, allowing the assessment to fit more integrally into their curriculum. Second, we extended the length of the tasks to endure over several days. Once the tasks were embedded in the classroom, it no longer mattered whether students would work at home or talk to others. Therefore, as a third departure, we included group tasks as well as individual tasks. This decision was motivated by several sources. First, there is the recognition by business and industry as well as the general public that it is important for people to be able to work as part of a team; most jobs are accomplished by a group of workers. Second, by making use of an interpersonal context, we also build upon Vygotsky's (1978) notion of the zone of proximal development. In this way, students are able to achieve a higher level of achievement earlier than they would

achieve by working alone. A fourth departure resulted from our recognition of the importance of sharing the scoring criteria with students and teachers as a routine part of the assessment. This allows the kinds of conversations alluded to in the earlier part of this paper.

Three Guiding Principles.

Three additional principles have helped to shape our assessment work. The first is that we view our assessment tasks as "bits of curriculum." They are intended to provide students with opportunities to "put their learning together" -- to integrate and synthesize separate bits and pieces of knowledge about science and mathematics and deepen their understanding of the big ideas in these disciplines. The second is that we are designing our tasks to represent what our students should know rather than what they may currently be learning in their classes. This means that for the next several years, the stakes for this assessment will be low, allowing Connecticut educators time to examine their curricula, instruction, and assessment strategies in order to bring them into closer alignment with the new vision of science and mathematics. The third principle is that we view ourselves and our teachers as learners in this development process. Despite the fact that we are starting out with a fairly well articulated new vision of science and mathematics, there are few examples of consonant curriculum or assessment available. Therefore, as we deepen our own understandings of how to develop appropriate learning and assessment tasks, it is a major unfinished goal of our project staff to document and share these understandings with others.

A Description of the Common Core of Learning Assessment

Our project has two major components, both designed to provide information about what Connecticut students know and can do in science and mathematics after twelve years of school. These are described below and summarized in Figure 6.

Figure 6
Connecticut Common Core of Learning Assessment Project in Science and Mathematics: An Analysis of Its Two Components*

DIMENSION	COMPONENT I	COMPONENT II
Policy Question	Consistent with our new view of science and mathematics education, what do Connecticut high school students who are currently enrolled in science and mathematics classes know and what can they do?	Consistent with our new view of science and mathematics education, what do Connecticut high school graduates know and what can they do in science and mathematics irrespective of what courses they have taken?
Number of Tasks Pilot Tested	Mathematics: 18 Science: 26	Mathematics: 81 Science Type 1: 106 Science Type 2: 45 Science Type 3: 22
Numbers of Classrooms in which Each Task Was Administered	Between 0 and 8	Between 4 and 8
Assessment Task Format(s)/(Types)	Group investigations requiring students to design and carry out a study, analyze and portray data and report the results in writing and orally. Individual tasks precede and follow the group work.	Mathematics Open-ended problems requiring written responses, justifications and explanations. Problems have multiple solutions and/or solution paths and may require using mathematics to make decisions.
Assessment Task Format(s)/(Types)		Science: Type 1 Responding to open-ended questions and problems requiring written answers, justifications, and explanations. Type 2—Constructing charts, graphs, and tables from data and interpreting qualitative information. Type 3—Students generally design and always conduct a hands-on investigation in the presence of a trained observer who interviews the student.
Time per Task	Several class periods with some out-of-school time.	Mathematics tasks and Science Types 1 and 2: Approximately 10-20 minutes per task. Science Type 3 tasks require between one and two class periods.

Figure 6 (Continued)
Connecticut Common Core of Learning Assessment Project in Science and Mathematics: An Analysis of Its Two Components*

DIMENSION	COMPONENT I	COMPONENT II
Pilot Sample	Volunteer high school science and mathematics in 20 states administered three tasks of their choice to their own students in grades 9-12 in biology, chemistry, earth science, physics, general math, algebra, geometry, and advanced mathematics.	In 65 volunteer Connecticut high schools, science and mathematics teachers administered 6 to 9 tasks to their own students, primarily juniors. Tasks were matrix sampled so that different students took different tasks.
When administered	At each teacher's discretion spread out over the school year.	Between May 13 and May 24, 1991.
Scored Elements	Group Work (written and oral student reports); Finishing by Yourself (individual tasks).	Types 1 and 2: Open-ended written responses, graphs, tables, charts. Types 3: Hands-on investigations.
Other Available Data Sources	Beginning by Yourself (individual task); Self-assessment of behavior in groups; Videotapes of some groups working on tasks; Students' reactions to the task; Teachers' reactions to the task; Student attitude questionnaires (fall and spring) including students' self-reported grades.	Students' self-reported overall grades and grades in mathematics and science for each course taken.
Scoring Dimensions	Qualitative judgements obtained on between 4 and 10 dimensions.	Mathematics and Science Types 1 and 2: To be determined Summer and Fall 1991. (Our challenge is to capture qualitative differences within several different justifiable approaches to each question.
Scorers	The students' own science and mathematics teachers.	Connecticut science and mathematics teachers at a central location.

BEST COPY AVAILABLE

Figure 6 (Continued)
Connecticut Common Core of Learning Assessment Project in Science and Mathematics: An Analysis of Its Two Components*

DIMENSION	COMPONENT I	COMPONENT II
Required Professional Development	Extensive professional development and continual support of teachers in using group work, understanding the appropriate role of the teacher during the assessment, understanding important scoring procedures and exercising common standards of judgement.	Mathematics and Science Types 1 and 2: None required to administer the tasks. To score the tasks, considerable training will be required. Science Type 3: One day of training is required to administer the investigations. A second day of training is required to score students' work.
Who Will Be Assessed in 1991-92	Volunteer teachers in Connecticut and other states.	A random sample of Connecticut high school juniors.

* Funded by the Connecticut State Department of Education and the National Science Foundation (SPA-8954692) Project Director: Dr. Douglas A. Rindone, Ed.D. (203) 566-1684 Principal Investigator: Joan B. Baron, Ph.D. (203) 566-5454

Component 1 is designed to answer the policy question "Consistent with our new vision of science and mathematics, what do our high school students who are currently enrolled in science and mathematics classes know and what can they do? Our biology tasks will be administered by voluntary biology teachers to their own students during the school year; the chemistry tasks will be administered by voluntary chemistry teachers to their own students. The same will hold true for physics, earth science, and all areas of high school mathematics. It is our intent that data from these classroom-situated tasks be useful to at least three important client groups.

- Of primary importance are the students themselves. By participating in rich tasks with multidimensional scoring criteria, students will be able to monitor their own progress.

- Second, classroom teachers can use the data in assessing their students' learning and in making changes in their curriculum and instructional strategies.

- Third, these data will contribute to our reports to policy makers on the condition of education in Connecticut. While certain features of the research design are limiting (e.g., the fact that the sample is non-random limits the generalizability of the results), the richness of the data should deepen our understanding of what students know and can do in science and mathematics.

An example of a Component 1 science task called The Soda Task is provided below in Figure 7.

Figure 7
The Soda Task

Part I: Getting Started by Yourself

Name _____ I. D. #_____

You will be given two samples of soda; one regular soda containing sugar and the other one diet soda containing an artificial sweetener. Your task is to identify each sample as diet or regular based on your knowledge of physics, chemistry, and/or biology. **As in any experiment, you are not allowed to taste any of the samples.**

Come up with a list of properties of the two sodas which might help to distinguish between the samples. Write down as many as you can think of.

Written for the Connecticut State Department of Education - Sponsored by the National Science Foundation

Figure 7 (Continued)
The Soda Task

Part II: Group Work

Names_____ I. D. #s_____

The criteria that will be used to assess your group work are found on the Objectives Rating Form - Group. Each member of your group will also fill out the Group Performance Rating Form.

1. Make a list of **as many possible tests as your group can** think of **which** might help to distinguish between the two types of soda. Briefly explain why you think they might work. Write your answers below.

2. Now select two tests from your list to carry out. They should be the ones which your group believes would be the most effective in distinguishing between the two soda samples. Explain why you chose each of them. Show that you understand the science involved in each test.

3. Write out a complete experimental plan for each of these two tests. It should be clear enough so that someone else could easily repeat your experiments. Include a list of all the materials and equipment that you will need. Show your plan to your teacher before proceeding.

After getting approval from your teacher, carry out your experiments.

4. Record all of the results of your experiments in a clear and organized way.

5. What conclusions can be made from your experiments?

6. Make an oral presentation summarizing your experiments and results. Each member of your group should be ready to participate in any part of the presentation. Your teacher will determine the order of the presenters.

7. After hearing all the oral presentations answer the following question; if you were diabetic and had to know whether a sample of soda had sugar in it, which test would your group trust the most? Which test would your group trust the least? Explain fully why you chose each of these using complete sentences.

Written for the Connecticut State Department of Education - Sponsored by the National Science Foundation

Figure 7 (Continued)

Part II: Objectives Rating Form - Group

Student I.D. # 's

1. _____
2. _____
3. _____
4. _____
5. _____

Title of the Task: The Soda Task Teacher Id. # _____

The group should be able to ...	*Overall or test 1* E	G	N.I.	U	*Test 2* E	G	N.I.	U
1. make a list of reasonable solutions to the problem.								
2. select tests based on scientific knowledge.								
Testing against knowns (if used) ◯ Adequate ◯ inadequate								
3. design a controlled experiment.								
4. gather pertinent data.								
5. draw conclusions consistent with the data.	◯ Yes ◯ No				◯ Yes ◯ No			
Discrepant Results (if reported) ◯ Resolved adequately ◯ Not resolved adequately								
6. select most and least effective tests based on their scientific validity.								
7. communicate the strategies and outcomes of a study through written means.								
8. collaborate effectively.								

E = Excellent G = Good N.I. = Needs Improvement U = Unacceptable

Figure 7 (Continued)

Part II: Objectives Rating Form - Oral Communication

Title of the Task: The Soda Task Teacher Id. # _____ Student Id. # _____

The student should be able to ...	E	G	N.I.	U
1. the content is well organized and appropriate to the task.				
2. presenters voices are clear, enthusiastic and loud enough to hear, with no distractions.				
3. presenters answer questions thoroughly and clearly.				
4. presenters maintain eye contact with the audience.				
5. visual aids are easily seen and understood.				

E = Excellent G = Good N.I. = Needs Improvement U = Unacceptable

Figure 7 (Continued)

STUDENT INSTRUCTIONS

Group Performace Rating Form
Connecticut Common Core of Learning Assessment Project

Using a Number 2 pencil, for each question, fill in the appropriate box to describe your behavior in the group during this task. Please note that items 3, 7, and 15 are different from the others; when you rate these items, be aware that you are pointing out a problem.

After you have completed your ratings, write the name of the task, its Task I.D. No. and the date below and circulate your self-ratings to each person in your group for his or her review and signature or initials. If any member of your group disagrees with your ratings of yourself, please discuss with that person the reasons for the disagreement and then decide whether or not you want to change your original rating

Name of Task _____ Task I.D No. _____ Date _____

	Signature or Initials of Other Group Members	Student I.D. No.
1	_____	_____
2	_____	_____
3	_____	_____
4.	_____	_____
5	_____	_____

When each member of your group has approved and signed your rating sheet, please submit this form to your teacher.

If you cannot agree on a rating or if you wish to make comments about this process, please use the space below. Do not write your comments on the other side of this sheet.

This space may be used for COMMENTS

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you for participating in this project

217
BEST COPY AVAILABLE

Figure 7 (Continued)

Student Name _____

Student I.D. Number _____

Check One

A. GROUP PARTICIPATION	Almost Always	Often	Some times	Rarely
1 Participated in group discussion without prompting.				
2 Did his or her fair share of the work				
3 Tried to dominate the group - interrupted others, spoke too much				
4 Participated in the Group's Activities				
B. STAYING ON THE TOPIC	**Almost Always**	**Often**	**Some times**	**Rarely**
5 Paid attention, listened to what was being said and done.				
6 Made comments aimed at getting the group back to the topic.				
7 Got off the topic or changed the subject				
8. Stayed on the Topic				
C. OFFERING USEFUL IDEAS	**Almost Always**	**Often**	**Some times**	**Rarely**
9 Gave ideas and suggestions that helped the group				
10 Offered helpful criticism and comments				
11 Influenced the group's decisions and plans				
12 Offered Useful Ideas				
D. CONSIDERATION	**Almost Always**	**Often**	**Some times**	**Rarely**
13 Made positive, encouraging remarks about group members and their ideas				
14. Gave recognition and credit to others for their ideas				
15 Made inconsiderate or hostile comments about a group member				
16 Was Considerate of Others				
E. INVOLVING OTHERS	**Almost Always**	**Often**	**Some times**	**Rarely**
17 Got others involved by asking questions, requesting input or challenging others				
18 Tried to get the group working together to reach group agreements				
19 Seriously considered the ideas of others				
20. Involved Others				
F. COMMUNICATING	**Almost Always**	**Often**	**Some times**	**Rarely**
21 Spoke clearly Was easy to hear and understand				
22 Expressed ideas clearly and effectively				
23. Communicated Clearly				

BEST COPY AVAILABLE

Figure 7 (Continued)
The Soda Task

Part III: Finishing by Yourself

Name_____ I. D. # _____

If you were given two samples of water, one of which is salt water and the other fresh water, which tests can you think of which might help to differentiate between the two samples. (You may use tests from the soda task or other ones.) Explain why you think each might work using complete sentences. Show that you understand the science involved.

Written for the Connecticut State Department of Education -
Sponsored by the National Science Foundation

Figure 7 (Continued)
Sample
Student Reaction Form

Date 10-4-90 Name of Perf. Task The Soda Task

If there is not enough room to answer the questions completely, please answer on the back. Thanks!

1. Did you enjoy working on this Performance Task? Explain why or why not.
I enjoyed working on this talk because it was a change from the step by step labs that we always do and I got to use my imagination

2. Describe something about this Performance Task that you liked.
I liked being able to try new and different procedures. I also liked the idea of using an everyday substance, like soda, in an experiment. It made science seem more practical.

3. Describe something about this Performance Task that you didn't like.
I didn't like having to ~~state~~ make up an experiment because I had never done anything like it before. (Now I think that making your experiment is fun.)

4. How did you feel about working in a group?
I liked working in a group. It was fun and I learned a lot about myself and others.

5. Would you like to do more group problem solving activities as a part of this class?
Yes, working in a group allows you to see problems in a different perspective. It also allows you to get a better social education because you have to listen to everybody else and cooperate if you want to get any work done.

6. How do you feel about the use of performance tasks to evaluate your knowledge and skills?
I like ~~it~~ them. ~~Performance tasks~~ taking a test can prove you remembered something, this doesn't mean you understand it. ~~It~~ In order to do a performance talk you must remember the information and understand it.

7. What, if anything, did you learn during this Performance Task?
I learned to always, always label your containers and if you want something to evaporate quickly only use a small amount of the substance. I also learned that working in a group is much easier and more fun than working alone.

BEST COPY AVAILABLE

Component 2 is designed to answer the policy question, "What do Connecticut high school graduates know and what can they do in science and mathematics irrespective of what courses they have taken?" These performance tasks will be administered to a random sample of high school juniors by someone other than their science or mathematics teacher. (We are assessing students at the end of Grade 11 rather than Grade 12 because their motivation is higher and we believe that they will take the assessment more seriously.) Students' work will be scored by teachers at a neutral scoring site. These data will be used to report on the condition of education in Connecticut and to allow educational decision makers at all levels to set programmatic priorities for science and mathematics education. A supplemental benefit of these open-ended assessment tasks is that they will provide models of alternative formats that teachers can use to assess the depth of their students' understanding of science and mathematics. Where possible, we have attempted to write items that have several correct solutions or solution paths. Some items require students to use the same data set to support different assertions. Two examples of Component 2 tasks will be provided below in Figures 8 (science) and 9 (mathematics).

Figure 8
Energized Object

For each of the following objects, name the kinds of energy involved and explain how they are involved.

1. Moving toy car

2. Burning candle

Figure 8 (Continued)

3. Bursting balloon

4. Growing Plant

Figure 9
McDonald's Claim

You and a friend read in the newspaper that 7% of all Americans eat at McDonald's each day. Your friend says, "That's impossible!"

You know that there are approximately 250,000,000 Americans and approximately 9,000 McDonald's restaurants in the U.S. You think the claim is reasonable.

Show your mathematical work and write a paragraph or two that explains your reasoning.

Neither of these components, by itself, can provide a complete answer to the question of what our students know and can do. However, when considered together, educators and policy makers will have a better understanding of both the condition of science and mathematics education in Connecticut and some steps that can be taken to strengthen these programs.

Accomplishments to Date

During the first two years of our project, we have developed more than 300 performance tasks, described in the section which follows.

Component 1: During the 1989-90 school year, following an intensive six-day training summer session, we worked closely with fifty teachers from ten states to develop performance tasks that could be used to assess students' understandings in high school science classes (i.e., biology, chemistry, earth science, and physics) and high school mathematics classes (i.e., general mathematics, algebra, geometry, and advanced mathematics, including advanced algebra, trigonometry, and calculus). At the end of the first year, we had available approximately fifty performance tasks at different levels of development.

During July 1990, we trained a cadre of ninety high school teachers and state education department personnel to try out and refine these tasks. Before leaving the workshop, teachers were asked to choose three tasks to use in their classrooms during the 1990-91 school year. For each task, they agreed to videotape one group of their students at work, score their students' group products and processes on a series of between five and ten pre-specified scoring dimensions, and score an individual task designed to determine the extent to which each member of the group really understood what the group had done.

Each Component 1 task has three sections that involve a blend of individual work at the beginning and end of the task and group work in the middle. At the beginning of the task, each student provides information individually about his or her prior knowledge and understanding of the scientific concepts and processes relevant to the tasks. (See Figure 7, The Soda Task, Part 1 for an example.) In the middle section of the task, by far the longest phase, students work as a team to produce a group product. Students plan together and work together. Throughout the tasks, interdependence is fostered by having each student feel responsible for telling "the whole story" from the development of the group's initial design to its final conclusions. Also, at various intervals, students are asked to monitor their success both as a group and as individuals working as part of a group. (See the Checklists provided in Figure 7, Part 2 for examples of these scoring checklists.) Following the group work, a related task is administered to students individually to see what each student learned from the group experience. In the cognitive and instructional psychology literature these have been referred to as "near-transfer" or application tasks. We recognize that these individual tasks do not fully represent the knowledge tapped by the larger tasks, but they are designed to provide the teacher and students with some evidence that the student can use the knowledge gained in the group experience on a new but very similar piece of the science or mathematics terrain explored in the group task. (See Figure 7, Part 3 for an example of this near-transfer task.)

In attempting to develop a series of assessment tasks suitable for Component 1, we have developed a set of characteristics of rich performance tasks (Baron, 1990 and Baron, in press). Some of these are described in Figure 10.

Figure 10
What Are the Characteristics of Enriched Performance Assessment Tasks?

Enriched performance assessment tasks:

- are grounded in real-world contexts
- involve sustained work and often take several days of combined in-class and out-of-class time
- are based upon the most essential aspects of the content of the discipline(s) being assessed; that is. they deal with "big ideas" and major concepts (e.g., energy, form and function, change) rather than peripheral or tangential topics (American Association for the Advancement of Science, 1989; National Council of Teachers of Mathematics, 1988)
- are broad in scope, frequently integrating several scientific principles and concepts
- blend essential content with essential processes, often requiring the use of scientific methodology and the manipulation of scientific tools and apparatus
- present nonroutine, open-ended, and sometimes loosely structured problems that require students both to define the problem and to determine a strategy for solving it; optimal problems afford both multiple solutions and multiple solution paths (Charles & Saver, 1989; Greeno, 1978; Resnick, 1989; Schoenfeld, 1976)
- encourage group discussion and "brainstorming," in which a problem is considered from multiple perspectives
- require students to deterrnine what data are needed, collect the data, report and portray them, and analyze them to discern sources of error
- call upon students to make, explain, and defend their assumptions, predictions, and estimates
- stimulate students to make connections and generalizations that will increase their understanding of the important concepts and processes

- are accompanied by explicitly stated scoring criteria related to content, process, group skills, communication skills, and a variety of motivational dispositions and "habits of mind" (Wiggins, 1989)

- spur students to monitor themselves and to think about their progress (as individuals, as members of a group, and as a complete group) in order to determine how they might improve both their investigational and group process skills

- necessitate that students use a variety of skills both for acquiring information (e.g., reading, listening, and viewing) and for communicating their strategies, data, conclusions, and reflections (e.g., speaking, writing, and graphic displays)

Baron, J. B. (1990b).

Over the past two years, we have been soliciting reactions from both the students and the teachers participating in our project. One student's reactions are found at the end of Figure 7. Although we have only begun to compile the large amount of data amassed thus far, we recognize the complexity, the difficulty, and the rewards inherent in developing meaningful and effective performance tasks. Other students' reactions were summarized by Claire Harrison (1991), a member of the CCL project team and are provided in Figure 11.

Figure 11
Student Reactions to Component 1 Tasks
Prepared by Claire Harrison
Connecticut Common Core of Learning Assessment Program

We have learned that when tasks worked well, students enjoyed the freedom and the challenge of designing and carrying out their own projects. They felt involved and intrigued, and liked not being given the answer. They liked applying and testing their knowledge, especially on a practical question. They enjoyed seeing their ideas work and their predictions confirmed, and sometimes mentioned feelings of pride and accomplishment. In order for this to occur, students needed a task that was sufficiently challenging. They also had to have an ideas of where to start and in what direction to head. Thus, they needed a level of prior knowledge about the topic. They also needed a task that was not too vague or confusing. Having a clear goal seemed important to some students.

Figure 11 (Continued)

A small minority of the students had difficulty dealing with the open-ended nature of the tasks. They were uncomfortable not knowing whether their work was correct. Some students found it helpful to be able to check their work with other group members. Whether students liked or disliked the task, most enjoyed working in a group. Working with others made the tasks more interesting and more fun. The students liked hearing the ideas and opinions of others, and finding out how others approach problems. A few mentioned enjoying having their thoughts listened to and accepted by others. Most felt they learned more by working in the group. Being able to help each other was also frequently mentioned as a positive aspect of group work. A few students did express concerns about group work. Most of these were related to the possible effect of the group on their work. They were concerned that being part of a group that worked poorly together, or in which not all members participated, would depress their own grades. Some, seeing the advantage to the group of having knowledgeable or skillful members, felt that this resource should be evenly distributed. A few students were concerned about group members who do not carry their own weight but benefit from the group's effort. A preference for working alone was expressed by a minority of students. Some of them felt they work better alone and some wanted to carry out their own ideas in their own way.

From a summary prepared by Harrison (1991) of twenty-nine teacher questionnaires returned in June 1991, we have learned:

> teachers use these assessment tasks as assessment, curriculum instruction and combinations of these. Teachers report that they are gaining important new insights about their students' skills and understandings -- expressing surprise at the difficulty their students encountered in doing the tasks. Teachers reported that they plan to use more cooperative learning and group work in their classes as a result to using these tasks. The major problem reported by the teachers involved time. Twenty-two of the twenty-nine teachers cited time as a constraint in using the tasks. Eight of these explained that the time taken to do the tasks made it difficult to cover the existing curriculum; several reported falling behind. This was a particular problem for teachers whose course of study or examinations are determined on a school-wide basis.

Component 2: The science assessment development work began in the summer of 1990 and continued throughout the fall with a selected group of Connecticut high school teachers and Department staff working together throughout the fall to write open-ended tasks. During the winter and early spring these tasks were critiqued by other Connecticut teachers and practicing scientists in Connecticut colleges and universities. In May 1991, we pilot-tested approximately 200 open-ended science items with eleventh graders in sixty Connecticut high schools. The items are composed of three types. The first consists of either a science passage to interpret or some open-ended questions to which to respond. See Figure 8 for an example of this item-type. The second item type consists of a data set to interpret. Students may be asked to construct a graph or a table and draw some conclusions from data. The third type of item is a "hands-on" experiment that students are required to design and conduct. While working, each student is observed by a trained external assessor (a retired science teacher from a different school district) who determines whether the student has designed a valid and reliable experiment and the extent to which he or she understands the relevant science content.

The mathematics tasks were developed largely by a team of mathematics educators within our department. They consisted of contextualized problems with several possible solution paths or strategies. Students were asked to communicate their reasoning to a specified audience (e.g., another student, a younger child, or an adult other than a mathematics teacher). Connecticut teachers were then convened to respond to the items and suggest improvements. During the winter and spring, the items were then reviewed by additional experts in mathematics assessment. In May 1991, we pilot-tested eighty-one open-ended items with eleventh graders in forty Connecticut high schools. (See Figure 9 for an example of a Component 2 mathematics task.)

The pilot test design provided us with between two and four classrooms of students responding to each group of items. Students responded to approximately seven tasks and also provided us with a list of courses they had taken and grades received in those courses. Teachers and students reported their reactions to the items.

Some Prerequisites for the Effective Use of Performance-Based Assessments

In reflecting on what we have learned over the past two years from listening to teachers and students participating in both components of the Common Core of Learning Assessment program, it

seems obvious that new assessment approaches by themselves are insufficient. We will need to supplement new assessments with:

- Significant and sustained professional development opportunities to provide time for teachers to: identify the "big ideas" in their discipline; understand and develop a new vision of learning and teaching; develop a repertoire of new instructional strategies, and develop a sense of efficacy;

- Permissions from state and school administrators that "less is more" and that the job of teachers in David Hawkin's words is not "to cover the curriculum but to uncover the curriculum (Duchworth, 1987);

- New curriculum materials that support depth over breadth;

- Appropriate stakes and incentives so that administrators, teachers, and students will be willing to take risks and try new approaches;

- Time for teachers to develop new assessment tasks and refine them through the many iterations required:

- Time for teachers to develop shared understandings of quality and to have conversations about how to provide their students with rich opportunities to foster it;

- Time for teachers to score students' work and develop common standards.

In addition to the foregoing:

- Other high stakes tests may also need to change. We frequently hear from teachers: "We think this is the right way to teach and assess but we are too busy preparing our students to take the College Board Achievement Tests, and

- Some restructuring may be required to provide opportunities for students and teachers to achieve the higher standards we value: e.g., different configurations of class time will be required for more sustained student projects and conversations. Finally, common planning time will be necessary for teachers to work with other teachers and/or other content experts to understand what quality is and how to best achieve it.

Implications of Using Performance-Based Assessment for Students with Limited English Proficiency

Performance assessment, as developed by Connecticut, is multi-faceted. It intentionally integrates content and procedural understandings with skills in problem solving, communication, and collaboration. It strives for ecological validity in determining what society values and then developing tasks which foster and assess those values most directly. This results in a strong emphasis on language skills. Students talk with one another in small groups and are called upon to communicate their findings to others at the end of their investigation. Their work rests on a foundation of content understandings. Before students can design an experiment, they have to have some knowledge about the subject of the experiment. If one uses a gate-keeper metaphor, content may serve as a gate-keeper for process, and communication skills may act as a gate-keeper for elucidating what one knows and understands. These gate keeper relationships are present for all students being assessed through the kinds of multi-faceted performance assessments advocated in this paper.

An interesting paradox surfaces in trying to build ecologically sound performance tasks. On one hand, as a society, we place high value on students being able to communicate their understandings effectively (e.g., NCTM Standards); on the other, we are concerned about the ability of minority students and students with limited English proficiency to do so. Which is more unfair -- creating high expectations for all students, while knowing that some will have difficulty, or creating relatively lower expectations for everyone, knowing that in their wake, some groups of students will not have access to demanding curricula? The answer to that question is related to the stakes imposed by the tests. If stakes are high and students are punished by poor performance on the assessments, it seems unfair to set expectations that will present hardships for certain subgroups. However, if stakes are low and better educational experiences are likely to result because of the mere existence of the assessments, then it seems unfair to deprive the groups most in need of enriched communicative experiences of those opportunities. This paradox must be addressed as states and local districts consider implementing performance-based assessments which require effective communication skills. (Linn, Baker, & Dunbar, 1991 include an interesting discussion of fairness.)

I will close as I began. Alternative assessments have grown in popularity, in part, because of the growing dissatisfaction with the fragmented and artificial multiple-choice tests that have been dominating our classrooms. Teachers have felt frustrated under the pressure to prepare their students for tests that are considered by them

of dubious value. As we work toward developing alternative forms of student assessment, we must take steps to provide adequate professional development opportunities, appropriate stakes and incentives, and sufficient time and space for these innovations to take root and grow. Those of us involved in this arena of school reform believe that this endeavor is doable, difficult, and worthy of pursuit.

Notes

1. Many of the ideas in this paper resulted from my work on the Connecticut Assessment of Educational Progress Program, the Connecticut Mastery Testing Program, and the Connecticut Common Core of Learning (CCL) Assessment Program funded by the Connecticut State Department of Education. The CCL program is also funded in part by the National Science Foundation. (SPA-8954692). Many external contractors assisted the CSDE in its work and the help of these organizations is gratefully acknowledged: Advanced Systems in Measurement and Evaluation (CAEP: Science, Business and Office Education, English Language Arts), Educational Testing Service and Scholastic Testing Service (CAEP: Foreign Language), National Evaluation Systems (CAEP: Art and Music), National Occupational Competency Testing Institute (CAEP: Drafting, Graphic Arts, and Small Engines), The Psychological Corporation and Measurement Inc., (CMT). I am grateful to my colleagues at the CSDE Common Core of Learning Assessment Program for their dedicated work. The science team consists of Jeffrey Greig, Michal Lomask and Sigmund Abeles; the mathematics team consists of Bonnie Laird Hole, Susan Dixon, and Steven Leinwand. Douglas A. Rindone has provided invaluable direction for the project with the able assistance of Claire Harrison, Steven Martin and Arlene Morrissey. However, any opinions expressed in this paper are my own and are not meant to represent the views of the funding agencies, the contractors, or my coworkers.

2. In 1989, the Connecticut State Department of Education received a grant from the National Science Foundation which supported Connecticut teachers and curriculum specialists to work collaboratively with colleagues from six other states (i.e., Michigan, Minnesota, New York, Texas, Vermont, and Wisconsin) and the Coalition of Essential Schools to develop performance assessments for high school mathematics and science. After the first year, teachers from sixteen large urban school districts in the Urban Districts' Leadership Consortium of the American Federation of Teachers (including Albuquerque, NM, Cincinnati, OH, Cleveland, OH, Dade Country, FL, Detroit, MI, Hammond, IN, Kansas City, MO, Los Angeles, CA, Newark, NJ, New Orleans, LA, Philadelphia, PA, Pittsburgh, PA, Rochester, NY, Saint Paul,

MN, San Francisco, CA, and Washington, DC) and five states from Project Re:Learning (i.e., Arkansas, Delaware, New Mexico, Pennsylvania, and Rhode Island) joined the Connecticut multi-state project.

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. (1990a, April). How science is tested and taught in elementary school science classrooms: A study of classroom observations and interviews. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Baron, J. B. (1990b). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In A. B. Champagne, B. E. Lovitts & B. J. Calinger (Eds.) This year in school science: Assessment in the service of instruction. Washington, DC: American Association for the Advancement of Science.

Baron, J. B. (in press). Strategies for the development of effective performance exercises. In R. Stiggins and B. Plake (Eds.) Applied measurement in education.

Baron, J. B., Forgione, P. D., Rindone, D. A., Kruglanski, H., & Davey, B. (1989, April). Toward a new generation of student outcome measures: Connecticut's Common Core of Learning Assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Brown, J.S., Collins, A., & Duguid, P., (1989, January-February). Situated cognition and the culture of Learning. Educational Researcher. 32-42.

Burstall, C. (1989. June; 1990, June). Update on the new national assessment in Great Britain. Paper presented at the Large Scale Assessment Conference sponsored by the Education Commission of the States and the Colorado State Department of Education, Boulder, CO.

Charles, R. I. & Silver, E. A. (Eds.). (1988). The teaching and assessing of mathematical problem solving (Vol. 3). Reston, VA: Lawrence Erlbaum Associates and the National Council of Teachers of Mathematics.

Connecticut State Board of Education (1987). Common Core of Learning. Hartford, Connecticut.

Connecticut State Department of Education (1982). Connecticut Assessment of Educational Progress 1981-82 Art and Music Summary and Interpretations Report. Hartford, Connecticut: Author.

Connecticut State Department of Education (1985). Connecticut Assessment of Educational Progress 1983-84. English Language Arts Highlight Brochure. Hartford, Connecticut.

Connecticut State Department of Education (1985). Connecticut Assessment of Educational Progress 1983-84. Business and Office Education Summary and Interpretations Report. Hartford, Connecticut.

Connecticut State Department of Education (1986). Connecticut Assessment of Educational Progress 1984-85. Science Summary and Interpretations Report. Hartford, Connecticut.

Connecticut State Department of Education (1988). Connecticut Assessment of Educational Progress 1986-87. Drafting, Graphic Arts and Small Engines Summary and Interpretations Report and Performance Criteria Handbooks. Hartford, Connecticut.

Duckworth, E. (1987). The having of wonderful ideas and other essays on teaching and learning. New York: Teachers College Press.

Frederiksen, J. R. & Collins, A. L. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Greeno, J. (1978). A study of problem solving. In R. Glaser (Ed.). Advances in instructional psychology (Vol. 1, pp.13-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Harrison, C. (1991) Summary of student and teacher reactions to Component 1. Connecticut State Department of Education (Internal documents).

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

National Council of Teachers of Mathematics (1988). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

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 (pp. 65-75). Andover, MA: The Network.

Resnick, L.B. (1988). Learning in school and out. Educational Researcher, 16(9), 13-20.

Resnick, L.B. (1989). Teaching mathematics as an ill-structured discipline. In R. I. Charles & E. A. Silver (Eds.), The teaching and assessing of mathematical problem solving (Vol. 3, pp. 32-60). Reston, VA: Lawrence Erlbaum Associates and the National Council of Teachers of Mathematics.

Schoenfeld, A. (Ed.) (1976). Cognitive science and mathematics education. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shepard, L.A., Why we need better assessments. Educational Leadership, 46(7), 4-9.

United States Department of Education. (1991). America 2000: An education strategy. Washington, DC: Author.

Vygotsky, L.S. (1978) Mind in society.: The development of higher psychological processes. (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans. and Eds.) Cambridge, MA: Harvard University Press. (Original work published 1935)

Wertsch, J. V. (Ed.) (1985) Culture, communication and cognition. Cambridge: Cambridge University Press.

Wiggins, G. (1989) A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Wolf, D. (1988) Opening up assessment. Educational Leadership. 45(4), 24-29.

Response to Joan Baron's Presentation

Mary Jean Habermann
New Mexico Department of Education

Thank you, Rene, for a short introduction. Twenty minutes is a short time to talk, and five minutes of introduction takes away from this time. In relation to this particular topic, I wish to identify, first, some of the strengths of the system of performance based testing from the perspective of the practitioner. Then, I would like to briefly describe the functions of language as part of learning and outline some applicable points described in the development of alternative assessment that we are doing in New Mexico, as applied to native American languages. My final comments will center on some of the implications of performance based testing for students in programs of bilingual education.

The Connecticut Assessment of Educational Progress and the Common Core of Learning Assessment Programs contain performance-based assessment tasks for high school students. I appreciate, Dr. Joan Baron, the extensive set of materials you sent me, which provided the rationale and supportive research base for this form of testing, designed, and I quote from the materials, "to determine what students know and can do." Dr. Joan Baron likens performance assessment to "a blurring of the edges among assessment curriculum and instruction."

As a former teacher, who has dedicated all of my professional career to teaching in and through two languages, I like that definition. Having been a classroom teacher for many years and also a bilingual specialist responsible for observing bilingual instruction given throughout the state with the New Mexico Department of Education, it gives me great personal and professional pleasure to discuss performance testing from this point of view. I am not an expert in evaluation nor do I claim to have deep understandings of the technical aspects of evaluation. My comments, then, in this area will relate to the purposes of teaching, and therefore, to assessing what is taught, first, for the average English speaking child, and later, in the context of bilingual learners or for those who are becoming bilingual.

I use the latter terms in reference to these students because I know that becoming bilingual is, indeed, an expansive intellectual experience for any individual, a means by which one is able to use two linguistic and cultural systems to negotiate one's world and one's place in it. Culturally speaking, a bilingual individual is able to live, act, and participate in cultural events conducted in English and/or a language other than English, whether through literature, traditions, government, music, art, or any arena. An individual who is bilingual

can be home in diverse language communities of the United States or in the countries of the world which speak the other language.

We all know that the addition of a second, third, or fourth language is expected and valued as a sign of a well-educated individual, well-cultivated individual in many countries of the world. To me, the term limited English proficient has always presented a much more limited view of the individual's intellectual and linguistic potential.

In terms of the testing process, tests can provide teachers one definitive means to ascertain whether students understand the concepts and skills being taught and also the degree to which they are learned; thus, they must be tied directly to the curriculum. Through testing, a teacher can continually reassess the teaching methods he or she utilizes and then reteach and recycle the skills and concepts needing attention.

Of the various concerns voiced by the general public regarding standardized achievement tests, one is the mismatch between achievement test results and the progress of students reported by teachers. Since teachers, however, use motivational factors and criteria to make judgments about student progress, and a paper and pencil achievement test does not and cannot, this disparity, then, will naturally exist. What these tests do provide is a measure of individual performance relative to a given set of standards. We must never forget that. It is relative to a given set of standards. These standards represent the skills and concepts deemed important for learning in the curriculum, and that curriculum represents a general American curriculum.

Annual assessment of student achievement using a standardized measure provides the teacher, the program, the district, and the state a status report, or "product" measure of a given performance at a given point of time compared to a stated expectation. With this orientation, comparability of achievement on a standardized measure can be established between the population tested and other groups nationwide, statewide, district-wide. The data produced is also useful to analyze performance trends of a given population, annually or longitudinally.

Now, a process orientation uses data generated by the measure diagnostically; that is to pin point and refine elements within the program of instruction to teach to those needs, not to teach the test...to teach to the needs. Since items tested represent items that may not have been taught, teachers have always known that test results do not necessarily represent an objective measure of what the student really knows, nor should they be interpreted as such.

It is indeed good to know that a state agency is designing and field testing an evaluation process, which ties what is taught to the items being tested. It is also tremendously important because this can foster greater accountability, as Dr. Joan Baron described, on the part of teachers. It gives the teachers responsibility for teaching, and it also gives them the tools to assess student learning using a uniform set of factors that are tied both to cognitive and affective domains.

The testing system I reviewed in the materials Dr. Baron sent to us has been designed for secondary English speakers. Once again, as a former elementary teacher in a program of bilingual education who also coordinated instruction in a secondary bilingual program, I saw how critical student involvement is for learning at all levels, for all students, regardless, of one's proficiency in the dominant language of the country. It is both valuable and valid for secondary students learning content area material. Why?

Society has changed dramatically in recent years and the demand upon the schools in preparing students to function effectively in this world have also changed. Students no longer need to just "know" facts and practice skills taught in the schools; rather, they need to know how to access information, how to evaluate it, and abstract and apply the "facts" directly to real life contexts. They need to learn how to think, how to problem solve, how to question, how to make judgments, and how to do so in a reasoned way. They need to know how to read and write, using standard grammatical forms for specific purposes, and they need to know the principles governing mathematics and science.

Students today live in a society where trends which influence them change as rapidly as they can flip the switch on the VCR, the TV, or the stereo system. This environment gives students, today, more control over their own interests. In a secondary classroom, many English speaking students seem to show difficulty attending to a lecture given about a topic unless it has immediate relationship to this instantaneous lifestyle. A teacher must almost become a magician to spark the interest of secondary students for the adult world they will enter. The type of assessment describe by Dr. Joan Baron is intended to involve secondary students then in the learning tasks while charging them with the responsibility for thinking and analyzing the material taught by the teacher.

Another aspect of the system that I find of tremendous importance is that it provides a focus on meaning. Rather then simply testing facts taught, this system tests students' ability to manipulate facts, to organize and share their knowledge, and then apply it, in highly contextualized settings.

In addition, the performance assessments prepared by the Department of Education of Connecticut appear to be very well-thought-out, based in well-founded research in testing, evaluation, and in the psychology of learning. It has also been validated through pilot testing.

I wish to commend the Connecticut Department of Education for it's leadership in this thoughtful and insightful initiative. Before I identify issues relative to this topic for bilingual learners or for those who are becoming bilingual, we must first focus our attention on the process of learning and also the relationship between learning and language, because this is a tremendously important connection.

What is learning? We talk about it all the time. There are many complex definitions, but one could say that learning takes place when the brain recognizes something in a new way. Just a little "aha"; the light goes on, so to speak. Learning is universal, and it is a unique characteristic of man resulting from his intelligence. It is, indeed, the genius of man to which we attribute the development of language because since the beginnings of time, man, a social being with intelligence, needed to communicate thoughts and ideas to others. Man's intelligence with language brought about the development of tools. These gave man leisure time for developing his artistic expression and also forms of governing, forms of educating, and forms of living as represented through the institutions within the society that evolved.

The schools represent the institution developed by man to transmit a universal body of knowledge valued by people. Now, the schools will implement a curriculum that encompasses this body of knowledge valued by society. And, in the schools, learning occurs primarily through the use of language. Whether it be Chinese, Swahili, Navaho, or English, language is the primary vehicle for learning, and students all over the world learn in and through the language they control.

For students with language and culture different from that of the schools, the desire is always the same -- that is, for the children to be successful and to accomplish learning. Bilingual, multi-cultural education recognizes that bilingual children stand to derive the same intellectual benefits that monolingual English speakers receive in the schools when instruction is given in and through their language. A well structured ESL program, part of a bilingual program, allows students to add this language to their intellectual repertoire, using methods and materials designed for second language learning.

Therefore, many of the psychological and linguistic principles of learning that apply to instruction in English will also apply to other languages. When we take this point to the point of evaluation, the

same holds true. Dr. Alan Ginsburg, this morning, said testing should not discourage bilingualism.

When the purposes of evaluation is to ascertain what "students know and what students can do," the language of the child becomes a tremendously important factor. For students who are bilingual or who are becoming bilingual, both languages must be used. The language of the child should serve as the means to demonstrate mastery and understanding of the material taught. Now, if the purpose of the test is to ascertain what command of English the students have in he subject matter areas, then, the design and content of the tasks must take on a different configuration, and the results must be analyzed in terms of lexicon, syntax, semantics for second language learners. I believe, however, that the process contained in the materials, Dr. Joan Baron sent us, would probably remain the same in terms of individual work, group work, and evaluation.

I wish to discuss some ways in which this process can be modified for bilingual students or for students who are becoming bilingual by citing an example of alternatives we have recommended in the state of New Mexico. A bit of background is needed.

The state of New Mexico is perhaps the only state in the nation where several languages and cultures are part of a population mosaic which includes American Indian, Hispanic, Anglo and other ethnic groups and whose constitution has provisions for the maintenance of a bilingual citizenry. It is also the only state where the Spanish language has been used continuously since the early Spanish settlements were established after 1538. The seven languages spoken by the American Indian people are an integral part of government, religion, and aspects of daily life among each of the tribes whose elders value the use of the language in the community and generally require proficiency in it for governance. This situation has existed in New Mexico since the dawn of the Native American civilizations. It is only in very recent times that these languages have been written. In fact, for the Pueblo languages, some of the tribal governments are only now moving in this direction. The oral tradition remains as the ever-present form to transmit the values of the culture from generation to generation. One could say their "literature" exists in the oral form.

The teaching and learning of English as a second language has been both a personal as well as an institutional need for a large number of the population since the incorporation of the territory into the national framework of the United States in the mid-1800s. The schools of the state are always searching for ways and means to incorporate methods and materials which can facilitate the acquisition of English for speakers of other languages.

Up until 1986, the state testing program, designed to assess the learning needs of students in grades 3, 5, and 8, had always been done in the English language. With the passage of the Public School Reform Act of 1986, the state formulated grade level competencies for all subject matter areas. The schools of the state were charged with designing local assessment measures in each grade level to find out whether students had acquired the competencies prior to promotion to a higher level and also to provide a remediation process for those who had difficulties. For graduation, students needed to demonstrate mastery of these competencies. We restructured the state testing program to include competency based components for grades 3, 5, and 8, which accompanied the CTBS and also designed a high school competency exam. Students who did not pass this test would be given a certificate of attendance rather than a diploma.

The State Board of Education, recognizing the large numbers of students with languages other than English at their disposal for learning, provided for the development of alternatives for these students.

In order to assist districts with these new elements in the standards and help them in cases where exemption would be necessary, the New Mexico Department of Education developed a technical assistance manual entitled Recommended Procedures for Language Assessment. We also prepared state-wide training institutes for district personnel involved in evaluation and in bilingual education. For the Spanish language, we identified standardized achievement measures currently available which correlated to the content tested in the state testing program and prepared the competency exam in Spanish to meet the needs of Spanish speakers of New Mexico.

We were faced with difficulties in terms of the American Indian languages where an oral form of the test would need to be devised. We recognized that the district would need to rely upon a person who is fluent and educated in the native language to test the student's mastery of the competencies and also seek a consultant with knowledge of testing to assist in this process.

In these cases we recommended the following procedures:

1. List each competency.

2. Analyze the concepts and/or skills required in each competency.

3. Determine items and procedures within the linguistic and cultural framework of the child which correlate to each competency.

4. Determine what constitutes mastery of the competency.

5 Pilot test.

6. Administer the instrument and document the results.

For the American Indian languages, we needed to use the language and culture of the child as the means to find out his or her knowledge of general American curricular items. In some cases, translation alone would not do because of the cultures involved. This allowed the schools to find out what the child knew of the competency within his world experience.

In summary, it seems to me that in using performance tests with students who are bilingual or who are becoming bilingual, there are elements which may need to be incorporated into the process. As provided by Dr. Joan Baron, performance tests, basically, "have three parts that involve a blend of individual work in the beginning and end in group work in the middle. The work in the middle section is done as a team to produce a group product. Through a variety of accompanying assessment tools, some written (such as checklists, optional journals, logs, portfolios) and some oral and visual (i.e., video tapes of discussions and oral presentations), students have continual opportunities to provide evidence of their deepening understanding and related reflections. In order to warrant several hours of group time, tasks must meet one of two criteria: they must provide a forum in which students can work together and talk together in ways that intensify their understanding of essential scientific or mathematic concepts and processes, and/or their structure must allow students to divide a large amount of work among the group members and report their findings to the group."

Most important to this test, then, is the inner action that occurs. Since language is the key to learning, and because culture represents a group's values about the content of the curriculum encoded through language, I believe the following elements must be part of a performance based testing program for this population.

First: If, indeed, the tests are "to find out what students know and can do," then they must utilize the language of the students, so they're able to negotiate the meaning inherent in the tasks. This means that written material must be prepared in the language other than English for students who have studied in this language and, for those who have not, this means that this must be negotiated somehow, orally, through a bridging of the concepts between the two languages. It means that team work among the students, in the middle part, may have to be done bilingually, and the teachers need to understand the meanings of that if they are to fulfill the purposes of this type of assessment. In cases where the content of the task may be alien to the culture of the child, restructuring of this content will be necessary if these tasks are to be intrinsically motivating and

have personal meaning. When the content of the task has no relevance, whatsoever, to the cultural framework of the child, we need to redesign those tasks so that they can build concept connections to the culture before we start teaching the general American curriculum. This is allied tremendously to meaning.

The second point I wish to make is that, since many of the principles of cooperative learning are being utilized in this plan, it would be wise to group English speakers with bilingual students. For students acquiring English, this will provide meaning-driven English language development in and among the four modalities of language (understanding, speaking, reading, and writing). Speakers of English will become sensitized to the other language and, perhaps, perk their interest in learning another language. For both groups, this will develop understanding among different ethnic groups and general appreciation for language and languages.

The third point is students must become sensitized to the fact that use of another language in the learning task does not apply lack of understanding nor the potential to understand English.

Fourth, I believe it will be necessary to design and pilot test a criteria to analyze student performance for these learners which will not penalize the student for English language manipulation that is not on par with English speakers. This criteria will need to be designed by a linguist who knows the language of the child and the semantic areas which may be affected. Finally, the English language arts component must contain tasks which assess English language performance in terms of the second language learner. To capture Dr. Jack Damico closing remarks, as tendered by Dr. Michael O'Malley earlier today, we will need to turn the research questions to target and evaluate "true linguistic performance," in terms of performance assessment. This, according to Michael O'Malley, means linguistic aspects are to be described by and through the tasks being done. Lastly is the fact that significant sustained professional development is needed for teachers implementing performance testing. This takes on a different dimension in the context of bilingual students, because we must not only provide teachers the training in what this means but also in the meanings of second language acquisition and the values of learning through two languages.

Response to Joan Baron's Presentation

Richard A. Figueroa
University of California, Davis

My apology to Dr. Joan Baron. I was asked approximately two weeks ago to change the nature of my presentation. Rather than address the issues she has raised, it was requested that I speak about California's emerging reforms in special education testing.

California's concerns about reforming the assessment process were "inspired" by a recent, federal challenge to a 1986 Injunction on the use of IQ tests with African American children. The injunction essentially broadened the 1979 *Larry P.* decision to cover not just black children being considered for Educable Mentally Retarded placement, but for any special education placement. The case is *Crawford v. Honing.*

In a hearing on this case in the U.S. Ninth Circuit Court (Spring, 1991), legal counsel for California informed the Court that the challenge to *Larry P.* (that African American children were unconstitutionally being singled out by denying them the right to an IQ test) may well be moot since the state was considering removing IQ from the diagnostic process in special education.

In the summer of 1991, Superintendent Honig's deputy, Dr. Shirley Thornton, asked me and my research team to help the state develop new policies and procedures in the area of assessment. [N.B., The statements in this article represent my own thinking on these topics and do not necessarily reflect those of the California State Department of Education]. I have gone through the whole cycle of being very pro-testing to gradually coming to realize that psychometric "diagnoses" for bilingual children, and possibly for all children, are really a needless, expensive mistake.

The rationale for removing IQ and possibly most psychometric tests in special education comes from four main findings.

The first is that now we can say, with considerable confidence, that we have found psychometric evidence of bias. The Court cases on test bias (*Larry P. v. Riles, Diana v. California State Board of Education, PASE v. Hannon, Crawford v. Honig*), since the 1970s, have drawn a lot of attention to this question. But most of the psychological community, especially the testing community, has been very successful in demonstrating that, in terms of psychometric evidence of bias, you cannot find it across ethnic groups. No matter whether you look at predictive validity, item analyses, reliabilities, or factor structures, you basically do not find evidence of psychometric bias.

Today I can report to you that we have begun to find this elusive quality of tests. We are finding it, or more accurately rediscovering it, right under our noses. In the early 1980s, Richard Duran began to alert us that Spanish language background seemed to have an impact on the predictive validities of college entrance test scores. In my own research (Figueroa, 1990) I found that IQ scores where very sensitive to bilingualism and that their predictive power declined in direct proportion to the degree of Spanish in the home. Because of the considerable implications from these data, I went back to the historical literature and found that, in fact, there is plenty of evidence sprinkled throughout the 1920s, 1930s, and 1940s showing similar outcomes for Japanese- and Chinese-speaking children. Recently, several studies have appeared with the same general findings (cited in Valdes and Figueroa, in press).

By the way, the latest edition of the Standards for Educational and Psychological Testing, for the first time, has a chapter on "Testing Linguistic Minorities." The opening statement is that for linguistic minorities, "every test given in English becomes, in part," an English language or literacy test. This is a momentous statement. It means that verbal (vocational, intellectual, achievement, personality) tests are biased when used with speakers of other languages. At the time of its publication this statement had little acknowledged, empirical support. Now, that support is more in evidence.

The second reason why IQ is under scrutiny in California is the tremendous misuse of the diagnostic process in special education. Hugh Mehan (et al, 1986) produced a superb little book titled Handicapping the Handicapped where he reports on his ethnographic study of the diagnostic process in special education in one school district in California. He found that school psychologists test until they find the "right" profile, the profile that verifies the referral for testing. He also found that school psychologists did not follow standardization procedures in testing. Poor practices in the administration and application of test scores was quite extensive. IQ anchored much of the socially constructed decisions in the "diagnoses" of learning handicaps.

The third reason for moving away form IQ is that the testing of children, particularly ethnic and bilingual children, really constitutes a form of medical malpractice. There is a group of adults known as school psychologists who have no medical training but who routinely make "diagnostic" decisions about medical conditions such as Mental Retardation, Attention Deficit Disorders, and neurological impairments (e.g., Learning Disabilities) on the basis of psychometric test scores. Some have suggested that the consequences of this professional activity are the wide national disparities in the prevalence rates for mild handicapping conditions. Some states have 3 percent of their public school population as Learning Disabled. Others have

7 percent. I would suggest that a plausible reason for such discrepancies is the practice of medicine without a license in the public schools. Some are suggesting that this Medical Model, which "looks" for the disabilities in the child and not in the curriculum, or the instruction, or system, may be just as implicated as the tests. Part of this speculation comes from the fact that even under the *Larry P. Injunction*, which proscribes the use of IQ with African American children, such children are still very over-represented in special education classes.

The final reason why psychometric tests are being reconsidered in the "diagnostic" process in California is financial. It costs the state approximately six hundred million dollars every three years to test the special education population. The unique quality of this expenditure is that it has absolutely no impact on instruction.

The reform of the special education assessment system in California begins with two initiatives. First the possible removal of IQ from all special education functions for all children in the public schools. Second, the removal of the current Medical Model which undergirds the assessment process. During the next two years, the state will undertake a multiple set of experiments aimed at determining which procedures will substitute the current assessment model and methods. The new system will be grounded on the following set of principles.

First, assessment will not focus exclusively on the child who is having problems in learning. As per the National Academy of Science's recommendation of the over-representation of ethnic children in special education, both the instructional contexts and the pupil's performance within these will be assessed.

Second, the current script which now governs testing, where an adult (often an unknown adult) presents a series of decontextualized, reductionist questions and tasks, will be changed. Rather than an unnatural communicative event where the tester cannot provide cues or feedback and where small verbal stimuli elicit small verbal and nonverbal responses, the new assessment procedures should provide for contextualized, verbally rich interactions over a long period of time.

Third, where the current methods elicit single language responses (since indeed there are no bilingual tests or norms available), the new procedures will allow for responses in L1, L2 or L1 and L2. As Valdes and Figueroa (in press) assert, the current monolingual testing practices may well be biased not just in what they do but also in what they fail to do, what they fail to account-for in bilingual pupils' mental repertoires.

Fourth, there can no longer be a single focus to assessment, such as "IQ intelligence." As constructivist frameworks point out, in mentation learners use multiple abilities for overcoming the limitations of short term memory, for using their own knowledge bases, for regulating their mental processes and for marshalling their available learning strategies. The new assessment tasks must allow for the use of multiple abilities and for the time necessary to engage them.

Fifth, "diagnosis" will no longer be a viable objective. Even if it were possible to determine who is not learning well because of subtractive bilingualism, because of poor instruction, because of the results of poverty, because of lack of schooling, because of limited English proficiency in an English-only classroom, because of a "communication handicap" or because of a "learning disability"; it makes little difference in terms of curricular or instructional needs (Rueda, 1989). A more viable objective would be the establishment of Optimal Learning Environments (Ruiz, Figueroa, Rueda and Beaumont, 1992) where pupils can "catch up" and return to the regular classrooms.

Sixth, the cadre of professionals engaged in this assessment process can no longer function as school psychologists currently do. The need is not for a testing technician. It is for an educational psychologist who is not afraid to know about curriculum and instruction; who can analyze the reading and writing process from children's work products; and who is willing to assess children in multiple contexts and in the psychopedagogical relationship described by Vygotzky.

As should be obvious by now, these reforms will extend quite beyond the area of assessment. The entire special education enterprise will be affected. It is very likely that even programs aimed at remedial interventions will also be impacted by these changes. As Ysseldyke and others have noted, children in these programs are indistinguishable from children in classes for the "mildly handicapped."

References

Figueroa, R.A. (1990). The Assessment of Linguistic Minority Group Children. In C.R. Reynolds and R. W. Kamphaus (Eds). Handbook of psychological and educational assessment of children: Volume I. intelligence and achievement. New York: Guildford Press.

Mehan, H., Hertweck, H. and Meihls, J.L. (1986). Handicapping the handicapped. Palo Alto: Stanford University Press.

Rueda, R. (1989). Defining mild disabilities with language minority students. Exceptional Children, 56, 121-129.

Ruiz, N.T., Figueroa, R. A., Rueda, R. and Beaumont, C. (1992). History and status of bilingual special education for Hispanic handicapped students. In R.V. Padilla and A.H. Benavides (Eds.) Critical Perspectives on Bilingual Education Research. Tempe, AZ: Bilingual Review Press.

Valdes, G. and Figueroa, R.A. (In press). The nature of bilingualism and the nature of testing. New York: Ablex.