DOCUMENT RESUME

ED 372 089 TM 021 751

AUTHOR	Jorgensen, Margaret
TITLE	Alternative Assessment from the Perspective of Six Different School Systems--An Experience in Consensus Building.
INSTITUTION	Educational Testing Service, Atlanta, Ga.
SPONS AGENCY	National Science Foundation, Washington, D.C.
PUB DATE	Apr 94
NOTE	52p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Decision Making; Documentation; *Educational Assessment; Educational Objectives; Educational Theories; Mathematics Instruction; Models; *Portfolios (Background Materials); *School Districts; Science Instruction; Standards; *Test Construction; Test Use; Training
IDENTIFIERS	*Aggregation (Data); *Alternative Assessment; Authentic Assessment; Consensus; Georgia

ABSTRACT

An overview is provided of the "Authentic Assessment for Multiple Users" project funded by the National Science Foundation to determine whether portfolio assessment can be structured to permit meaningful aggregation for multiple hierarchical users. The research focus was in the context of science and mathematics instruction in grades three through six in six Georgia county school systems. The study uses a model that articulates content-dependent characteristics such as rationale, standards, and judgment, as well as content-independent characteristics such as activity and media. Experiences in these school districts indicate that the theoretical model for consensus building in constructing portfolio assessments appears to be working and provides a structure for decision making that is useful for both novices and more experienced assessment developers. Emerging from the research is a notion of structured portfolios that calls for a core of structured documentation strategies with varying content and activities that depend on the classroom or the student group. While evidence is emerging that it is possible to define portfolio contents to ensure utility, it is also becoming apparent that the task is far from easy. Two figures and 25 tables present study findings. (Contains 10 references.) (SLD)

Purpose

The pur ose of this paper is to provide an overview of the "Authentic Assessment for Multiple Users" and to report how six diverse and distinctly different public school systems found a common ground in the area of alternative assessment -- one that met the needs of the teachers/researchers and one that supports aggregation.

Project Overview

The "Authentic Assessment for Multiple Users" project, funded by the National Science Foundation, was designed specifically to determine whether portfolio assessment can be structured to permit meaningful aggregation for multiple hierarchical users. This research focus is in the context of science and mathematic instruction at the third-, fourth-, fifth-, and sixth-grade levels in six Georgia school systems. The term "portfolio assessment" was used because from the onset, this research was intended to produce multiple sources of documentation of student learning, those that in combination provided an adequate and complete description of each student while, simultaneously providing a meaningful basis for aggregate analysis.

For this research, "portfolio assessment" is considered to be a data collection device that can and should contain samples of student work about which meaningful judgments can be made. The specific operational definition is:

> A (student) portfolio is a purposeful collection of student work that exhibits to the student (and/or others) the student's efforts, progress, or achievement in (a) given area(s). This collection must include:
>
> - student participation in the selection of portfolio content
> - the criteria for selection
> - the criteria for judging merit and
> - evidence of student self-reflection[2]

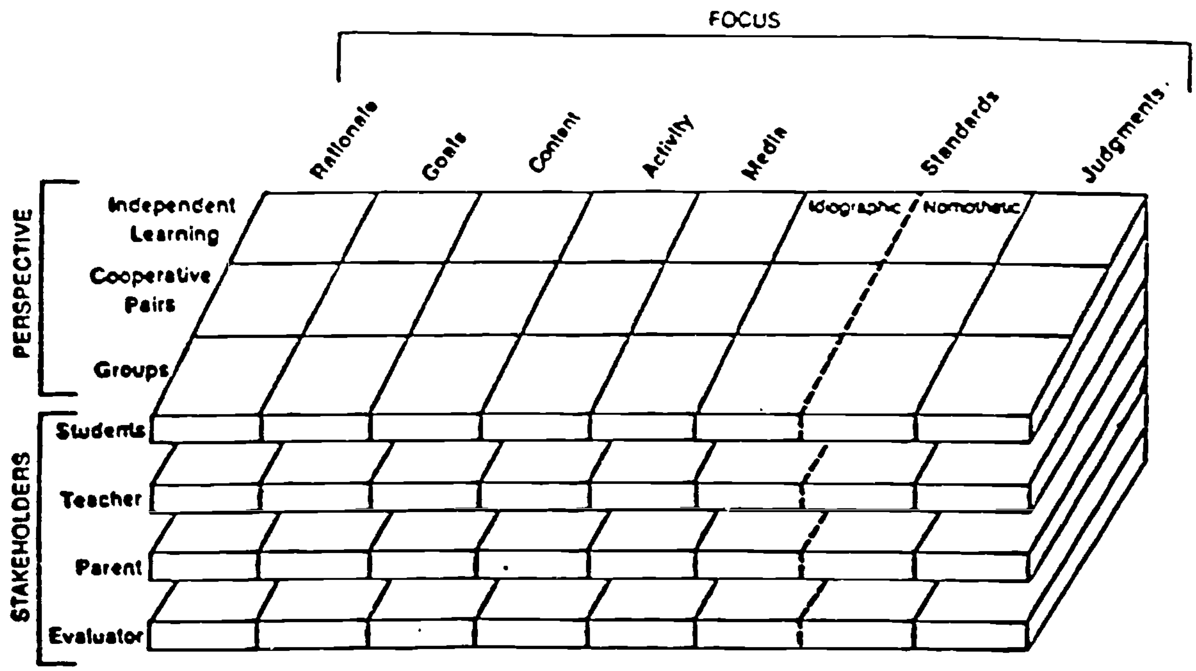
These collections were interpreted to be of virtually unlimited variety given state-of-the-art technology, creativity, instructional relevance, and sound measurement practice. Implicit in this concept, however, is that collection, selection, and reflection are desirable descriptors of both what goes into the portfolio to become assessments and how the stakeholders use the portfolio entries.

---

[2]Arter and Spandel, June 1991

Theoretical Framework

The development framework of portfolio assessments for multiples users in this study derives from the work of Paulson and Paulson (1990). Beginning with their Activity, Historical, and Stakeholder dimensions, the principal investigator for this proposed research reconceptualizes these dimensions to articulate the evaluation context, the situation in which the learner is placed, and a more inclusive definition of stakeholder. Thus, the model under study articulates the content-dependent characteristics such as rationale, standards, judgment per Paulson and Paulson (1990), and the instructional objective and content areas as well as some content-independent characteristics such as activity and media. The situation in which the assessment occurs is described in terms of student groupings (i.e., independent learning, study by cooperative pairs, group work). And the stakeholder dimension is expanded to include parents. This framework is used to guide the assessment developers through a decision-making process that results in a consensus about all dimensions of a portfolio design that can be adopted by multiple users in both hierarchical and horizontal environments.



The model under study has theoretical appeal because it suggests a structure within which clearly articulated decisions can be made. And if decision rules are articulated, the "rules" for aggregation should follow. This study is examining the practical utility of this model.

The **Focus Dimension** introduces critical controls for portfolio assessment. It specifies the rationale, educational objectives, content area(s) to be tapped, eligible activities (i.e., experiments, narrations, simulations, drawings, speeches), eligible storage devices (i.e., paper, diskette, audiotape, videotape), standards (both idiographic and nomothetic), and the type of judgments that will be made after the activity (i.e., grades or scores to be assigned).

The **Perspective Dimension** identifies the setting in which the behavior occurs. It defines the level or degree of autonomy in which the behavior is made manifest. For example, the teacher developing the portfolio assessment would specify which type of activities would be most appropriately undertaken by cooperative pairs, by small or large groups, or by the individual student. This dimension has particular importance in determining the types of standards and judgments that can be made with the information collected.

The **Stakeholder Dimension** clarifies the intended audience. For example, if a portfolio assessment is designed for classroom use rather than for multiple users, a different emphasis in the standards and in the judgments made should be expected. Students should set personal standards, perhaps using baseline samples of their own work, and make judgments about personal growth. In assessments designed to go beyond a single classroom, this type of standard would not be useful.

The paradigm for this research project provides the teachers/developers with a framework for portfolio assessment. It provides a structure for planning that theoretically should optimize the possibility that the assessment will work effectively for multiple users and that its application will produce meaningful aggregate data. Further, this model defines the elements of portfolio assessment independent of specific context, content, grade level, learner characteristics, or activity. It also views the assessment as multidimensional, clarifying variables that interact in the design, implementation, and evaluation of student behaviors.

This adaptation of the Paulson and Paulson model is being used to structure a process of consensus building among teachers, students, parents, and evaluators. Each portfolio assessment entry is being developed by consensus with each perspective represented in the model. These perspectives emphasize the summarizing and integrating of information for evaluating curriculum and for instructional decisionmaking. Consensus is built regarding the dimensions of the portfolio that are likely to impact meaningful aggregation. For example, the participants are guided through the model with the understanding that the product of their work must be an assessment activity that support use by each member of the team. This means that the decisions about what constitutes a portfolio and its purpose(s), when entries are made, who selects entries, how they are "scored," what standards are used, and how the aggregated portfolio information at the student, classroom, and school levels are communicated and used must be made by a consensus of users at each level of the model.

Project Partners

The project partners include Educational Testing Service (ETS) staff, ETS advisors, school system representatives and school-based teams, external advisors, and external evaluators.

ETS Staff

At the time the project was funded and through the first two years, the project staff included Roberta Camp, Ted Chittenden, Marty McDevitt, and Terry Salinger. Ms. Camp and Dr. Salinger are both well-known in the area of portfolio assessment. Ms. Camp was heavily involved in the ARTS Propel project. Dr. Salinger is a traditional test developer as well as a frequent consultant to school systems in the area of language arts portfolios. Dr. Chittenden is a science educator, test developer, and consultant in the general area of documentation of student learning to inform instruction. Ms. McDevitt is an experienced test developer in both traditional and innovative types of language arts assessments. Dr. Margaret Jorgensen, the principal investigator for this research, is also an experienced test developer with considerable experience working with teachers and administrators in the area of performance-based assessment for classroom use.

As the project moves into its final year, the project needs have changed. Instead of expertise in defining the assessments, we are now in need of expertise in scoring and managing the information from the student performances. Concurrent with this new need, Ms. Camp and Dr. Salinger have left ETS. Thus, to better meet the current needs of the project, we recruited Ms. Barbara Voltmer, Director of the Essay Scoring Office at the ETS Bay Area (California) Office. Ms. Voltmer will assist us in training, scoring, and the analyzing student performances.

Similarly, we have found it necessary to increase contact time between the school teams and subject matter specialists. Thus, science and mathematics experts have joined the project as consultants to work directly with the school teams.

Internal Advisers

The internal advisers include Henry Braun, Vice President for Research at ETS; Nancy Cole, Executive Vice President for ETS; and Richard Noeth, Vice President for the Field Service Division of ETS. Each of these individuals was involved in the decision to propose this work to the National Science Foundation and their support of this project is evident in their continuing role.

External Advisers

The external advisers bring to the project unique and important perspectives from outside the measurement community. Dr. Anneli Lax has recently retired from the Courant Institute of Mathematical Sciences at New York University. Dr. Richard Lesh is current both a Senior Research Scientist at ETS in the area of mathematics education and consultant with the National

Science Foundation. Dr. Michael Padilla is Chair of the Science Education Department of the University of Georgia as well as being active in other significant projects related to reform.

External Evaluators

Drs. Pearl and Leon Paulson, the developers of the model upon which our theoretical model is based, are serving as external evaluators. What they might lack in objectivity is more than offset by their knowledge about portfolios, about measurement, and about the notion of aggregation as an important outcome of portfolio use.

School Partners

The project began with six Georgia school systems: Clarke County, Dade County, Fulton County, Gwinnett County, Marietta City, and Richmond County. In terms of expenditures for education, enrollment data, pupil-teacher ratios, racial and ethnic diversity, and level of teacher training, these systems are diverse and likely to represent a reasonable cross-section of the state. As indicated in TABLE 1, there is considerable variability in the demographics and financial commitment to education across these systems.

TABLE 1

School Systems	Cost per Child (based on 90-91 data)	Student Count (FTE)	Number of Schools	Number of Teachers	Percentage of Minority Students	Percentage of Advanced Degrees in Teacher Pool
Clarke County	$4,901.08	10,294	15	650	52%	79%
Dade County	$3,654.71	2,210	4	150	1%	20%
Fulton County	$5,293.33	47,000	53	2,500	49%	56%
Gwinnett County	$3,767.50	72,500	60	4,100	14%	58%
Marietta City	$4,888.36	5,480	9	2,500	50%	60%
Richmond County	$3,790.78	34,506	54	1,951	64%	40%

Each of these six systems had some exposure to innovative assessment practices prior to participation in this project: All are either involved in or moving towards system-wide use of portfolio assessment. However, the level of knowledge about implementing an innovative assessment program as well as about the underlying assumptions of such a shift in assessment practice varied, which is representative of school systems both in Georgia and across the country.

These systems were recruited for participation in this project at the time that the preliminary proposal was being prepared for submission to the NSF. The science coordinator for each system was the contact person. Each contact person reviewed the preliminary proposal sent to the NSF and they received full copies of the complete proposal at the time that it was mailed to the NSF.

Following notification of the award, the science coordinators from each of the six systems were invited to a planning meeting (March 5, 1992). At this time, they were queried as to whether or not they were still interested in participating in the project and able to do so. Their responses were all positive. In fact, although the project could support only the work of a team of four from each system, all systems volunteered the participation of the science coordinator throughout the course of the project. And one system requested that multiple teams be included from that system. All system participants were reminded that this project was indeed research and that preliminary positive results should be found before expanding the scope of work. However, the enthusiasm and belief in portfolio assessment were clearly expressed and noted by all.

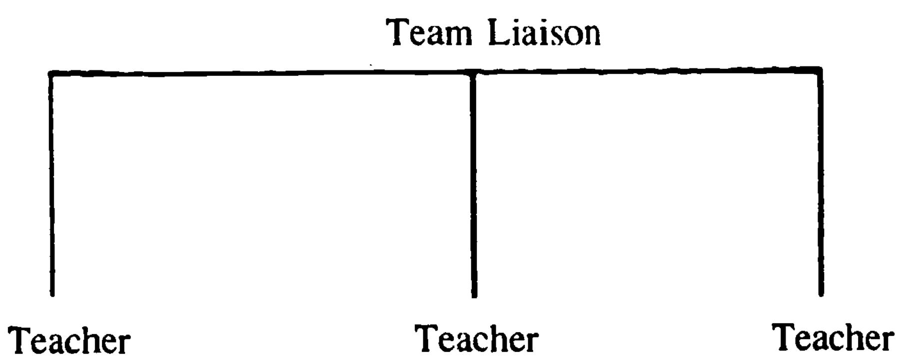
This planning meeting was critical in reaffirming the systems' commitment to the project. By so doing, each system publicly acknowledged that the teachers and the students who would be participating in the project may require special consideration regarding system-wide plans for both instruction and assessment. They also agreed to support the absence of teachers from the classroom for project-related meetings as well as the obligation to obtain written permission from all participants for all aspects of this project. Although these issues may seem trivial, they contribute to the visibility of this project in the local school setting. This visibility is part of the risk that each system was willing, indeed enthusiastic, about taking to move their systems forward in the area of innovative science and mathematics assessment.

The project was structured so that each system science coordinator would recruit a school-team liaison. That individual would serve as the communication link between the ETS project staff and the three teachers who completed each school team. The school-team liaison could be recruited from any position or role at the school level that the system coordinator thought appropriate. Five of the six school-team liaisons are building level administrators. One is an instructional lead teacher.

The grade-level focus for this project is three through six. The content-area focus is science and mathematics or an interdisciplinary or thematic approach to these areas.

The relationship among partners on this project is depicted in Figure 1.

FIGURE 1



The Team Liaison is the primary contact between the ETS project staff and the school teams.

The system science coordinator recruited the team liaison with an interest in maximizing the success of the project. The team liaison then recruited the teachers in consultation with the system science coordinator. As indicated in TABLE 2, the teachers were identified primarily because of their willingness to participate, their instructional expertise, and their commitment to quality and to change.

TABLE 2

School Systems	Why Teachers Were Selected
Clarke County	• Volunteers
Dade County	• Teacher leaders in grades 3, 4, and 5 • All members of the Total Quality Management Team
Fulton County	• Teachers looking for new challenges • Teachers considered experts in hands-on instruction • Teachers challenged by exceptionally able students
Gwinnett County	• School population characterized by diversity and at-risk students • Teachers committed to change • Teachers interested in mathematics and science
Marietta City	• Teachers committed to change • Teachers creative and open to try new things • Teachers willing to spend extra time
Richmond County	• Teachers with good mathematics background and hands-on experience • Racially balanced team

Teachers as Researchers

It is important to recognize that the teachers who choose to become partners with ETS on this project have demonstrated time after time a willingness to take chances, to be creative, and to work very, very hard. And, although we have had five resignations over the course of the first two years of the project, three were for personal reasons and two were for professional choices.

One of the teachers from the Gwinnett County team has taken a job in Chatham County (Savannah, Georgia). Rather than lose her from the project, the project staff offered her the opportunity to continue if she had the support and commitment from her new employer (county level and building administrator). Shortly after arriving in Chatham County, we received verification from the school system and building principal that the continued participation was welcome and supported. So, this project has now expanded to include an additional Georgia county.

Even more gratifying is the situation of a second Gwinnett County team member. This individual was invited to visit the Kazakh-American Lab School in Almalibak, Kazakhstan this summer. Subsequent to that visit, she was appointed to the position of Curriculum Developer for this school. And, as part of her new job, she will be responsible for developing authentic assessments to document student progress with an emphasis on portfolio assessment. (Detailed information about this school is available in Appendix B). It is with considerable excitement that the project now includes an American educator facing instructional and curricular reform in such a challenging environment.

These two individuals who resigned in order to take other jobs are indicative of the commitment to the project generally expressed across the group. It also speaks to the rich potential of this type of research project, which, ultimately, shapes and reinforces teachers to think as scientific investigators.

The Clarke County team has changed their liaison three times with a teacher now assuming that role. In addition, two teachers resigned and were replaced. Richmond County had one teacher resign, and she was replaced. The difficulty that the project staff had in contacting and interacting with either of the two designated liaisons caused problems for the team itself. Not only was the team short one person because the liaison was not available for most of the project meetings, they also experienced lags in communication from the project staff and within their team. Exacerbating the situation was the fact that two of the Clarke County team were located at an elementary school and one at the middle school. Ultimately, the project staff initiated a request to Clarke County that four teachers form the team and that two of them become the liaisons (one from each school). This strategy seems to have improved the morale of the team as well as their collaborative products.

Budget

The project was funded effective January 15, 1992, and will continue through June 30, 1994. The total budget is $ 445,506. The project-year budgets are:

TABLE 3

Year 1	$ 186,545
Year 2	$ 180,096
Year 3	$ 78,865

The scope of work for the first project year was originally planned to begin in July 1991. Due to delays in the funding process, the actual start-up of the project was January 15, 1992. This delay impacted the project rather significantly because of the schedules of the participating school systems. As a result the work was adjusted as follows:

TABLE 4

ACTIVITY	ORIGINAL DATES	REVISED DATES
YEAR ONE		
Task 1: Planning	07/01/91 - 08/01/91	01/15/92 - 02/15/92
Task 2: Training	09/01/91 - 11/01/91	03/01/92 - 04/30/93
Task 3: Consensus Building	11/01/91 - 06/30/92	08/01/92 - 04/30/93
Task 4: Process Monitoring	10/01/91 - 06/30/92	04/01/92 - 04/30/93
Task 5: Project Management	07/01/91 - 06/30/92	01/15/92 - 03/30/93

YEAR TWO		
Task 6: First Year Implementation	01/01/92 - 05/31/93	05/01/93 - 12/31/93
Task 7: Process Monitoring	01/01/92 - 06/30/93	05/01/93 - 01/14/94
Task 8: Revision and Reflection	01/01/93 - 06/30/93	05/01/93 - 07/31/93
Task 9: Project Management	07/01/92 - 06/30/93	05/01/93 - 01/14/94
YEAR THREE		
Task 10: Tryout	01/01/93 - 05/01/93	01/15/94 - 05/15/94
Task 11: Stakeholder Meeting	05/01/93 - 05/01/93	06/01/94 - 06/30/94
Task 12: Evaluation	07/01/93 - 06/01/94	01/15/94 - 06/30/94
Task 13: Dissemination	07/01/93 - 06/01/94	01/15/94 - 06/30/94
Task 14: Project Management	07/01/93 - 06/30/94	01/15/94 - 06/30/94

Work Plan

At this point in the project, the participants have been supported for 55 hours of large-group work, an average of 33 hours of on-site work, and 20 hours of scoring (including training). Across all six teams, this amounts to more than 2500 hours of work on this project. There is no doubt, however, that the participants each spent additional hours engaged in discussion and work related to this project. Evidence of this has been reported during project work sessions at ETS, on audiotapes which reveal that the teams continue discussion during lunches, etc., and on their Daily Reflections written documents.

TABLE 5 indicates the times and duration of support for the school teams.

TABLE 5

MEETINGS	HOURS	ON-SITE[3]	PURPOSE
August 10-12, 1992	20	0	Training
September 26, 1992	6	0	Training
November 17, 1992	6	8	Strategy Development
January 7, 1993	7	8	More Strategy
January 31, February 1, 1993	14	0	Rubric Preparation and Assessment Refinement
April 17, 1993	7	4	Debriefing
May-June, 1993	0	8	On-site Revision of Assessments and Rubrics and Review of Exemplars
June 11-13, 1993	20	5	Review of Student Products and Refinement of the Assessments and Rubrics
September 27, 1993	7	0	Planning for the Final Year
December 8-9, 1993	16	20	

[3] Some teams, and individuals on teams, requested special time allocations to complete assignments. These requests were always honored.

Training Highlights

The project began in August 1992 with a three-day training session. The session included an overview of the project and a brief introduction to the notion of education reform as well as a discussion about the climate for assessment reform which prompted development of "Authentic Assessment for Multiple Users." Joel Barker's video "Discovering the Future" was shown to set the tone of teacher as explorer in the quest for assessment strategies that would really tie instruction to assessment and enhance the teaching/learning environment. A consultant on the topic of consensus-building also spoke to the group early in the session.

The dynamics of the three-day session can be capsulated by the phenomenon of empowerment. The focus was to move through the theoretical model from the perspective of the teacher as stakeholder. Thus, the groups were to reach consensus at the school-team level on the Rationale for the project and the Goals, Content, Activities, and Media from the perspectives of teachers only. Entry into the model was selected at this point to mediate anxiety about the unknown, with the thought that tying the research to familiar territory would anchor the research partners.

The content base was provided through Science for All Americans (1989) and the National Council of Teachers of Mathematics Standards for Curriculum and Evaluation (1989). These documents governed the presentation of important foci for assessment. These were coined the "Big Ideas":

- Being familiar with the natural world and recognizing both its diversity and its unity
- Understanding key concepts and principles of science
- Being aware of some of the important ways in which science, mathematics, and technology depend upon one another
- Knowing that science, mathematics, and technology are human enterprises and knowing what that implies about their strengths and limitations
- Having a capacity for scientific ways of thinking
- Using scientific knowledge and ways of thinking for individual and social purposes

Three key features of mathematics as embedded in the Standards:

- "Knowing" mathematics is "doing" mathematics
- Some aspects of "doing" mathematics have changed during the last decade, e.g., computers
- The changes in technology and the broadening of areas in which mathematics is applied have resulted in growth and changes in the discipline of mathematics itself

In addition, the notion of hard content (complex, not necessarily difficult) derives from the work of Archbald, Tyree, and Porter (1991):

> "Hard content means not just the facts and skills of academic work, but understanding concepts and the interrelationships that give meaning and utility to the facts and skills....The emphasis is on students learning to produce knowledge, rather than simply reproduce knowledge."

The strategy for training was as follows: Each school-based team was sent a list of guiding questions (see Appendix A) in advance of the training sessions. In addition, they were sent reading materials to facilitate responses to these guiding questions. The reading materials were selected because they represented state-of-the-art assessment approaches in science and/or mathematics. The guiding questions were used during the training session to anchor the participants and their understandings of innovative assessment practices and to encourage ownership in the research project. A questionnaire was administered at the beginning of the training session. Reflection opportunities were also used.

The school-based teams worked together to reach consensus first on the guiding questions and then on the cells in the model along the teacher continuum from Rationale through Media. (Standards and Judgments were to be considered once the participants had a clearer understanding of the complex cognitive outcomes to be tapped through portfolio assessment.) Once consensus had been reached within a school-based team, the six teams were disassembled into two large teams comprised of two individuals from each of the six original teams. It took two days to reach consensus within these two large groups on the Rationale and Goal statements for this project.

A review of the Rationales and Goals identified by each of the two groups is somewhat indicative of the struggle with perspective that was observed by the project staff: Group 1 began and remained student-centered. Group 2 began teacher-centered and only showed slight movement away from the traditional "teacher as doer/enforcer - students as sponge" paradigm (see TABLES 6 and 7).

TABLE 6

GROUP 1

RATIONALE:	With the recognition of the technological and societal changes and challenges of the 21st century, there is the realization of the need for change in assessment of students' progress in math and science. The use of portfolios is a means of integrating teaching and assessment, thereby enhancing scientific literacy.

GOALS:

1. To become complex thinkers, able to critically observe, investigate, formulate problems, produce solutions and evaluate outcomes
2. To become effective learners, able to identify and analyze strengths and areas for future growth in individual and group settings
3. To become self-confident and able to take risks with diminished fear of failure
4. To become collaborators in a variety of settings with diverse groups of people
5. To become experiential learners, integrating curriculum with real-life situations
6. To become responsible participants in a global society, promoting quality of life

TABLE 7

GROUP 2

RATIONALE:	To develop a method of standardization measuring student progress and achievement To increase students' responsibility for their own learning

GOALS:

1. To improve student learners' attitudes about math and science

2. To encourage innovation, higher-order thinking, creativity, and risk-taking

3. To implement a more interdisciplinary, authentic curriculum through hands-on activities and physical manipulation

4. To develop an understanding of science and math concepts by use of the scientific process

5. To produce students who are effective communicators

6. To encourage students to become self-evaluators through reflection

7. To produce students who are self-motivated and have high self-esteem

8. To provide parents a broader understanding of their child's progress

Thus, project consensus did not occur at the initial project training session. As a result, after conversations with the systems coordinators, a follow-up training session was scheduled for September 26, 1992. This session was to be used to document large-group consensus on Rationale and Goals and to move into thinking about documentation of student learning in ways consistent with the Rationale and Goals of this project.

During the period of time between the initial training session and the September session, the ETS project staff reviewed the videotapes of the training session and the written documentation in an effort to propose a compromise Rationale and set of Goals which would be adopted by consensus. These were presented to the research partners in the following form:

TABLE 8

CONSENSUS RATIONALE AND GOALS

RATIONALE:	With the technological and societal changes and challenges of the twenty-first century, there is the recognition of a need for change in assessment of students' progress in mathematics and science. The selection of portfolio entries for the evaluation of student progress allows for the documentation and evaluation of valued student outcomes. The collection, selection, reflection, and aggregation processes necessary in the development of a portfolio serve as a model, enabling all stakeholders to make purposeful evaluations.

GOALS:

To develop students who are:

- **creative and strategic thinkers**

  adept at using higher-order thinking skills, innovative in their approach to problem solving, and able to formulate questions, develop solutioı ., and evaluate outcomes

  (G-1: 1, G-2: 2,4)[4]

- **reflective thinkers and self-evaluators**

  able to evaluate their own learning through the identification and analysis of their strengths and able to determine the need and direction for growth as individual learners and as cooperative learners

  (G-1: 2, G-2: 6)

- **self-motivated learners**

  willing to take risks and self-confident as learners, embracing a positive attitude about math and science

- **effective communicators**

  (G-1:5)

- **effective collaborators**

  in a variety of settings with diverse groups of people

- **responsible global citizens**

[4]The codes that follow reference the group number and goal number used to create the consensus goals.

Of considerable interest is the discussion regarding the use of the phrase "Experiential Learner" and the distinction regarding the separation between the world of school and the world of work and whether "real-world" indicated that the world of the school was not "real." The compromise was to avoid use of "real-world" references.

Once the Rationale and Goals were accepted by the group through a consensus-building process, the school teams were directed to brainstorm behaviors which would serve as evidence that the students were "effective collaborators," "effective communicators," etc. That is, what specific learner outcomes would serve as evidence that the goals of the project had been attained? The brainstorming of the school teams then led to a large-group discussion, the results of which are reported in TABLE 9.

TABLE 9

To develop students who are <u>Reflective Thinkers and Self-evaluators</u>: • knows his/her learning style, strengths, and weaknesses • knows how to use the identified strengths/weakness of others • continually monitors and evaluates own progress and makes changes accordingly • shows willingness to regroup and try again based on self-evaluations • demonstrates willingness to articulate steps (approaches) to problem situation • demonstrates ability to recognize the act of transference from one learning situation to another
To develop students who are <u>Creative and Strategic Thinkers</u>: • uses systematic procedures/processes things systematically • uses multiple solutions • shows persistence • is inquisitive • uses open-ended approaches • uses trial and error problem solving • juggles multiple strategies • has rational plan • demonstrates flexible thinking • is able to let go/cut losses • is open minded • builds on previous knowledge • is able to access information from multiple sources

To develop students who are Self-directed Learners:

- exceeds basic requirements
- uses wait time effectively (finds something meaningful to do after completing tasks)
- makes choices and sticks to choices
- pursues own interests
- desires knowledge for self-fulfillment (rather than grades)
- moves outside of individual comfort zone
- takes initiative
- extends learning to home
- tries things in a new way
- assesses progress

To develop students who are Effective Communicators:

- is able to orally explain
- can show written evidence of work through narration, description, persuasion, and exposition
- can show visual evidence of work through diagrams, drawings, and graphs
- demonstrates ability to learn through listening and following directions
- demonstrates ability to gather information through reading and being read to
- uses technology to communicate
- uses appropriate vocabulary for math and science
- uses effective presentation skills

To develop students who are Experiential Learners:

- is involved in student-directed activities
- shares information and "things" from own environments
- initiates student experiments
- shows evidence that classroom learning is being transferred to out-of-school experiences
- has role-playing abilities
- seeks audiences
- articulates to audiences

To develop students who are <u>Effective Collaborators</u>: • recognizes and accepts self-worth and that of others • believes that the collaborative result will be better than any single effort • demonstrates respect for self and others by accepting responsibility for collaborative participation • recognizes the rights of all members to participate and have a voice
To develop students who are <u>Responsible Global Citizens</u>: • interprets, evaluates the relationship between current events, issues in daily life • shares knowledge with others • practices environmentally friendly behavior • beginning with the classroom, practices getting along with others, adhering to a set of rules - expands to school and community • demonstrates awareness of, value of diversity • participants in service activities • participants in the democratic process • identifies values, demonstrates a responsible course of action

Wit . these "evidentiary behaviors" as focal points, the school teams were challenged to develop documentation strategies[5] for portfolios that would provide archival evidence of the project goals. Their charge was to develop between four and six strategies which would, in some combination, capture evidence of the seven goals. Each team then reported a collection of documentation strategies to the group on January 7, 1993.

[5] The project staff used the phrase "documentation strategy" rather than assessment to avoid the subtle limitations which may be placed on each individual because of their existing "assessment paradigms."

In thinking about and preparing these strategies, the research partners were asked to focus on these questions:

1. What were they trying to describe and how?
2. What were they trying to document and how?
3. What were they trying to model and how?
4. Whom were they trying to inform and how?

In addition, the research partners were asked to keep in mind the fact that this research focuses on portfolio assessment. As such, the strategies must, in fundamental ways, have the characteristics of assessments. Thus, they should be systematic procedures for observing behavior and describing it with a numerical scale or category system.[6]

The documentation strategies presented in January, 1993, tended to be primarily interviews. Across all teams was a clear preference for one-to-one questioning to determine learning outcomes. Other documentation strategies included logs, letters, and lab reports.

As each group presented their documentation strategies to the large group, it became clear that without some guidance as to variations in strategies, the predominant tool would be interviews. Thus, in an effort both to maximize the possibility that at least some of the strategies would lead to reliable scoring and meaningful aggregation and to enable the group to see the impact of more than one type of assessment strategy in their classrooms, the project staff guided the selection of documentation strategies to be refined for the spring field test. The project staff also constructed two documentation strategies for use in the field test.

The determining guideline plan for the selection of documentation strategies to be refined and implemented was variation. The four dimensions for variation are time, content-dependence, stimulus complexity, and response complexity.

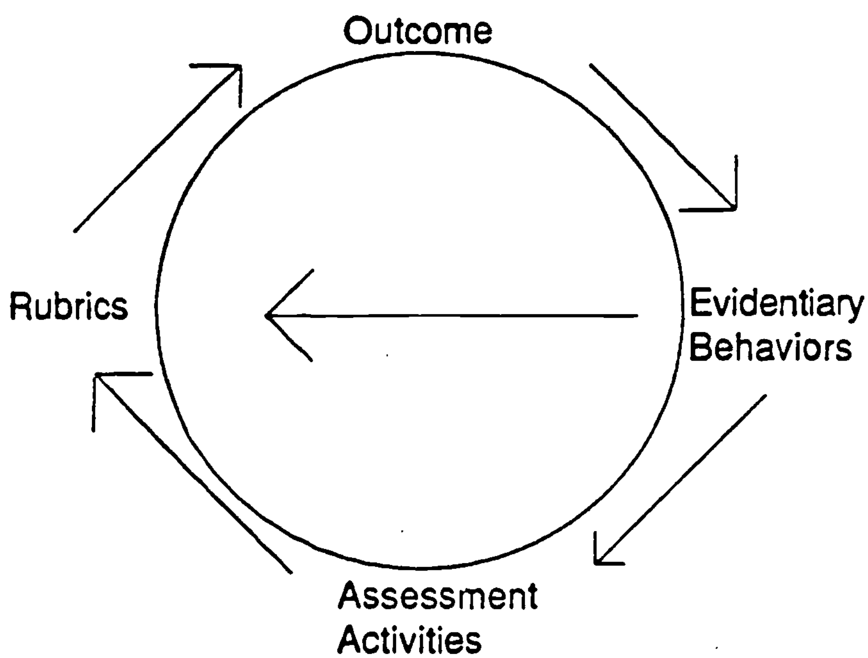
Time refers not to assessment time per se but to the amount of instructional time which would be culminated by the assessment. **Context complexity** refers to the degree to which the assessment is tied to a specific body of content rather than to broad principles or processes or concepts. **Stimulus complexity** refers to the cognitive complexity of the activity or task itself which is the "stimulus" for the resulting documentation of student learning. And, **response complexity** refers to the cognitive complexity required by the student as the evidentiary behaviors are evoked. This perspective reflects an attempt to sample across these dimensions. The six documentation strategies which were refined and prepared for field testing do reflect these four dimensions.

---

[6] L. J. Cronbach, 1970

In addition to preparing the "final" versions of these documentation strategies for field testing, the research parti.   were also challenged to develop "first tries" at a scoring rubric to be used in informing the students and parents of the valued evidence. And they were asked to map the evidence to be collected back to the project goals and to the scoring rubrics. This process of mapping appears to be an extremely valuable step in the development cycle, as it causes the developer to revisit the purpose of the assessment, the structure of the assessment and of the evidence to be collected, as well as how the evidence is going to be scored. Thus, with the mapping process, the development cycle is complete (see FIGURE 2).

FIGURE 2



At this time in the research, the "teacher as stakeholder" dimension has been explored through the development of nomothetic standards and judgments. For some research partners, work has begun on moving into the "parents students, and evaluators as stakeholders" dimensions. However, in general, it is accurate to report that the work has progressed slowly. It is also accurate to report that consensus has been less of an issue than has the design of relevant and relatively context-free assessments.

Currently, eight assessr ·nts have been developed and field-tested. Each school team administered each of these assessments during April and May, 1993. Every effort was made to ensure that some students in every class had an opportunity to perform on each assessment. The number of student responses by gender and assessment is indicated in TABLE 10.

TABLE 10

	ASSESSMENT									
	System	1	2	3	4	5	6	7	8A	8B
MALE	Dade	5	11	8	2	7	11	2	0	0
	Clarke	15	10	5	7	6	8	7	3	4
	Mar.	29	26	25	14	28	26	20	8	0
	Gwin.	31	33	21	31	28	25	0	26	26
	Rich.	35	40	47	28	41	15	0	0	12
	Fulton	25	24	7	4	4	5	1	4	4
	Total	**140**	**144**	**113**	**86**	**114**	**90**	**30**	**41**	**46**
FEMALE	Dade	10	14	5	9	7	8	3	0	0
	Clarke	10	9	8	6	5	6	7	9	3
	Mar.	34	30	31	18	30	29	11	14	0
	Gwin.	35	35	30	37	28	22	0	32	32
	Rich.	29	30	45	21	29	13	0	0	9
	Fulton	29	30	14	11	12	0	2	9	9
	Total	**147**	**148**	**133**	**102**	**111**	**78**	**23**	**64**	**53**

Grand Total by Assessment	287	292	246	188	225	168	53	105	99

(n= 1662); 26 Gende- Missing

1 = Science Observation, **2** = Retelling Applied to Word Problems, **3** = Letter Writing,
4 = Continuum of Progress Toward Goals, **5** = Toys in Space, **6** = Problem Solving, **7** = Interview,
8A = Experiment-Type (Response About Group), **8B** = Experiment-Type (Response About Individual)

Rubric Development and Revision

A scoring session was originally scheduled for June 11-12. However, as the project staff reviewed the student responses and the exemplars selected by the developing team for use in training scorers, it became clear that there was not sufficient information provided about the context complexity, i.e. the nature of the instruction, the specific instructional activities engaged in by the students, and the length of time spent in the instruction phase of learning.

The teams' responsibilities, prior to the June meeting, were to identify five representative samples of student work which characterize each score point in their rubric. Rather than use these immediately to build training materials for the scoring session, they became the focal point for discussions about which assessments evoke which kinds of responses. These samples were also the focus for discussions exploring whether or not there are certain developmental properties of the evidence which crosses rubrics (and therefore, which cross assessments).As a result, the June 11-13 meeting was used to reflect upon the scoring process, rethink the role of rubrics for each assessment, and to begin to think about a rubric or set of rubrics that might work across all categories of assessments included in the structured core notion.

The exemplars selected by the development team for each assessment served as the basis for discussion of the student responses at the June meeting. This discussion provided insights into validity links among the assessments. In turn, these validity links will be examined empirically and may spark insights into problems or successes in interrater reliability (e.g., Vermont Study, Rand, 1992) when the scoring does take place. Instead of scoring the responses, the school teams (development groups) were charged with working on-site in their teams to examine student responses across classes and schools for each assessment and to make revisions to the scoring rubrics developed for each assessment. Particular attention was paid to whether or not the student responses reveal information about science and/or mathematics knowledge or processes.

As part of preparation for the June meeting, the project staff assembled science and mathematics educators who had not been an active partner in this project. Each of these individuals was asked to work with a school team and to provide two specific resources: First of all, they were to be the subject area experts and to critique and refine any instructional flaws based on content or on the habits of the mathematics and science disciplines. Second, they were to bring a fresh perspective to the question: "What information do we expect to evoke from each assessment and how do we need to be able to communicate that information?"

Throughout this meeting, each team revised not only the scoring rubric for their assessment activity, but also the assessment activity itself for future implementation. With the revised rubric and re-selected exemplars, the project is ready to begin scoring of the student products in October, 1993. Based on the results of the scoring, assessment revision will be conducted by each team at their school site for project-wide implementation in January, 1994.

Each team of teachers selected three sample papers for each of the score points in the revised rubrics. The rubrics and sample papers were further revised by project staff in consultation with subject area specialists. These rubrics were field tested with live papers supplied by the schools and, in particular, those selected as sample papers.

In general, the final changes made to the rubrics included:

- rewording to eliminate ambiguous language,
- rewording to eliminate overlap between score points,
- eliminating constructs that were no longer included in the task, and
- simplifying the layout of the rubrics and ease of use.

It was important for all concerned to maintain the original intent of the teachers/developers throughout the revision process. Voluminous documentation of comments made throughout the revision process facilitated this effort. And, as an additional check, the revised rubric was applied to the sample papers originally selected by the teachers/developers. The most able student response based on the original rubric continued to be the most able response using the revised rubric, for example, and so forth at the other score points.

Prior to the live scoring session held in December, 1993, a project staff member and a teacher/developer scored approximately 25 papers for each of the tasks, one at a time. This exchange was designed as a pre-reading session. While it did not follow a traditional format, the purpose was to determine if the rubrics could be successfully used for more than a few papers and to further refine the rubrics as necessary. The rubrics for each task were reviewed and discussed and the two readers scored papers independently. The scores were discussed and resolution reached. Further revision, mostly fine tuning, were made and additional papers were scored independently by the two pre-readers. From this scoring session, sample papers were chosen to be used for training readers during the scoring session. When appropriate, the original sample papers were used. Samples were chosen based on consensus of score and representativeness of the types of papers readers would likely encounter. Three sets of sample papers for each task were assembled. Each set provided examples of all score points.

Scoring

The training materials for the scoring were compiled and two teachers from each of the six teams were invited to participate as readers over a two-day session. Several other individuals, not directly involved with the project, also participated as readers. This was done to provide some evidence about the transferability of the training materials to relatively naive individuals.

Training began by reviewing the each rubric one at a time and examining a set of scored sample papers. The first set of sample papers served to establish an understanding of the score points for the rubric. The scores were given and the reason for the score was discussed. The assigned scored for the second and third set of papers were not given and the readers were asked to independently score each paper. The scores were recorded and discrepancies (as compared with the "true score" estimates of the teacher/developer team, were discussed and resolution reached. The purpose of this session was to bring all readers to the same frame of reference with regard to positions along the scoring continuum.

Training began by reviewing the rubric and examining a set of scored sample papers. This first set of sample papers served to establish an understanding of the score points for the rubric. The "true scores" were given and the reasons for those scores were given. The assigned scores for the second and third set of papers were not give because the readers were asked to independently score these papers. Discussion and resolution followed.

Randomly built batches of approximately ten papers were given to each reader. When scoring was completed on each batch, they were returned, scores recorded manually and then covered, and the batch was delivered to a second scorer. When scoring was completed by the second reader, papers with score discrepancies of more than one score point were routed to a third reader.

Three of the tasks were scored during day one of the scoring session, two on the second day, and the remaining three were scored off-site on two additional days. All papers were scored twice.

The interview task was administered to only a few students during the data collection stage of this project. In those instances, a video tape recording was made of the interview. However, because of the poor quality of the amateur teacher/developer/interview as camera person, the teachers/developers decided to reconstruct the entire instrument so that it would be a more efficient tool. This reviewed instrument was field tested and those interviews were scored in early April.

Preliminary Data

TABLE 11

Toys in Space

	Score	Percent Exact Agreement	Simple r Between Raters	Kappa	Estimated Intraclass Correlation
All cases (N=250)	Prediction	74.4%	.79	.52	.78
	Drawing	68.4%	.62	.41	.56
	Narrative	50.4%	.64	.36	.63
	Contrast	64.0%	.78	.52	.78
	Question	57.6%	.66	.40	.64
	Total Score				
Grade 3 (n=59)	Prediction	79.9%	.55	.30	.51
	Drawing	67.8%	.40	.27	.19
	Narrative	54.2%	.74	.40	.74
	Contrast	67.8%	.75	.56	.74
	Question	52.5%	.61	.33	.61
	Total Score				
Grade 4 (n=91)	Prediction	85.7%	.93	.64	.92
	Drawing	69.2%	.67	.50	.63
	Narrative	50.5%	.60	.37	.60
	Contrast	61.5%	.74	.48	.73
	Question	60.4%	.55	.42	.55
	Total Score				

Grade 5 (n=100	Prediction	61.0%	.71	.44	.70
	Drawing	68.0%	.65	.36	.61
	Narrative	48.0%	.60	.31	.59
	Contrast	64.0%	.82	.53	.81
	Question	58.0%	.76	.42	.73

TABLE 12

TOYS IN SPACE
DISTRIBUTION OF RESPONSES

Value	Grade 3	Grade 4	Grade 5	Overall
0	0	0	2	2
1	0	0	4	4
2	0	4	1	4
3	1	5	2	8
4	1	7	4	12
5	4	8	4	16
6	13	7	13	33
7	9	7	7	33
8	12	15	28	55
9	20	27	17	64
10	15	27	17	64
11	7	30	24	57
12	14	20	19	58
13	16	13	19	48
14	5	7	17	29
15	0	3	9	12
16	1	0	6	7
17	0	2	1	2
18	0	1	0	2

TABLE 13

RETELLING

Score	Percent Exact Agreement	Within 1 Scorepoint	Simple r Between Raters	Kappa	Estimated Intraclass Correlation
Overall (N=303)	76.9%	.99%	.86	.68	.86
Grade 3 (n=100)	81.0%	.99%	.86	.73	.87
Grade 4 (n=97)	69.1%	.96%	.81	.56	.81
Grade 5 (n=106)	80.2%	.99%	.88	.72	.88

TABLE 14

RETELLING TABLE RESPONSE DISTRIBUTION

Overall	Grade 3	Grade 4	Grade 5	(N=303) Total
None	6	8	2	16
Attempt	51	34	25	110
Same	76	82	70	228
Most	57	55	75	187
Complete	10	15	40	65

TABLE 15

LETTER WRITING

Score	Percent Exact Agreement	Simple r Between Raters	Kappa	Estimated Intraclass Correlation
Overall (N=270)	42.2%	.59	.23	.58
Grade 3 (n=262)	41.9%	.62	.22	.62
Grade 4 (n=109)	40.4%	.59	.21	.59
Grade 5 (n=99)	44.4%	.55	.26	.51

TABLE 16

LETTER WRITING TABLE RESPONSE DISTRIBUTION

Label	Value	Grade 3 (n=62)	Grade 4 (n=109)	Grade 5 (n=99)	Overall (n=270)
No Attempt Made	0	13	17	8	38
Minimal Understanding	1	17	24	15	56
Limited Understanding	2	39	49	54	142
Satisfactory Understanding	3	43	84	70	197
Good Understanding	4	11	38	42	91
Exceptional Understanding	5	1	6	9	16

TABLE 17

SCIENCE OBSERVATION

Score	Percent Exact Agreement	Simple r Between Raters	Kappa	Estimated Intraclass Correlations
Overall (N= 309)	46.0%	.63	.28	.62
Grade 3 (n= 89)	50.6%	.53	.32	.52
Grade 4 (n= 123)	43.1%	.67	.25	.65
Grade 5 (n= 97)	45.4%	.58	.25	.56

TABLE 18

SCIENCE OBSERVATION TABLE OF DISTRIBUTION RESPONSES

Label	Value	Grade 3	Grade 4	Grade 5	Overall
No Response	0	3	48	8	59
Poor	1	17	32	32	81
Fair	2	64	87	49	200
Good	3	60	63	80	203
Very Good	4	29	16	22	67
Excellent	5	5	0	3	8

TABLE 19

COMPARISON OF EXPERIMENTS

	Score	Percent Exact Agreement	Simnple r Between Raters	Kappa	Estimated Intraclass Correlation
Overall (N=70)	Understands Concepts	52.9%	.68	.31	.60
	Extends Learning	70.0%	.65	.46	.60
	Communicates	57.1%	.55	.34	.52
Grade 3 (n=0)	Understands Concepts	NA	NA	NA	NA
	Extends Learning	NA	NA	NA	NA
	Communicates	NA	NA	NA	NA
Grade 4 (n=21)	Understands Concepts	66.7%	.53	.44	.50
	Extends Learning	64.1%	.69	.64	.64
	Communicates	57.1%	.36	.07	.34
Grade 5 (n=49)	Understands Concepts	46.9%	.70	.25	.60
	Extends Learning	59.2%	.63	.37	.58
	Communicates	57.1%	.56	.38	.52

TABLE 20

COMPARISON OF EXPERIMENTS
DISTRIBUTION OF RESPONSES

Overall	Value	Grade 3	Grade 4 (n=21)	Grade 5 (n=49)	Total
Understands Concepts	0		0	2	2
	1		20	25	45
	2		19	49	68
	3	N/A	2	16	18
	4		1	6	7
Extends Learning	0		0	10	10
	1		39	46	85
	2		3	35	38
	3	N/A	0	6	6
	4		0	1	1
Communicates	0			6	6
	1		29	35	64
	2		11	38	49
	3	N/A	2	18	20
	4		0	1	1

TABLE 21

PROBLEM SOLVING

	Score	Percent Exact Agreement	Percent Agreement Within 1	Simple r Between Raters	Kappa	Estimated Intraclass Correlation
Overall (N=190)	Understands Problem	64.2%	96.3%	.79	.34	.45
	Plans/Reports Solution	61.1%	97.9%	.64	.40	.63
	Analyzes Results	62.6%	96.8%		.44	.66
Grade 3 (n=80)	Understands Problem	65.0%	96.3	.53	.30	.52
	Plans/Reports Solution	57.5%	100%	.67	.33	.66
	Analyzes Results	NA	97.6%	.69	.46	.62
Grade 4 (n=32)	Understands Problem	65.6%	100%	.48	.36	.47
	Plans/Reports Solution	62.5%	93.8%	.63	.43	.62
	Analyzes Results	71.9%	100%	.82	.57	.79
Grade 5 (n=78)	Understands Problem	62.8%	94.9%	.38	.30	.36
	Plans/Reports Solution	59.1%	97.4%	.59	.44	.59
	Analyzes Results	57.7%	94.9%	.588	.37	.58

TABLE 22

CONTINUUM OF PROGRESS (N=145)

Score	Percent Exact Agreement	Simple r Between Raters	Kappa	Estimated Intraclass Correlation	
Focus					
A	98.0%		.83		
B	88.0%		.50		
C	66.0%		.32		
D	57.0%		.21		
Strategies					
A	81.0%		.79		
B	81.0%		.20		
C	75.0%		.49		
D	75.0%		.49		
E	81.0%		.80		
Summarizes					
A	89.0%		.78		
B	84.0%		.63		
C	88.0%		.62		
D	98.0%		.66		
E	87.0%		.54		
F	75.0%		.38		
G	98.0%		-.01		
Applies					
A	92.0%		.29		
B	95.0%		.35		
C	73.0%		.37		
D	91.0%		.47		

Issues

Consistent in the perspective of a vocal minority of the school teams is the notion that innovative assessments will ensure that all students demonstrate complex cognitive behaviors. This perspective leads to the development of lengthy and, in fact, quite burdensome, documentation strategies intended to provide students with every opportunity to produce evidence, refine evidence, collaborate, and then refine. In this way, an assessment never ends. Instead it is continuous.

In response to this perspective, the project staff has encouraged the design of assessments that are sensitive to individual differences with respect to ways of thinking and ways of doing. We have encouraged the development of assessments that enable students to be selective in terms of the response mode and to encourage the teachers to facilitate the involvement of students in the selection of the stimulus itself. However, it seems reasonable that an assessment should be constrained by time in some way. Recognizing that the assessment may be intended to take place over an extended period of time (e.g., multiple class periods), at some point the "end" for the purpose of scoring must be defined. That is, of course, not to say that there is no future, no hope for improvement.

It is also the position of the project staff that the assessments must not be more of a burden than they are a source of meaningful information. In other words, the amount of effort required in the documentation of evidence must not exceed the value of the evidence provided. Thus, it is appropriate to question the "value" of one-on-one interviews in terms of burden to administer for both student and interviewer and burden for documenting the interviews and the consequent burden of summarizing or scoring the documentation.

To remind the research partners of these issues, the wisdom of both the measurement community and the world of science are cited. First, the observation reputed to be that of Albert Einstein:

> "Not everything that counts can be counted,
> and not everything that can be counted counts."

Second, the warning from Richard Snow:

> "No matter how you try to make instruction
> better for someone, you will make it worse
> for someone else."[7]

---

[7]Richard Snow, Abilities, Motivation, and Methodology: The Minnesota Symposium on Learning and Individual Differences, 1989. (Snow's Law of Conservation of Instructional Effectiveness).

Both of these observations helped the project partners refocus on the measurement properties of portfolio assessment. This is critical because it is so easy to slip from assessment models to instructional feedback models. This project focuses on the former and, as such, is trying to define a portfolio strategy which behaves as good measurement. By that is meant it provides systematic information about student behavior, which can be summarized (and therefore aggregated) in a meaningful manner. Implicit in this notion is that the information provides a meaningful, descriptive picture of learning upon which a judgment can be made. That suggests that the information is representative of the varieties of learning that occur within the school environment. It is important that any assessment is subject to constraints of time or other parameters which will eventually reflect certain limitations.

A second issue of concern is the absence of evidentiary behaviors for any Goal that articulates student learning in terms of the knowledges and processes of science or mathematics. Although the Goals derive from the philosophy underlying the NCTM Standards (1989) and Science for All Americans (1989), the direct and explicit linkages are missing. Thus, immediate work must begin on expanding the evidentiary behaviors to articulate theses explicit linkages.

Discoveries Along the Way

As the project staff has worked with the school teams, four categories of problems have emerged. These are misunderstanding the model, interpersonal dynamics, inability to internalize portfolio assessment, and frustration with the complexity of the project. In TABLE 9, these problems have been listed along with "solutions" tried during the course of the project.

TABLE 23

Problems	Solutions
Misunderstanding Models and Groups	Clarify Provide Specific Examples Revisit Modeled Behavior
Group Dynamics	Restructure Groups Set "Rules" and Time Limits
Paradigm Paralysis	Barker Film ("The Business of Paradigms" and "Visions")
Frustration	Ownership and Pride Tension Between Generic Approach and Content Demands

Each of the problems listed above are fundamental obstacles to reform of any kind. The "Misunderstanding Models" is characteristic of a lack of knowledge. This lack can be addressed by infusing information. But, as this project has revealed, it has been essential to clarify, provide specific examples, and to directly model the desired behavior. To support variety of knowledge presentation, we have provided information via videotape, printed materials, oral presentations and analogies, expert speakers for the large group, and expert consultants to work with the school teams. We have encouraged discussion, have reviewed the Daily Reflections for the purpose of raising discussion points, and have encouraged informal contact over the telephone or through letters, faxes, etc.

Relative to "Group Dynamics," the major obstacle was removed when the Clarke County liaison responsibility was switched from an administrator to team teachers. Interestingly enough, the teachers have not experienced any negative consequences and have continued to have rhetorical support and no real interference. However, it is clear to the project staff that, without the motivation and commitment of these and the other team members, the project would not have been as successful or rewarding. Certainly, all of the researchers involved in this project have demonstrated extraordinary commitment.

Relative to "Paradigm Paralysis," this group has experienced the same inertia as any group (or individual) does when facing a new challenge; we tend to seek solutions from our experience rather than looking beyond our experience to other generalizable or transferable situations. Yet, it is exactly that behavior or generalizing and transferring which we desire to evoke in students. We have not seen any pattern in what causes individuals to make paradigm shifts. Some have moved because of frustration. Some have moved because of creative thinking. Some have moved because they have been sparked by others. The nudges which each project research has had to use to move away from our comfort zone to take risks and seek new paradigms serves as examples for the teachers to use as they, in turn, nudge their students to seek new paradigms.

The first benchmarks or indicators of paradigm shifts came in January, less than six months from the start-up of the project. At this time, the project staff reflected on the conversations occurring during the large-group meetings, the following shifts were documented (see TABLE 24):

TABLE 24

TIME LINE	
8/92	1/93
Less reflective	More reflective
Narrow perspective	Broader perspective
Simplistic understanding	Complex understanding
Has not been influenced	Has been influenced
Simplistic definition of innovative assessment	"Rich" definition of innovative assessment

These shifts in paradigms have continued to be evident as the teams have continued their work but are most marked in the six-month interval referenced above.

Finally, relative to "Frustration," this project has confirmed in the minds of the project staff that defining, describing, and implementing portfolio assessment (or perhaps any type of innovative assessment system) will cause frustration simply because there are no easy answers. And, in some cases, there are no answers at all. The science of innovative assessment is just beginning to emerge. Frustration will accompany that emergence and we had better learn to use that as a lever for moving forward rather than as a reason to fall back into our comfort zone of traditional assessment only. Some of the quotations from the teams listed in TABLE 12 indicate both the frustrations and the resolution of these frustrations.

TABLE 25

"It becomes clearer through our team efforts." (January 7, 1993)
"I'm really beginning to figure out our task." (January 7, 1993)
Mapping..."helped, clarifying the link between our documentation strategy and the Big Ideas..." (January 7, 1993)
"People are saying the same things but aren't able to hear each other." (January 7, 1993)

Whether or not these solutions have or will work to remove or lessen the problems is still an unanswered question. Some of the evidence lies in the successful use of the assessments. Some lies in the use of portfolio assessment consistent with this model after the project has ended. Some lies in the personal shifts made by the project partners. And, there is evidence[8] of shifts in thinking among the school team members. The first source suggests that the strategy for consensus-building and for using the assessment activities does work. What is not yet certain is whether the assessments are all scorable and whether than scoring can be done reliable and, finally, whether the results can be aggregated and remain meaningful. Other evidence is not available at this time.

Determination of Project Success

From the perspective of the project staff, the following will provide evidence of success:

- scorability of data across systems
- stakeholders' perception of meaningful information
- increased measurement sophistication of school-based teams
- continued commitment of school-based teams to innovative assessment to inform instruction

[8]Extracts from Daily Reflections

- articulation of a structure for portfolio assessment
- refinement of training techniques to more effectively work with school-based educators
- additional external support for extended work in portfolio assessment
- continued dialogue with measurement, curriculum, and instructional leaders across the country
- extended involvement and conversations among educators, scientists and mathematicians, and employers

Conclusion

The theoretical model for consensus building within the context of constructing portfolio assessments appears to be working. It provides a structure for decision-making which is useful in focusing the efforts of both novices and more experienced assessment developers. It reinforces (or allows the reinforcement) of constructs of interest (i.e., big ideas, habits of mind, and the like).

Emerging from this research is a notion of "structured portfolios." This notion calls for a core of structured documentation strategies. These strategies are structured in terms of the assessment stimulus and the evidence sought from students. However, the content and instructional activity which precede the assessment vary from classroom to classroom or even from student group to student group. The content becomes the contextual vehicle for eliciting the evidence which is documented through use of the structured assessment activity.[9]

During the first and second years of the project, the teams were challenged to brainstorm three or four assessment activities for use across the six school systems. These activities were to capture evidence about more than one project goal. They could be individual or collaborative in nature and could varying in their format. Thus, each team became the author of one assessment activity which would then be used across all six school systems.

During the early presentations of each team's favorite assessment "idea," it became clear that each team had operationalized the concept of portfolio assessment to allow them to use interviews as a method of documenting learning. There was considerable belief among the

[9]The single exception to this may be "Toys in Space" which is tied to specific content.

research partners that being able to have a dialogue between teacher and student was the "fairest" and "most valid" measure of what the student "really had learned." After much discussion to focus the purpose of the assessment with the format of the assessment and after reviewing the other assessment ideas, the teams were encouraged to consider a combination of assessment formats so that, across all six assessment activities, there would be a planned variation as to format, structure, type and amount of evidence, etc. The six assessments ultimately adopted by the teams for use in this project, the assessments vary along four dimensions: stimulus complexity, response complexity, context dependence, and amount of instructional time sampled.

The project staff developed two assessment activities for use across the school systems as well. These were specifically developed to contrast in format and structure from those developed by the school teams.

The structured portfolio should facilitate meaningful aggregation while embodying such powerful characteristics of innovative assessment as multiple strategies for problem solution and multiple solutions. At the same time, the structured portfolio entries are sufficiently well-defined and controlled so as to yield evidence which can be used in a comparative manner (over time, over students) and in an absolute manner (against performance standards).

To fulfill the concept of portfolio assessment in a manner more in keeping with the student-centered literature, this structured assessment core will be complemented by work samples representative of idiosyncratic student preferences, teacher preferences, classroom-specific experiences, and so forth. In addition, we will be working on the documentation of both student and teacher reflection in the final year. Specifically, we will use the Paulson's (1992) concept of reflection to develop project-wide strategies for documenting reflection. This concept separates reflection into four varieties: documentation (when students tell why they selected something for their portfolios), comparison (when students make comparisons of any kind), integration (when students review a body of work from a personal perspective), and presentation (when students reflect on their work from the perspective of others).

This approach of blending a structured core with the idiosyncratic selections of students and teachers is an extension of the Kentucky[10] model:

	On Demand	Extended
Uniform		
Local Option		

[10] 1991-92 Technical Report, Kentucky Department of Education, 1993.

Just as the Kentucky model calls for uniform and local option assessments, the structured part of the portfolio described above is uniform across students, schools, and systems. The "Local Option" component of the Kentucky model is analogous to the idiosyncratic portion of the portfolio assessment model described in this paper. Similarly, the structured portfolio assessment activities represent "on demand" assessments, whereas the idiosyncratic portions of each student's portfolio may be extended activities.

It is important to keep portfolio as instructional tool distinct from portfolio as assessment tool. Likewise, it is important to keep distinct portfolios as a collection to be judged as a whole versus portfolios to be judged as a collection of individual "things" which are judged independently and then merged/aggregated. Beyond those two issues, one need strive to reconcile the complexity desires and the practical limitations of resources. It is also important to use the big ideas underlying reform as clarifying variables to enhance the process of schooling and, in turn, of assessment.

In closing, as this research enters its final year, the project staff and research partners take heart again from the observations of others:

> "You can't expect these things to be perfect
> the first time around.[11]"

And:

> "Truth emerges more readily from error than
> from confusion."[12]

Emerging from this research is evidence that the process of defining types of entries which are both useful as a basis for judging student learning and which support the concept of portfolio assessment facilitates change in teachers' view and conduct of instruction. Similarly, there is emerging a realization of how difficult it is to develop assessments that honor the idiosyncratic nature of portfolios. It is both frustrating and rewarding to see the project partners struggle with the gap between traditional curriculum mandates and their new vision of science and mathematics assessment which has emerged from this project.

We are moving forward on our adventure which began with a vision of an assessment model which would empower teachers and students by leaving decisions about what should be

---

[11]Douglas I. Tudhope, Chair, Vermont State Board of Education in R. Rothman, "RAND Study Finds Serious Problems in Vt. Portfolio Program," Education Week, December 16, 1992

[12] Francis Bacon

taught and when at the classroom level while providing assessment frameworks which would represent the perspective of important student outcomes or big ideas across many classrooms and which would lead to meaningful, aggregatable data. We invite others to join us as we complete this adventure.

References

Arter, J., & Spandel, V. (1991, Spring) "NCME Instructional Module: Using Portfolios of Student Work in Instruction and Assessment," *Educational Measurement: Issues and Practice*, Vol. 11, No. 1, pp. 36-44.

Barker, J.A. (1990) *Future edge: Discovering the new paradigms of success.* New York: William Morrow Publisher.

Eisner, E. (1992) "The Misunderstood Role of the Arts in Human Development," *Phi Delta Kappan*, Vol. 73, No. 8, pp. 591-595.

Kentucky Department of Education. (1993). *Kentucky Instructional Results Information System.* Dover, NH: Advanced Systems for Measurement and Evaluation.

National Council of Teachers of Mathematics (1989) *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: The Council.

Paulson, F. L., & Paulson, P. R. (1990, August) *How do portfolios measure up?* Paper presented at "Aggregating Portfolio Data," Northwest Evaluation Association, Union, WA.

Paulson, F. L., & Paulson, P. R. (1992) Four varieties of self-refection. In Portfolios: Voices from the field, (K. Yancey, Ed.). Urbana, IL: National Council of Teachers of English.

Porter, A.C., Archbald, D. A., Tyree, A. K. (1991). Reforming the curriculum: Will empowerment policies replace control? *The politics of curriculum and testing*, pp. 11-36. Bristol, PA: The Falmer Press.

Ruthergold, F. J., & Ahlgren, A. (1989, 1990). *Science for All Americans.* NY: Oxford University Press for the American Association for the Advancement of Science, Inc.

Snow, R. (1989) *Abilities, Motivation, and Methodology.* Paper presented at the Minnesota Symposium on Learning and Individual Differences.

APPENDIX A

Guiding Questions

APPENDIX B

Description of The Kazakh-American School

The Kazakh-American School is a joint venture bi-lingual project which brings together Kazakh and expatriate children in this former Soviet Central Asian (Turkic) republic. The development of curriculum is vital for meeting the needs for intellectual, physical, and character growth as future citizens in this emerging society.

Duties of this position include: development of Social Studies, Mathematics, and Science curriculum, keeping in mind the possibility of "packaging" this curriculum for use in state schools around the republic; developing authentic assessment of student progress, both in English and in Kazakh, emphasizing portfolio assessment: building professionalism through the in-service development nd training of both Kazakh and American teachers: providing the in-service development and training of both Kazakh and American teachers; proving support for all teachers involved with the school; and to assist with administration as needed.

The Kazakh-American School seeks to provide for its students the highest intellectual and artistic growth in an environment of excellence, support and concern. The School seeks to train leaders in this emerging democracy and developing economy.1 Since citizenship, enterprise, and research are important skills for future developers of this society, our School seeks to instill critical thinking, strong communication skills, and open-mindedness.

APPENDIX C

Available Data

For those interested in digging into the process data for this project, there is a wealth of "stuff" available for scrutiny. Specifically, the following data sources are available upon request:

daily refections
logs
audiotapes
video logs of all group meetings
assessment drafts (including rubrics)
videotapes of all project meetings
video snapshots
monthly updates
written responses to guiding questions
questionnaire responses
letters
progress report
project participants (to interview)
student performances/responses
video snapshots (five 20-30 minutes excerpts of the project)
OERI video (20 minute project summary)